Journal Of Harmonized Research (JOHR)

Journal Of Harmonized Research in Engineering
8(3), 2020, 81-88

ISSN 2347 – 7393

Original Research Article

MOVIE SUCCESS PREDICTION USING RANDOM FOREST (RF) CLASSIFIER

Chintu Kumar

Research Scholar, Sri Satya Sai University of Technology and Medical Sciences, Sehore (M.P.), India

Abstract: The prediction of movie success the exhaustive goal of various Film Industries. Whereas, it becomes more critical to execute the movie successfully. To predict the movie success various data mining and machine learning techniques such as GuassianNB, Multinomial NB, BernoulliNB, KNeighnors Classifier, Decision Tree, Logistic regression has been developed but, in this work, we use random forest classifier for the prediction of movie success with reduced cost and schedule. The random forest classifier selects the dataset randomly from the available dataset and the generate the decision tree of the selected dataset and then apply the voting on the prediction results and whose score and accuracy will be maximum that will indicates the success of movie. For the sample of IMDb dataset, we use online resource of kaggle and the experimental results is generated from the widely used machine learning programming language Python which helps in the analysis of the proposed methodology. The performance of proposed methodology is measured using the parameters such as Score, accuracy, precision, recall value, F1 score, mean absolute error and mean square error. The comparative analysis of the proposed methodology is done among the existing approach GuassianNB, MultinomialNB, BernoulliNB, KNeighnors Classifier, Decision Tree, Logistic regression. The score and accuracy value of our proposed methodology is 70% while other is less. Similarly, the F1 score, precision and recall value of proposed methodology is 69%, 66% and 66% while the BernoulliNB, MultinomialNB and logistic regression is comparatively very less. Similarly, the comparative analysis of proposed and existing approach is done using mean absolute error and mean square error and the value is 32% and 36% which is very less. These results of proposed methodology improve the success rate of movie success.

Keywords: *RF, Logistic Regression, Prediction, IMDb, GNB, MNB, BNB, Precision, Python*

For Correspondence:
lkchoudhary421@gmail.com.
Received on: March 2020
Accepted after revision: July 2020
Downloaded from: www.johronline.com

Introduction: Movie industry is a huge sector for investment but larger business sectors have more complexity and it is hard to choose how to invest. Big investments come with bigger risks. The CEO of Motion Picture Association of America (MPAA) J. Valenti mentioned that 'No

one can tell you how a movie is going to do in the marketplace. Not until the film opens in darkened theatre and sparks fly up between the screen and the audience' [1]. As movie industry is growing too fast day by day, there are now a huge amount of data available on the internet, which makes it an interesting field for data analysis. Predicting a movie success is a very complex task to do. The definition of a movie success is relative, some movies are called successful based on its worldwide gross income, and some movies may not shine in business part but can be called successful for good critic's review and popularity.

A movie revenue depends on various components such as cast acting in a movie, budget for the making of the movie, film critics review, rating for the movie, release year of the movie, etc. Because of these multiple components there is no formula that helps us to provide analysis for predicting how much revenue a particular movie will be generating. However, by analyzingthe revenues generated by previous movies; a model can be built which can help us predict the expected revenue for a particular movie. Such a prediction could be very useful for the movie studios which will be producing the movie so they can decide on different expenses like artist compensations, advertising of the movie, promotions in various cities, etc. accordingly.

Plus, it allows investors to predict an expected return-on-investment (ROI). Also, it will be useful for many movie theatres to estimate the revenues they would generate from screening a particular movie.[2] Now a day's, online review system has become one of the most important part of any business approach. Posting reviews online for products bought or services received has become a trendy approach for people to express opinions and sentiments, which is essential for business intelligence, vendors and other interested parties. Social media contains rich information about people's preferences. There are many elements impacting the movies of a film, for instance, number of screens for the motion picture, publicizing, time, actors, directors, budget, genre and number of motion pictures that are released in specific duration or time frame and even within past years, months and days. Motion pictures or movies have turned into a vital piece of our lives as manner for passion, compassion, enthusiasm and entertainment. [3] Films have likewise been a noteworthy medium for culture trade between various nations and districts and are therefore an irreplaceable resource for the world. Given this, the motion picture industry has turned into a business and it has enormous market benefit and potential. As an outcome, the information and research about the motion picture industry is getting to be noticeably more profound. Capacity to precisely foresee the movies potential returns over investment based on total cost of ownership for a motion pictures will enable the film line decides the publicity cost and time of demonstrating the motion picture to expand the benefit and returns to investment made therein. The issue of foreseeing the movies gross of a releasing film has been broadly handled in the past from a measurable perspective. There are many elements impacting the movies of a film, for instance, number of screens for the motion picture, publicizing, time, actors, directors, budget, genre and number of motion pictures that are released in specific duration or time frame and even within past years, months and days. Under this scheme we are centering to build up a strategy in light of affiliation using machine learning to upgrade the forecast of success of the movie.

In this paper, we propose a random forest classifier for prediction of movie success and this classifier is compared with the existing machine learning techniques Guassian NB, Multinomial NB, Bernoulli NB, KNeighnors Classifier [5], Decision Tree [6], Logistic regression [7] . With the increasing in huge amount of data a good data analysis is required and machine learning approach is one of them. Nowadays this approach is extensively used for data analysis purpose. The comparative analysis of proposed and existing technique is done using various performance measuring parameters such as Precision, recall, F1 score, mean square error, mean absolute error, root

mean square error etc. and it is analyzed that our proposed approach outperforms than the existing approach. It means our approach is much better in the prediction of movie success rate. The organization of rest paper is done as follows: In section 2, we present the review of literature work and section 3 briefly discuss our proposed model for the prediction of movie success. In section 4, explaining the result analysis of our proposed model. In last section, we present the overall conclusion of the proposed work with their future aspects.

Related Work: *Kumar, and Kumar (2018)* proposed framework predicts the achievement of a motion picture in light of its gainfulness by utilizing chronicled information from different sources. Examination comes about with motion pictures amid years' time frame demonstrated that the framework beats benchmark techniques by a substantial edge in anticipating motion picture productivity.[8]

Chaudhari et al. (2016) developed a tool, which can predict the success of movie being a hit or flop. As this factor is important for everyone involved in the movie, for example: If a movie is flop, it exacerbates the image of actor or director. The tool will use searching algorithms and then use of bespoke system to predict the percentage of success of movie which is yet to be released. Their analysis of the data collected from various resources like IMDb, Kaggle. The results showed that the proposed approach improved the classification accuracy as compared to a fully independent setting.[9]

Meenakshi et al., (2018) developed a system based upon data mining techniques that may help in predicting the success of a movie in advance thereby reducing certain level of uncertainty. An attempt is made to predict the past as well as the future of movie for the purpose of business certainty or simply a theoretical condition in which decision making the success of the movie is without risk, because the decision maker has all the information about the exact outcome of the decision, before he or she makes the decision.[10]

Quader et al. (2017) proposed a decision support system for movie investment sector using machine learning techniques. This research helps investors associated with this business for avoiding investment risks. The system predicted an approximate success rate of a movie based on its profitability by analyzing historical data from different sources like IMDb, Rotten Tomatoes, Box Office Mojo and Metacritic. They showed Neural Network gives an accuracy of 84.1% for pre-released features and 89.27% for all features while SVM has 83.44% and 88.87% accuracy for pre-released features and all features respectively when one away prediction is considered.[11]

Kenvi Shah et al. (2019) used a Linear Regression can predict the approximate ratings the movie can receive once it is actually released and hence classify a movie as a hit or a flop. A large amount of data representing feature films is maintained by the Internet Movie Database (IMDb). The empirical model demonstrated good statistical results when the dependent variable was binary and interval.[12]

Proposed Methodology: In this section, we are describing the proposed methodology used for the prediction of movie success using machine learning technique. There are various machine leaning technique such as k-nearest neighbor, logistic regression, decision tree, Naïve Bayes classifier etc. Among various machine learning technique our proposed method uses the random forest model for the efficient prediction of movie success which is describing below.

Workflow: Most of the earlier work are predicting the IMDB Score. Means they are using different attribute to predict the IMDB Score hence they taken this as a Regression Problem. But our main focus is to predict the success rate. So, we divide the whole Range of IMDB in five different categories so that we can take it as Classification Problem and hence we can also increase the past scores. (To get a good Score in Regressor requires proper dataset but in classification a good score can easily raise). So, we had classified the movies in a category followed as:

Table 1: Range of IMDb Rating

IMDB Rating	Score (My system of scoring)
0-2	0
2-4	1
4-6	2
6-8	3
8 and so on	4

This is all about the workflow as it's all divided in the following steps which made this easy.

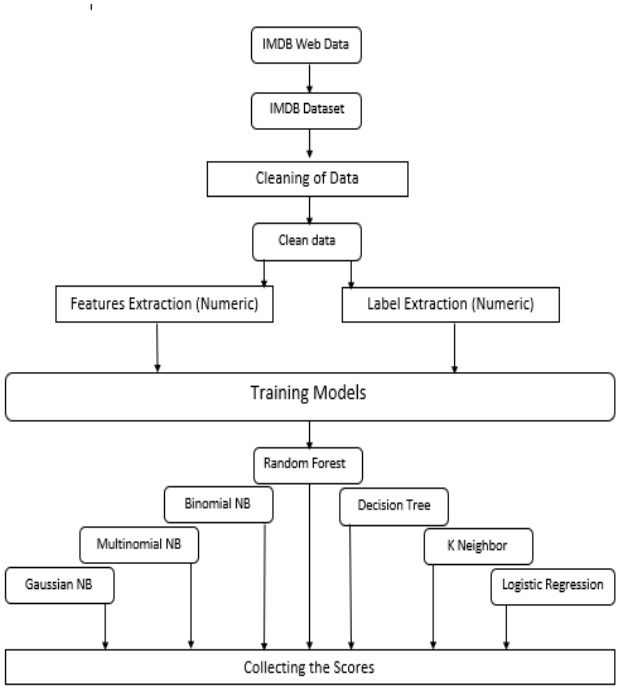

Fig.1: Data flow diagram of Proposed methodology

To predict the movies success rate, our methodology follows the subsequent steps which are discussing below:

Step1: Collecting database

In the first step we need dataset to work upon. As we are familiar with a website named as Kaggle, which is the best place for all kinds of datasets. So, talking about our aim we need a dataset which comprises of every single detail about the movie. Hence, we have the official; dataset of IMDB movies which is collection of every detail of around 4000 +n movies.

Discussing about the attributes or the details of each movie, as we have in following attributes:

Table 2: Attributes of Movie dataset

Table 2: Attributes of Movie dataset

color	director_name	num_critic_for_reviews	duration
director_facebook_likes	actor_3_facebook_likes	actor_2_name	actor_1_facebook_likes
gross	genres	actor_1_name	movie_title
num_voted_users	cast_total_facebook_likes	actor_3_name	facenumber_in_poster
plot_keywords	movie_imdb_link	num_user_for_reviews	language
country	content_rating	budget	title_year
actor_2_facebook_likes	imdb_score	aspect_ratio	movie_facebook_likes

So, with the help of this dataset we will try to implement a Classification model which can easily predict the Movie success rate as 0 with least and so on**.**

Step2: Cleaning Database

In this step we already had collected the database so this is the time where we need to filter or clean the dataset. So in this we take Care of few things as following.

As cleaning the data is first task in ML & DS workflow. Without this we will face many issues in exploring the required terms. As cleaning just not actually means to clean the data. It exactly means filtering and modifying your data such that it is easy to explore, understand and model.

1. First task is to remove all the Nan or Empty values.
2. Then we need to handle the missing value.
3. Handling the Outliers

So, in this process we clean the data and make it ready for the Training purpose.

Step3: Picking Features (Necessary)

In this step, we need to the features or we can say that we need to select the columns which we need to feed in the following model.

As in given dataset we have three types of data

1. Numerical data
2. Categorical Data
3. Composite data

As all the classifiers are best suited for the Numerical Values so we will be going to select all the numerical columns for the training purpose.

Step4: Training Different Models

From the last step we have training dataset and resulting IMDB scores now we need train different model or we can say that we need to train different classifiers so that prediction can be taken out. In this we are using following Models.

1. Random Forest Regressor
2. Logistic regression
3. Decision tree
4. K Neighbors Classifier
5. Gaussian Naïve Bayes
6. Multinomial Naïve Bayes
7. Binomial Naïve Bayes

Here we have used a few of best classifiers.

Step5: Printing all the Scores

In this step we printed all the scores of all used classifiers hence get to know that Random Forest is the best regressor.

Step6: Result

As a result, we final implemented a classifier which can predict the Success rate of any IMDB movie.

Result Analysis: In this section of the research work, we perform the result analysis on different measuring parameters like score, accuracy, precision, recall, f1-measure, mean absolute error and mean square error and comparison is done between the proposed methodology (random forest classifier), logistics regression and K neighbors' classifiers.

Comparison of Score: The score parameter is used prove the rating score to the movie and the comparative analysis of this parameter is done among different machine learning such as GuassianNB, MultinomialNB, BernoulliNB, KNeighnors Classifier, Decision Tree, Logistic regression and our proposed method (random forest). The simulation results of our proposed method and existing method is shown in table 3 and it is 70% which is much more about the other exiting approach. The analysis is done using the comparison graph shown in figure 2 and it is found that our proposed method has higher value than the others. It means that the proposed method is more success in the prediction of movie success or hit.

Table 3: Comparative analysis of score parameter between Random forest and existing method

Comparison of Score			
S. No.	Name of Classifier	Short Name	Score
1	GaussianNB	GNB	0.32
2	MultinomialNB	MNB	0.12
3	BernoulliNB	BNB	0.66
4	KNeighborsClassifier	KNS Model	0.61
5	Random Forest (Proposed Method)	RF	0.7
6	Decision Tree	DT	0.61
7	Logistic regression	LR	0.65

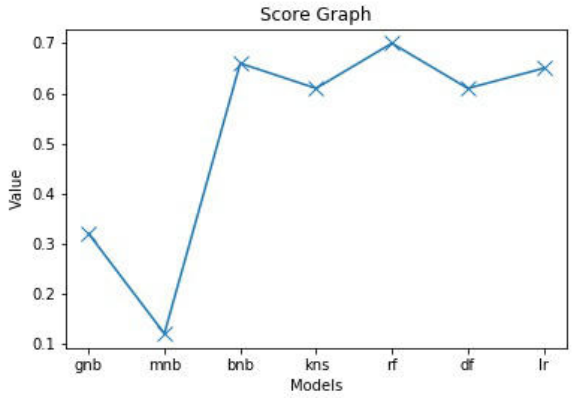

Fig. 2: Comparison of Score parameters

Comparison of Accuracy: This section presents the comparison of accuracy parameter to show the prediction accuracy of movie and the comparative analysis of this parameter is done among different machine learning such as GuassianNB, MultinomialNB, BernoulliNB, KNeighnors Classifier, Decision Tree, Logistic regression and our proposed method (random forest). The simulation results of our proposed method and existing method is shown in table 4 and it is 70% which is much more about the other exiting approach. The analysis is done using the comparison graph shown in figure 3 and it is found that our proposed method has higher value than the others. In this the value of accuracy is equivalent to score parameter. If score of the movie will high accuracy of the movie prediction will high. And it is analyzed that the proposed method is more success in the prediction of movie success or hit.

Table 4: Comparative analysis of accuracy parameter between Random forest and existing method

Comparison of Accuracy			
S. No.	Name of Classifier	Short Name	Accuracy
1	GaussianNB	GNB	0.32
2	MultinomialNB	MNB	0.12
3	BernoulliNB	BNB	0.66
4	KNeighborsClassifier	KNS Model	0.61
5	Random Forest (Proposed Method)	RF	0.7
6	Decision Tree	DT	0.61
7	Logistic regression	LR	0.65

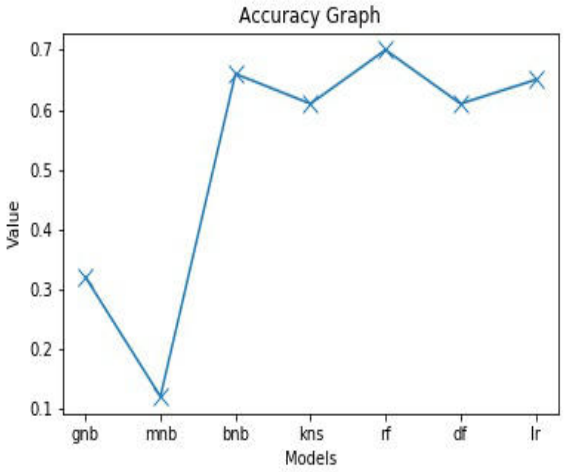

Fig.3: Comparison of accuracy parameters

Comparison of Precision Score: This section presents the comparison of precision score parameter to show the prediction accuracy of movie and the comparative analysis of this parameter is done among different machine learning such as GuassianNB, MultinomialNB, BernoulliNB, KNeighnors Classifier, Decision Tree, Logistic regression and our proposed method (random forest). The simulation results of our proposed method and existing method is shown in table 5 and it is 66% which is much more about the other exiting approach. The analysis of precision parameter is done using the comparison graph shown in figure 4 and it is found that our proposed method has higher value than the others. Due to the higher precision value it is analyzed that the proposed method is more success in the prediction of movie success or hit.

Table 5: Comparative analysis of Precision parameter between Random forest and existing method

Comparison of Precision Score			
S. No.	Name of Classifier	Short Name	Precision Score
1	GaussianNB	GNB	0.59
2	MultinomialNB	MNB	0.07
3	BernoulliNB	BNB	0.43
4	KNeighborsClassifier	KNS Model	0.56
5	Random Forest (Proposed Method)	RF	0.66
6	Decision Tree	DT	0.61
7	Logistic regression	LR	0.43

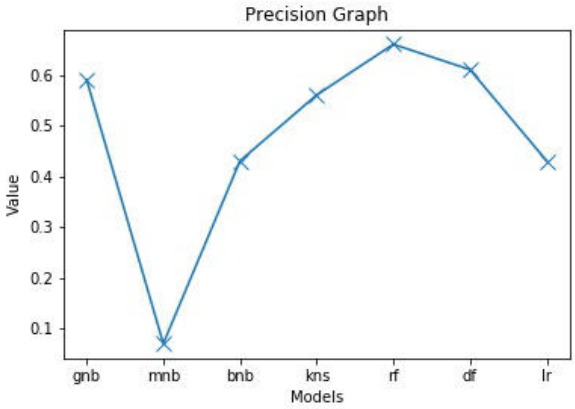

Figure 4: Comparison of Precision Score

Comparison of F1 Score: This section presents the comparison of F1 score parameter to show the prediction accuracy of movie and the comparative analysis of this parameter is done among different machine learning such as GuassianNB, MultinomialNB, BernoulliNB, KNeighnors Classifier, Decision Tree, Logistic regression and our proposed method (random forest). The simulation results of our proposed method and existing method is shown in table 6 and it is 66% which is much more about the other exiting approach. The analysis of F1 score parameter is done using the comparison graph shown in figure 5 and it is found that our proposed method has higher value than the others. Due to the higher precision value it is

analyzed that the proposed method is more success in the prediction of movie success or hit.

Table 6: Comparative analysis of F1 score parameter between Random forest and existing method

Comparison of F1 Score			
S. No.	Name of Classifier	Short Name	F1 Score
1	GaussianNB	GNB	0.23
2	MultinomialNB	MNB	0.09
3	BernoulliNB	BNB	0.52
4	KNeighborsClassifier	KNS Model	0.57
5	Random Forest (Proposed Method)	RF	0.66
6	Decision Tree	DT	0.61
7	Logistic regression	LR	0.52

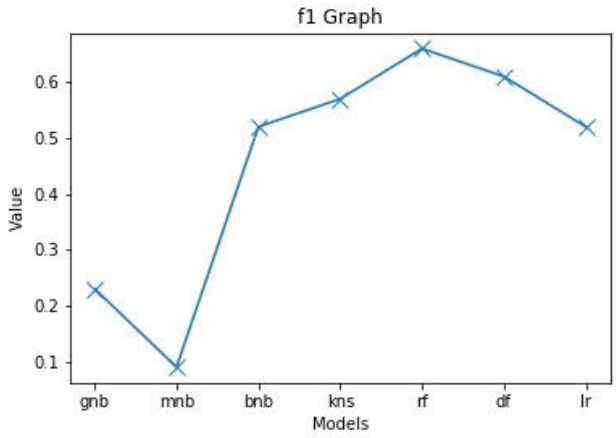

Figure 5 : Comparison of recall value parameters

Overall Comparison: The overall comparative analysis of proposed method is done with existing method using the measuring parameters precision, recall, F1 and accuracy score. The simulation results of these parameters are shown is table 7 and analysis is shown through figure 6. The value of different parameters of our proposed method is extremely higher than the already existing method which is about 8%. After overall analysis it is found that the proposed approach for the success rate prediction of movie is much better than the existing method.

Table 7: Overall comparative analysis of parameter between Random forest and existing method

Overall Comparison					
S. No.	Method	Precision	Recall	F1	Accuracy
1	Proposed Method	0.66	0.66	.69	.70
2	Existing Method	0.61	0.60	.58	.62

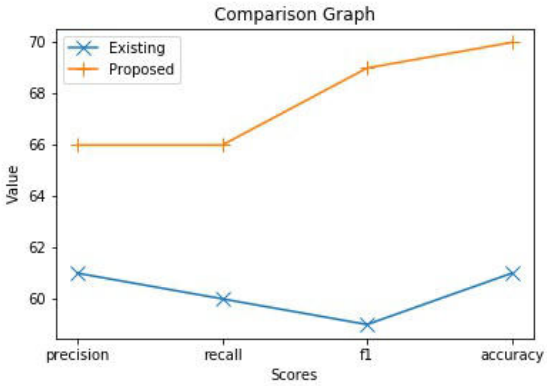

Figure 6: Overall analysis of Proposed and Existing method

Conclusion & Future Work: The success of movie not only depend on the features or attributes related to movie. The number of audiences also plays a significant role for a movie to become hit or successful. In a specific year how many no. of tickets sold indicates the number of viewers in that year. From the experimental results, we found that the is it difficult to apply any other technique of data mining to the IMDb dataset. These datasets require proper cleaning and integration which consume lots of time in the analysis. In addition, most of the data of these dataset is in textual and numerical form which makes the extraction of data difficult. Due to this source data cannot be integrated easily. So, for this we apply machine learning technique which can integrated these datasets properly. In this dissertation, we use machine learning technique for our experimentation which have powerful classification like GuassianNB, MultinomialNB, BernoulliNB, KNeighnorsClassifier, Decision Tree, Logistic regression and random forest for

classification and regression. The experimental analysis of proposed method and existing method is done using the performance measuring parameters like precision, recall, F1 score and accuracy, etc. For the simulation of proposed method and existing method Python language uses which is easy implement and consume less computation time than other language. The result generated for proposed method after simulation for the accuracy and score parameter is 70% which is much more than the existing method. Similarly, the analysis of proposed and existing method is done using precision and recall parameter and the value of proposed methods is 66% which is also more than the existing method. Later the analysis of proposed and exiting method is done using F1 score and the value of F1 score of is 69% and 58% which is about 11% more than the existing method. Based on these parameters we can easily predict the success rate of movies. In future work, it is necessary to add these factors also which can improve the prediction rate of movie success. And also use the hybrid technique of machine learning by using the essential features of random forest and some other technique like logistic regression which will improve the accuracy for successful of movie prediction.

References

[1] B. R. Litman & H. Ahn. (1998). Predicting financial success of motion pictures. In B. R.

[2] Litman (Ed.), the motion picture mega-industry. Boston, MA: Allyn & Bacon Publishing, Inc.

[3] Meenakshi *et al.,* "A Data mining Technique for Analyzing and Predicting the success of Movie", National Conference on Mathematical Techniques and its Applications, 2018. Journal of Physics: Conf. Series 1000 (2018) 012100, doi: 10.1088/1742-6596/1000/1/012100.

[4] [Online]. Available: https://www.edureka.co/blog/what-is-machine-learning/

[5] Bei-Bei CUI, "Design and Implementation of Movie Recommendation System Based on Knn Collaborative Filtering Algorithm", ITM Web of Conference, 2017. DOI: 10.1051/itmconf/2017 1204008

[6] B. C. . U. P.E.tg off, "Multivariate decision trees: machine learning," no. 19, 1995, pp. 45–47.

[7] Aashya Khanduja, "Logistic Regression for Predicting Movie's Success", ICML 2018.

[8] Hemant Kumar, Santosh Kumar, "Predicting Movie Success or Failure using Linear Regression & SVM over Map-Reduce in Hadoop", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 6, Issue 6, June 2018, pp. 6426-6433.

[9] Chaudhari *et al.*, "A Data Mining Approach to Language Success Prediction of A Feature Film", International Journal of Engineering Sciences & Management Research, 2016, pp. 1-9.

[10] Meenakshi *et al.,* "A Data mining Technique for Analyzing and Predicting the success of Movie", Journal of Physics: Conf. Series 1000 (2018) 012100 doi: 10.1088/1742-6596/1000/1/012100. Pp 1-9.

[11] Quader *et al.,* "A Machine Learning Approach to Predict Movie Box-Office Success", 20th International Conference of Computer and Information Technology (ICCIT), 22-24 December, 2017.

[12] Kenvi Shah *et al.,* "Movie Success Prediction using Data Mining and Social Media", International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue: 03, Mar 2019.