Using Academy Awards to Predict Success of Bollywood Movies using Machine Learning Algorithms

Salman Masih[1]
Department of Computer Science
University of Sialkot
Sialkot, Pakistan

Imran Ihsan[2]
Department of Computer Science
Air University
Islamabad, Pakistan

***Abstract*—Motion Picture Production has always been a risky and pricey venture. Bollywood alone has released approximately 120 movies in 2017. It is disappointing that only 8% of the movies have made to box office and the remaining 92% failed to return the total cost of production. Studies have explored several determinants that make a motion picture success at box office for Hollywood movies including academy awards. However, same can't be said for Bollywood movies as there is significantly less research has been conducted to predict their success of a movie. Research also shows no evidence of using academy awards to predict a Bollywood movie's success. This paper investigates the possibility; does an academy award such as ZeeCine or IIFA, previously won by the actor, playing an important role in movie, impact its success or not? In order to measure, the importance of these academy awards towards a movie's success, a possible revenue for the movie is predicted using the academy awards information and categorizing the movie in different revenue range classes. We have collected data from multiple sources like Wikipedia, IMDB and BoxOfficeIndia. Various machine-learning algorithms such as Decision Tree, Random Forest, Artificial Neural Networks, Naïve Bayes and Bayesian Networks are used for the said purpose. Experiment and their results show that academy awards slightly increase the accuracy making an academy award a non-dominating ingredient of predicating movie's success on box office.**

***Keywords*—Machine learning; supervised learning; classification**

I. INTRODUCTION

Bollywood releases a couple hundreds of movies every year with anticipation to reciprocate their investment that will only be possible if a movie succeeds. However, making a movie successful is not that easier, obviously a diverse audience with one sort of genre and cast could not help a lot. For instance, every individual has different expectation from a movie, some like comedy others do not and some prefer an actor. Well, making such diverse audience happy or entertaining them is quite challenging problem as entertainment is something that can't be quantified at all.

So, what we are supposed to do now? Apparently, a movie industry, Bollywood, must release a movie that is quite entertaining for the audience and will eventually become a success. Therefore, the question is how to predict degree of entertainment of a movie or its success or failure. Is there any way to predict movie success before its release or even before its production starts? Jack Valente, President and CEO of (MPAA)[1] once said, "No one can tell you how a movie is going to do in the marketplace. Not until the film open is darkened theater and sparks fly up between the screen and audience". This statement has just enlightened us about the complexity involved in predicting a success of a movie. Just to make it clear, success means a financial success at box office.

In year 2014, success rate of the movies particularly in Bollywood was quite disappointing ranging from 9-10% and wasted around 23.50 Billion Indian Rupees according to empirical data. The whole Bollywood industry only survives due to infrequent and limited blockbusters whereas majority of the movies could not recoup the total cost of production. Now, question arises that what are those ingredients, which help a movie to become a successful venture rather than a flop. Is it genre, actor, director, script, writer, music or combination of different elements?

Several factors that supposedly make a movie success on box office, few of them are traditional such as genre, leading actor/actress, director, production budget. Some non-traditional factors such as views of movie trailer on YouTube, likes of movie Facebook page, number of followers of leading role on Twitter. This situation provides us opportunity to investigate the impact of determinants of success. Since, filmmakers intend to make movie with high degree of entertainment and reciprocating the audience expectation. It becomes quite risky venture for them.

A significant research has been conducted for past decade. Researchers have used the traditional elements such as advertisement budget, number of opening theaters, and production studio etcetera. They have also extensively exploited social media for predicting the financial success of a movie such as number searches for the title of movie, tweets, Facebook likes and many more.

Most of the previous work [1, 2, 3] focused on post-release or post-production and with the help of word-of-mouth data they have shown good accuracy level. It is worth mentioning here that prediction made at post-release stage even with higher accuracy is less significant for all the stakeholders. We focus

[1] www.mpaa.org

on pre-production prediction to make the idea more persuasive. We have observed that winning a 'Zee Cine' and 'IIFA' award is quite competitive for any actor. It really fascinates us to understand the relationship between awards and movies success. As a result, being a good avenue to be explored, this paper identifies the significance of awards won by leading actors/actresses specifically 'Zee Cine' and 'IIFA' for predicting movies success.

II. Literature Review

Experts in different domains that include economists, marketing strategists, word-of-moth (WOM) experts and neural network scientists have conducted a significant amount of research. It is unfortunate that Bollywood has never been a focal point for research as most of the experts have considered Hollywood for building and evaluating their models. There is another research gap that none of the state-of-the-art approaches has considered academy awards such as IIFA and Zee Cine as a determinant for predicting success of movies. Authors in [4] did the pioneer work in the domain of movie revenue prediction. Their approach used ANN (Artificial Neural Networks) to predict movie success and they were the first one who converted the problem into classification by making different classes of revenues. They have used the following variables for prediction, MPAA rating, competition, star value, genre, special effects, sequel, and number of screens. The most recent work we studied is [5] who have employed the same number of variables with DANN (Dynamic Artificial Neural Network) and few more variables such as production budget, pre-release advertising budget, runtime and seasonality. Search engine query data has also been used for predicting movie success [1].They have employed a simple regression using movie query data from Google and Income, Rate number of theaters from box-office mojo and number of words in title. This research could not achieve better results that show that simple linear regression and the variables they have used are not significant for making movie a blockbuster or flop. There is another research [6] which has used linear regression with different variables such as movie revenue, pre-launch and post-launch period and music-trdscore (trend score of soundtrack of a movie searched over the Google during pre-launch week) and music-existing (which means whether the soundtrack used in movie is existing one or not) which results still can be improved. The word-of-moth [6] experts have also exploited tweets and build a hybrid model for revenue forecasting that shows that number of tweets can predict the movie revenue. Several machine learning algorithms have also been tested to predict the movie revenue [7] using most frequently used variables such as director, actors and genre could not achieve much accuracy. The last paper we have reviewed so far [8] had tried to examine social media influence on profit of a movie which has shown that facebook.com likes are not a good determinant for forecasting movie profit. Now all the work has missed something that is prediction time. Predicting a movie just before release does not mitigate the possibility of loss as all the resources have been invested. [5] has initiated the new research direction which is predicting a movie before its launch and this study is moving in the same direction. There are some authors who have also explored many other ideas for early movies predictions [9] has considered the date of release, [10] tried to blend all previously shown predictive power in different studies [11].

III. Proposed Methodology

Studies have revealed most of the prediction is based at post-production level without using academy awards information either for predicting or evaluating purposes. Based on this research gap, whether the awards won by the leading actors play any role in predicting the movie success at pre-production level or not, the adopted methodology is shown in Fig. 1.

The very first methodological step was to collect the data. Data were collected from three different resources and furthermore these three sources were also used for pre-processing in case of missing, erroneous values. Later, required attributes were separated from the unsolicited data. In the next step, the whole data were preprocessed which includes removing noise, class assignment. Once we have pre-processed data, several models were trained and tested using experimental setting. WEKA[2] was employed which has several renowned classifier and clustering algorithms. All the selected classifiers have been previously explored by different research to evaluate their hypothesis. Numerous research studies analyzed different independent variables to predict the success of movies. Conclusions of different variables predictive power was considered while selecting the data and independent variables. Let's investigate them in detail.

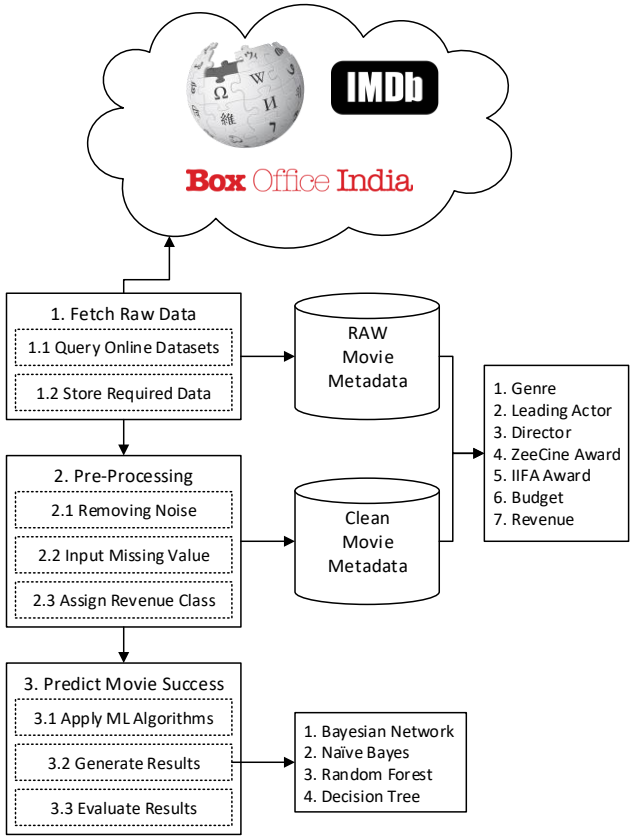

Fig. 1. Methiodology.

[2] http://www.cs.waikato.ac.nz/ml/weka

A. Data Acquisition

Data was collected from three different sources that include Wikipedia[3], IMBD[4] and BOI[5]. Initially, titles of movies, genre, director and cast were retrieved from Wikipedia and then awards of leading roles were collected from both IMDB and Wikipedia. Then the budget and revenue of each movie was retrieved from BOI.

B. Variable Description

The dependent variable or response variable in this paper was profit. We followed the footsteps of pioneering research work in this domain [12] by converting the simple point estimation problem into a classification. We assigned a class to each movie according to specified range. Table 1 defines class name and its associated profit range.

Table 1 shows a discrete number of classes. There can be multiple reasons for converting the values of continuous variable into a discrete number of classes according to specified ranges. First and most important preference for using a discrete number of classes as compared to continuous values is better knowledge representation [13] making data simplified and reduced through discretization. Secondly, it is quite easier to infer the variables if they are divided into several ranges [14] and learning algorithms get faster as they do not need to check every single value & just pass through different intervals [15]. For each range, a total of six explanatory variables were selected for experimental purposes. Our selection of variables is purely based on previous research studies conducted in domain of movies revenue prediction with an addition of Awards. Genre, Director and Leading role are nominal attributes and Budget, Zee Cine and IIFA awards are numeric attributes. A brief description about each explanatory variable has been given below.

TABLE I. CLASS VS PROFIT RANGE

Class	Profit Range (In Millions)
A	>= 900
B	< 900 and >= 800
C	< 800 and >= 700
D	< 700 and >= 600
E	< 600 and >= 500
F	< 500 and >= 400
G	< 400 and >= 300
H	< 300 and >= 200
I	< 200 and >= 100
J	< 100 and >= 000
K	= 0
L	< 0

[3] www.wikipedia.org
[4] www.imdb.com
[5] www.boxofficeindia.com

1) Genre: It is quite difficult to define a genre of a movie as story told in 2.5 to 3 hours may not follow only one genre and recently most of the Bollywood directors have started blending three to four genres together to make movie entertaining one and capturing the more audience. For instance, a bollywood movies titled "PK", released in 2014 had three different genres i.e. 'Comedy', 'Drama' and 'Romance' according to IMDB, however, it was purely a 'Drama' movie. Despite all the facts, genre is one of the most commonly used as explanatory variable for movies revenue prediction, but its contribution is yet to be concluded [16]. However, some studies have concluded that only a few genres have predictive power to forecast the movies revenue [17]. Based on the study conducted on Bollywood movies [18] that concluded that genre does impact the movies box office performance. We were intrigued to evaluate its contribution with or without awards in our case.

2) Leading role: Leading Role, actor or actress is another import factor that influences the performance of movies on box office. It has been widely used by majority of the research studies to evaluate its predictive power [12,19,20]. However, charisma of leading role does not work in many cases. A leading role played by either actor or actress was taken as a parameter in our study but we did not followed the conventional method of calculating the weights of actor by their mean salaries, number of follower on social media or ratting on different websites like IMDB as female gets less salary than their male counterparts, some actors are not much active on social media but still successful. Based on these facts we have taken only name of leading role played either by actoress or actor to check it predictive power with or without awards.

3) Director: Director is another important and underappreciated factor that may influence a movie success. A story not well directed can cause a huge loss to a movie, so director's impact should be considered while making any decision related to movies. Therefore, we included the director as another important parameter in our study. We used the name of director for a movie only. A number of research studies have found no predictive power of director in forecasting movie success [16,18] although few studies have different results than the majority [21]. We believe that director has a positive influence on success of a movie. For example, Bollywood's directors "Rohit Shetty" and "Tigmanshu Dhulia" both have directed five films in five years. It is quite surprising that "Rohit Shetty" got all of five hits but "Tigmanshu Dhulia" was unable to deliver a single hit at the box office. This case has raised many questions about the director's impact for movies performance at box office. Therefore, we included this parameter in our study.

4) Production budget: Production budget has been regularly included in research studies and came out as a powerful predictor of movies' revenue. Empirical data suggests that high budget movies tend to generate high revenue and yet it does not comply with high profit.

Increasing the budget may help to increase the revenue but not the profit. Moreover, average budgeted movies are more profitable than high budget movies. For example, a Bollywood movie "Boss", produced on the budget of 700 million INR, could only earn 850 million INR. Whereas another movie "Chashme Baddoor", produced on the budget of 200 million INR, earned around 628 million INR. However, looking at the past research as majority of the studies have included the budget an explanatory variable, our study explored the predictive power of budget in relation with awards.

5) Zee cine awards: Zee Cine Awards or ZCA for short according to their Wikipedia page[6] founded in late 90s and has been successfully conducted around 21 awarding ceremonies till date. ZCA has three types of awards namely, 'Jury's Choice Awards', 'Viewer's Choice' and 'Technical Awards'. Jury comprising of veteran actors & actresses and organized by ZCA, makes these awards more competitive and credible. Viewer's choice awards are awarded based on votes from general audience, a true representation of public opinion and value of an actor. These both characteristics make awards an optimal choice for predicting movies success. We collected all the awards won by the leading role either actor or actress under any capacity. We opted out technical awards because most of movie budget goes to a leading role whereas dancing crew, makeup artist and set designer get a very slight share of budget.

6) IIFA awards: International Indian Film Academy Awards or IIFA awards[7] as per their Wikipedia page started in back 2000. They have three types of awards but unlike Zee Cine they have a three different categories namely 'Special Awards', 'Popular Awards' and 'Technical Awards'. Voting procedure is same as Zee Cine except nominees are scrutinized by the member of jury before getting public opinion. However, still have characteristics and qualify for being included in our forecasting model. IIFA has been held 18 times till date and the recent one was held in Bankok, Thailand on 22-24 June 2018. We have included all the awards won by leading role under any capacity.

C. Pre-Processing

After data acquisition, data was cleaned by initially removing the unwanted values like brackets and punctuation marks around the name of actor, director and genre. There were some missing entries as well that were filled in manually by searching various source websites. Production budgets and revenue of movies were in crore INR. We converted both production budget and revenue into millions and calculated the profit by subtracting production budget from total revenue earned. Later, profit range classes defined in Table 1 were assigned. Finally, all records ware saved in .csv format.

[6] http://en.wikipedia.org/wiki/Zee_Cine_Awards

[7] http://en.wikipedia.org/wiki/International_Indian_Film_Academy_Awards

D. Classifiers and Experimental Settings

Various methodologies have been practiced by different studies over the years starting from linear regression to neural networks. We have chosen Naïve Bayes, Bayesian Networks, Decision Tree (J48) and Random Forest for our experimental purposes. Naïve Bayes and Bayesian network all selected classifiers have been used to build predictive models in domain of movies [22,23]. Naïve Bayes out-performed its counterpart decision tree J48 algorithm and has shown the same accuracy equal to neural networks [12, 16]. Despite its idealistic attribute's independence supposition, its performance has been surprising in many experimental studies [24]. Bayesian networks were frequently spoken as Bayes nets are probabilistic graphic models. They represent the conditional dependency of random variable via acyclic graphic graph. Bayesian network has been popular in the domain of text mining, language processing and forecasting.

We have chosen statistically rigorous experimental design methods for objectively analyzing the performance of models known as k-fold cross-validation also referred as rotation estimation. In k-fold the whole dataset (D) is randomly divided into folds of equal size (D_1, D_2, D_3… Dn). The model is trained and tested *k* number of times. This way almost every instance of the dataset gets a chance of being included in the training and testing data. According to the nomenclature of data mining 10-folds are highly recommended for splitting the data to train and test the classifier [12], it also considers the bias and variance tradeoff. There are many other approaches through which data are split to train and test the model. Split ratio is the mostly practiced in machine learning. In split ratio normally 60% of data is used for training and 20% testing and 20% for cross- validation.

E. Performance Evaluation Metrics

We employed Precision, Recall, F-Measures and weighted averages of each measure were calculated to evaluate accuracy of classifier. Precision tells us out of total instances classified by the classifier as positive how many were positive. What percentage out of all positive examples was picked up by the classifier is calculated with the help of recall. In ideal setting the precision and recall would be equal to 1.0 which implies completely an accurate model. However, keeping the balance between both things is quite difficult and especially achieving high precision. F-Measure is breakpoint between the both recall, and precision also written as F-Score or F_1-Measure.

$$\text{Weighted Average of Precision} = \frac{\sum_{i=1}^{n} P_i W_i}{\sum_{i=1}^{n} W_i} \quad (1)$$

$$\text{Weighted Average of Recall} = \frac{\sum_{i=1}^{n} R_i W_i}{\sum_{i=1}^{n} W_i} \quad (2)$$

$$\text{Weighted Average of F-Score} = \frac{\sum_{i=1}^{n} F_i W_i}{\sum_{i=1}^{n} W_i} \quad (3)$$

IV. Experiment and Results

The experiments performed, and the results tabulated are divided into three levels. First level of experiment described findings using single feature experiment combined with awards. The second level used two features in combination with awards to tabulate results. And in the last layer, n-number

of features were used with awards. Let's investigate the selected dataset and its statistics first, before investigating the results of each experiment setting in detail.

A. Dataset Statistics

Exploratory data analysis is highly recommended in statistics community to get initial insights about the data. Therefore, an exploratory data analysis was performed to summarize statistics of the whole five-year dataset with different classes according to the profit movies earned has been shown in the Table 1. Collected dataset has 522 movies starting from year 2013 to year 2017 with only 6% earned 1000 million or more getting Class 'A' and majority of the movies as super flops resulting in class 'L'. The mean and standard deviation of both 'A' and 'L' classes were 135.347, 528.7 respectively.

Initially, all the genres were included in data set but later through re-sampling of the data in WEKA, the rare one was removed automatically to reduce the class bias. Majority of the releases in past five years had genre as '*Comedy*', '*Romance*' and '*Drama*'. The second popular genres were '*Adult*', '*Crime*' and '*Social*'. A total of 409 directors, around 297 different actors were included in complete dataset.

B. Single Feature with Awards

In order to assess the predictive value of awards and of their combination with other features, single feature power combined with awards was tested in first experiment to see whether accuracy increases or decreases with or without inclusion of awards. Genre has always been included in many research studies previously and found to be a significant contributor of movie success. In our experiment, first single feature selected was '*Genre*' to predict success of movie and achieved 0.53 F-Score with Random Forest classifier performing the best. Fig. 2 shows the results with four different classifiers with Naïve Bayes and Bayesian Network both with low accuracy.

As seen in Fig. 2, Naïve Bayes, Bayesian network performed at same level but J48 was worst with F-score of 0.48. Next step was to evaluate the difference in predictive power of genre when it is combined with the awards. F-Score was significantly increased up to 13% by combining the genre and awards in case of Random as shown in Fig. 3. Bayesian Network, Naïve Bayes also showed an increment in F-Score but only few percent. It was surprising that J48 had not shown any changes in F-Score.

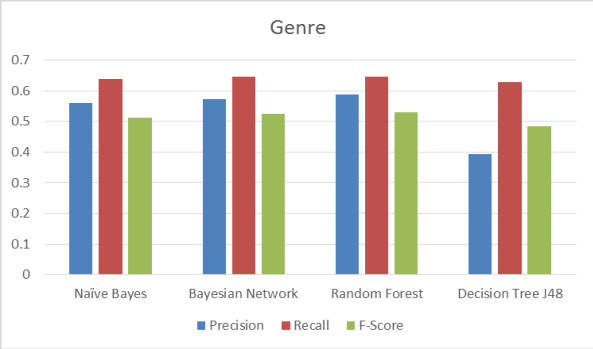

Fig. 2. Predictive Power of Single Feature–Genre.

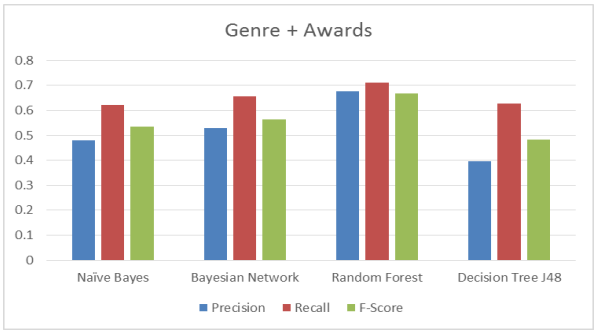

Fig. 3. Predictive Power of Single Feature–Genre with Awards.

The second determinant to be evaluated with its predictive power was '*Director*'. Most of studies have stopped using director attribute as it did not show any predicative power [25]. However, our experiment found that only '*Director*' had shown more predictive power than the '*Genre*' and awards combined as shown in Fig. 4 using Random Forest classifier. We were compelled to disagree with previous studies regarding the predictive power of director as we found almost 3% improved accuracy when combined '*Director*' with awards. Fig. 5 shows the results.

Leading role or star has remained a crucial ingredient for movie success and it has similar importance when it comes to prediction of movie success. Despite of different methodologies employed to calculate star weights, results are still comparable. However, we found a bit different results than the previous studies. Leading Role has less predictive power than director achieving up to 0.73 F-Score alone with Random Forest as shown in Fig. 6. Awards affected the results when combined with the leading role by increasing accuracy for Naïve Bayes, Bayesian Network and Random Forest as shown in Fig. 7.

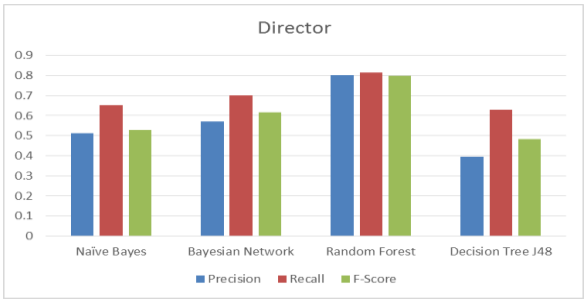

Fig. 4. Predictive Power of Single Feature–Director.

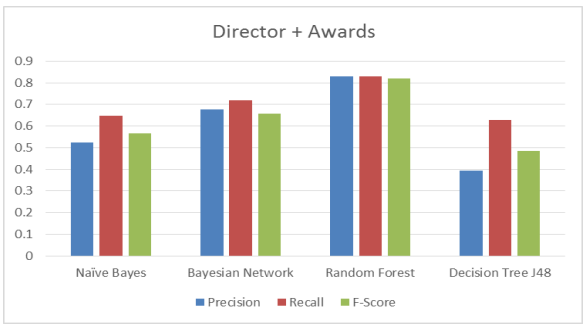

Fig. 5. Predictive Power of Single Feature–Director with Awards.

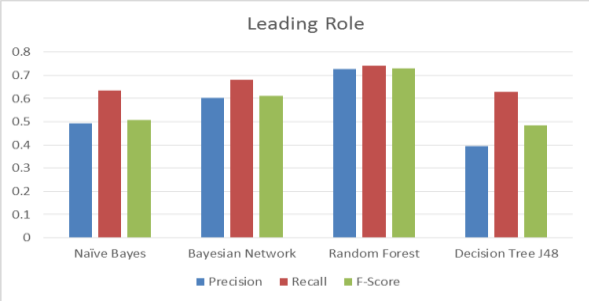

Fig. 6. Predictive Power of Single Feature–Leading Role.

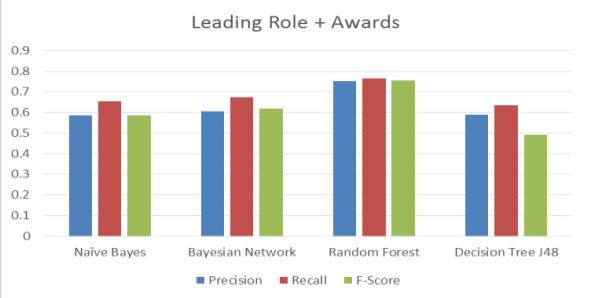

Fig. 7. Predictive Power of Single Feature–Leading Role with Awards.

Budgets are generally considered the true determinants of movies as high budget means expensive cast and specifically popular leading role. Our experiment suggests that '*Budget*' has less predictive power than '*Director*' attribute as shown in Fig. 8. Moreover, combining awards and budget showed less accuracy than director and awards combined. Budget has more predictive as compared to other attributes alone or combined awards except director as shown in Fig. 9.

We also performed an experiment using the feature '*Awards*' only and the results are shown in Fig. 10.

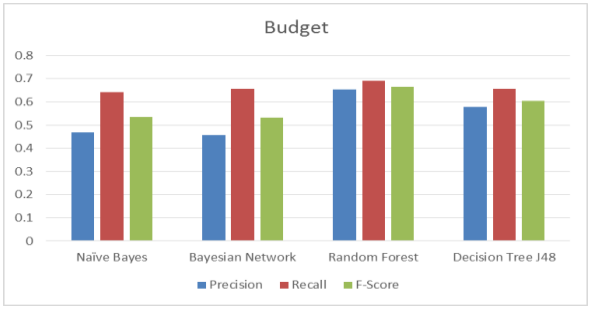

Fig. 8. Predictive Power of Single Feature–Budget.

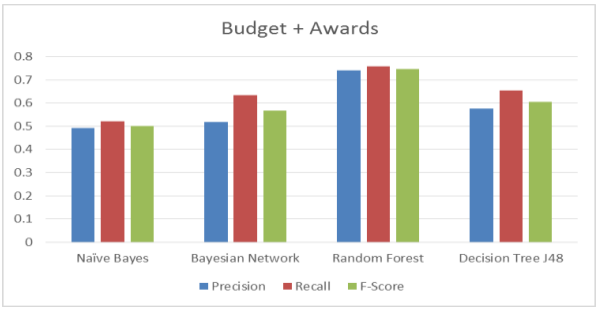

Fig. 9. Predictive Power of Single Feature–Budget with Awards.

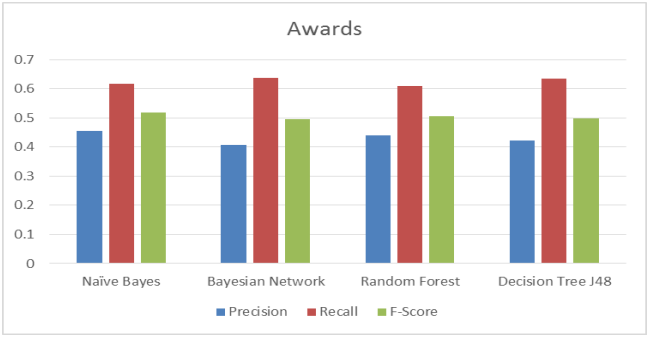

Fig. 10. Predictive Power of Single Feature–Awards.

Based on the results, it is evident that '*Awards*' have similar predictive power as compared to '*Genre*' but less than '*Director*' and '*Leading Role*', yet awards improve the accuracy when they are combined with other features.

C. Bi-Feature with Awards

We have so far seen all the plausible combination of single feature with awards and in this next experiment we tried a bi-feature combination with awards. Some abbreviations have been in used in results and these abbreviations are:

1) *A: Awards*
2) *B: Budget*
3) *D: Director*
4) *G: Genre*
5) *LR: Leading Role*

The gist behind using bi-feature combination was to evaluate the power of awards when they were combined and what was the best set of features to make accurate predictions. For instance, if we had combined '*Genre*' with '*Director*' and then adding '*Awards*', what improvement can be calculated in accuracy. Similarly, '*Director*' with '*Leading Role*' would make any difference or not. If we were able to make prediction with same accuracy using only one feature rather than two features, then it would be useless to use more feature. Therefore, we tried several feature combinations and evaluated their combined effects with awards on accuracy.

In the experiment, initially '*Genre*' and its different plausible combination with other attributes with and without inclusion of '*Awards*' was tested and results are shown in Fig. 11. Experiment suggested that '*Genre*' and '*Director*' had more predictive power than any other combination of attributes and adding award improved the accuracy a bit but not that significant. '*Genre*' and '*Budget*' also showed the same accuracy and its effect when '*Awards*' are added. '*Genre*' and '*Leading Role*' had significantly less accuracy. Next, we experimented upon all plausible combination using '*Director*' as shown in Fig. 12. The results show that '*Director*' and '*Budget*' provide the maximum predictive power.

Next combination was '*Leading Role*' with its plausible combinations and result shown that it works best with '*Budget*'. Adding '*Awards*' showed a minor improvement in accuracy. Combining '*Leading Role*' with '*Budget*' had almost the same accuracy of '*Director*' as shown in Fig. 13.

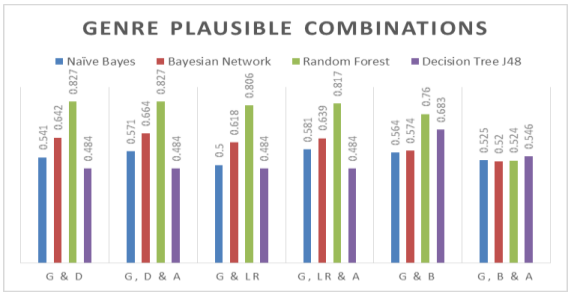

Fig. 11. Predictive Power of Bi-Feature Combination with Genre.

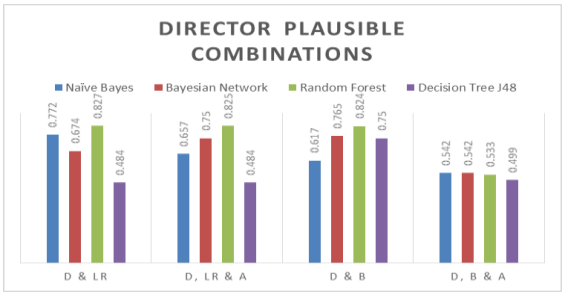

Fig. 12. Predictive Power of Bi-Feature Combination with Director.

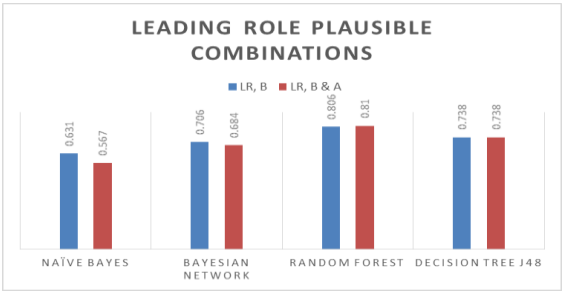

Fig. 13. Predictive Power of Bi-Feature Combination with Leading Role.

D. N-Features with Awards

Next experiment was to use N number of features combined. 3 and 4 number of features were combined with '*Awards*'. There was a slight difference in accuracy when '*Awards*' were combined in case of Random Forest however J48, Bayesian Network and Naïve Bayes had shown quite different results as depicted in Fig. 14.

Finally, we tried to evaluate the effects of award when they were combined with rest of the features. The F-score without awards using Random Forest classifier was 0.825 and it improved up to 0.83 when awards were added as another feature. It was quite intriguing that rest of the classifier could not capture the difference and in case of Naïve Bayes classifier the F-score was dropped to few percent. Bayesian Network did not show much difference. J48 surprisingly improved with and without awards. We ended with a conclusive experiment which evidently proved that awards do have predictive power though a slight one when combined with other parameters. A total of four classifiers were tried and Random Forest performed the best. The results of all features with and without awards are shown in Fig. 15 and Fig. 16, respectively.

Combining all the results for Single-Feature, Bi-Feature and Tri-Feature with and without Awards shows that Random Forest performs better than Naïve Bayes, Bayesian Network and J48. Moreover, Random Forest performs best for Tri-Feature is used in conjunction with Awards as shown in Table 2.

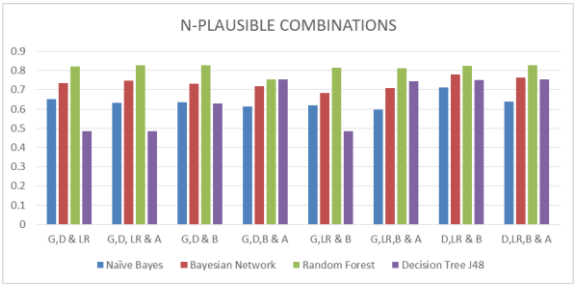

Fig. 14. N-Features Combination with Awards.

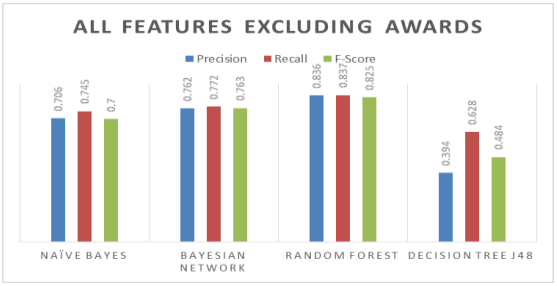

Fig. 15. All Features Excluding Awards

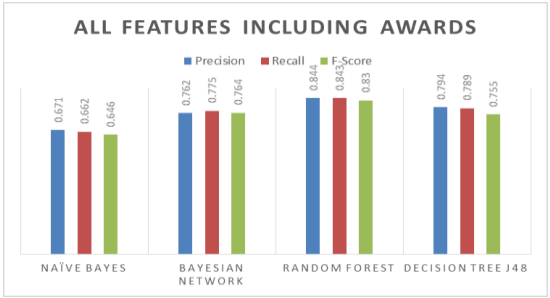

Fig. 16. All Features Including Awards.

TABLE II. F-SCORE AVERAGE

Algorithm Used	Single Feature		Bi – Feature		Tri – Feature	
	Without Awards	With Awards	Without Awards	With Awards	Without Awards	With Awards
Naïve Bayes	0.4890	0.5348	0.6023	0.5828	0.6673	0.6315
Bayesian Network	0.5580	0.5565	0.6523	0.6483	0.7340	0.7260
Random Forst	0.6615	0.7183	0.8037	0.8145	0.8290	0.8325
Decision Tree J48	0.4703	0.5815	0.5762	0.6005	0.6710	0.6708

V. Discussion and Conclusion

Predicting movies success has always been a quite challenging and interesting problem for the researchers due to its high association with unpredictability. It has attracted researchers from different domains which includes computer scientist, econometricians marketing strategists and WOM experts. It is quite unfortunate that only a limited number of studies had tried the Bollywood for predicting their success. We did not find any reason and we do not want to suppose as well. However, unfamiliarity with sophistication of computer application in predicting movies could be a plausible cause. Majority of the previous research had focused on the post-production prediction and especially WOM experts are inclined to make such predictive models with higher accuracy. Predictions made after production even with high accuracy are of limited and do not help that much to influence the movie revenue.

In our research, we have evaluated the predictive power of two commercial awards won by the leading role and influence of their power on the other parameters previously explored for predicting the movie success. Moreover, unlike the previous research we have tried to measure the accuracy of models at pre-production level. Genre has been included in most of the study as success determinant in movies domain and we did it as well to reevaluate its predictive power. It turned out that the genre has good predictive power but not as much as other parameter had shown in our case. Its performance increased when combined with awards up to 13 percent which is quite significant. Our research suggests that genre should be included in further studies as well. The next parameter 'Director' showed significant predictive power and disagreed with many previous conclusions that director did not play any role for predictive power for predicting movie success. However, our results showed director alone has more predictive power than a leading role with and without awards. This finding is the major contribution of our study.

Leading Role as recommended by previous studies a major movies success determinant. Our results show it has less predictive power than both budget and director which put this parameter at third position in our research. Accuracy increased when awards were combined with the leading role but still this improvement did not dominate the both budget and director. Budget is one of the most widely known ingredients of success. It has almost included in every forecasting previous studies. Empirical data shows that increasing the budget may not always help to produce a success of product as in most of the cases medium level budgeted movies are more likely to succeed. Well, talking about its predictive power, it has won the race with all other attributes except director. It has shown less predictive power than director with and without awards.

Results show that awards have equal predictive power as compared to genre but did not win the race with director, leading role and budget. Awards combined with the director parameter have shown the highest predictive power than in other combination in all our experiments. To conclude, we can say that awards have good predictive power when they are combined with director and combining them with other parameters have also shown significant improvement in accuracy.

References

[1] L. Chanseuung and J. Mina, "Predicting Movie Income Using Search Engine Query Data," in Conference on Artificial Intelligencce and Pattern Recognition, Kuala Lampur, Malaysia, 2014.

[2] B. DeSilva and R. Compton, "Prediction of foreign box office revenues based on wikipedia page activity," in WebSci, Bloomington, Indiana, 2014.

[3] D. Delen and R. Sharda, "Predicting the Financial Success of Hollywood Movies Using an Information Fusion Approach," Industrial Engineering Journal, pp. 21(1), 30-37., 2010.

[4] R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks," Expert Systems with Applications, pp. 243-254, 2006.

[5] M. Ghiassi, D. Lio and M. Brian, "Pre-production forecasting of movie revenues with a dynamic artificial neural network," Expert Systems with Applications, pp. 42(6), 3176-3193., 2015.

[6] X. Haifeng and G. Nadee, "Does Movie Soundtrack Matter? The Role of Soundtrack in Predicting Movie Revenue.," Department of Information Systems, National University of Singapore, Singapore., pp. 1-10, 2014.

[7] P. Sharang and S. Mevawala, "BoxOffice: Machine Learning Methods for predicting Audience Film Ratings," The Cooper Union for Advancement of Science and Art., 2014.

[8] S. Shruti, S. Deb Roy and W. Zeng, "Influence of social media on performance of movies," University of Missouri, Columbia, Missouri, 2014.

[9] M. T. Lash and K. Zhao, "Early Predictions of Movie Success: The Who, What, and When," Journal of Management Information Systems, pp. 33(3), 874–903, 2016.

[10] S. Darekar, P. Kadam, P. Patil and C. Tawde, "Movie Success Prediction based on Classical and Social Factors," International Journal of Engineering Science and Computing, pp. 50-62, 2018.

[11] J. Hofmann, M. Clement, F. Völckner and T. Hennig Thurau, "Empirical generalizations on the impact of stars on the economic success of movies," International Journal of Researc in Marketing, pp. 442-461, 2017.

[12] R. Sharda and D. Delen, "Predicting Box-Office Success of Motion Pictures with Neural Networks," Expert Systems with Applications, pp. 30, 243-254, 2006.

[13] H. Simon, "The Sciences of the Artificial," Cambridge, MA, 1981.

[14] H. Liu, F. Hussain, T. Chew and M. Dash, "Discretization An Enabling Technique," Data Mining and Knowledge Discovery, pp. 6(4), 393–423., 2002.

[15] J. Dougherty, R. Kohavi and M. Sahami, "Supervised and Unsupervised Discretization of Continuous Features," in ICML, Los Altos, CA., 1995.

[16] A. Elberse and J. Eliashberg, "Demand and Supply Dynamics for Sequentially Released Products in International Markets: The Case of Motion Pictures," Marketing Science, pp. 22(3), 329-354, 2003.

[17] A. Ainslie, X. Drèze and F. Zufryden, "Modeling Movie Life Cycles and Market Share.," Marketing Science, pp. 24(3), 508-517, 2005.

[18] M. Fetscherin, "The Main Determinants of Bollywood Movie Box Office Sales," Journal of Global Marketing, pp. 23(5), 461-476, 2010.

[19] S. Basuroy, S. Chatterjee and A. Ravid, "How Critical are Critical Reviews? The Box Office Effects of Film Critics, Star Power, & Budgets," Journal of Marketing, pp. 67, 103-117., 2003.

[20] N. Terry, M. Butler and D. De'Armond, "The Determinants of Domestic Box Office Performance in The Motion Picture Industry.," Southwestern Economic Review, pp. 32, 137-148., 2005.

[21] B. Litman and L. Kohl, "Predicting financial success of motion pictures: The ’80s Experience," Journal of Media Economics, pp. 2, 35–50, 1989.

[22] R. Neelamegham and P. Chintagunta, "A Bayesian Model to Forecast New Product Performance in Domestic and International Markets," Marketing Science, pp. 18(2), 115-136, 1999.

[23] M. Sawhney and J. Eliashberg, "A parsimonious model for forecasting gross box-office revenues of motion pictures.," Marketing Science, pp. 15(2), 113-131., 1996.

[24] H. J. George and L. Pat, "Estimating Continuous Distributions in Bayesian Classifiers," in UAI, San Mateo, 1995.

[25] R. Nelson and R. Glotfelty, "Movie Stars and Box Office Revenues: an Empirical Analysis," Journal of Cultural Economics, pp. 36, 141-166, 2012.