Movie Success Prediction Using Popularity Factor from Social Media

Dipak Gaikar[1], Riddhi Solanki[2], Harshada Shinde[3], Pooja Phapale[4], Ishan Pandey[5]

[1] *Asst. Professor, Dept. of Computer Engineering, Rajiv Gandhi Institute of Technology, Maharashtra, India*
[2] *B.E. student, Dept. of Computer Engineering, Rajiv Gandhi Institute of Technology, Maharashtra, India*
[3] *B.E. student, Dept. of Computer Engineering, Rajiv Gandhi Institute of Technology, Maharashtra, India*
[4] *B.E. student, Dept. of Computer Engineering, Rajiv Gandhi Institute of Technology, Maharashtra, India*
[5] *B.E. student, Dept. of Computer Engineering, Rajiv Gandhi Institute of Technology, Maharashtra, India*

---***---

Abstract - *Hollywood and Bollywood film industries have reached staggering heights in terms of volume of business, movies produced and it's reach. With so much at stake, it is of commercial interest to develop a model to predict the success of the movie. In this paper, we have developed a mathematical model for predicting the rating and success classes such as hit, flop and neutral of the movies. In order to do this we have used a machine learning and data mining algorithm. The algorithm used for classification is k-NN. Popularity factor of various movie parameters like actor, actress, director, writer, budget etc. is collected which helps in the movie success prediction. This project helps the director or producer of the movie to pre-decide the parameters such as actor, actress etc. of the movie. This project also helps the user to decide whether to book ticket in advance or not based on upcoming movie prediction.*

Key Words: Prediction, Data Mining, Machine Learning, k-NN, IMDB, Movie, Facebook, Twitter, Social Media, Followers.

1. INTRODUCTION

The Indian movie industry produces the maximum number of movies at 1000 per year which is higher than any other country's movie industry. However, movies that taste success and are ranked high are very few. Given the low success rate, models and mechanisms to predict success and ratings of a movie can help de-risk the movie-making businesses significantly and increase the average returns. Various stakeholders such as actors, producer, directors etc. can use these models to make more informed decisions. Various social media websites are excellent resources to find the popularity factor about any component of a movie. Data mining techniques enable us to uncover information which will both confirm or disprove common assumptions about movies, and also allow us to predict the success of a future movie based on the historical data of its components before its release. The paper organized as follows : Section II describes related work, Section III describes the role of data gathering, preprocessing and data mining. Mathematical models are discussed in section IV. Interpretation is shown in section V & Section VI concludes the paper. General Design is shown in Figure 1.

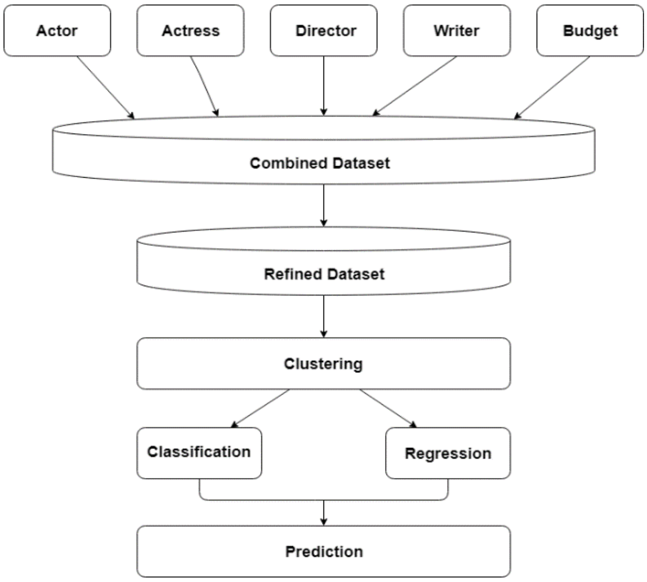

Fig -1: General Design

2. OBJECTIVES

- To obtain the Hollywood and Bollywood datasets of different parameters used in our project.
- To perform an efficient pre-processing task to make data suitable for analysis.
- To predict the rating of movies and determine whether it is hit, flop or neutral.
- To help the Director pre-decide the components/parameters of the movie.
- To help the User to decide whether to book the tickets of an upcoming movie in advance or not.

3. RELATED WORK

Knowing the movie rating prior to the release of a movie is an essential requirement of the movie makers, a project performing movie prediction needs to incorporate certain methodologies and algorithms. We have studied many papers based on the same.

In 2014 Nitin, Pranav, Sarath and Lijiya [1] collected their datasets of movies ranging from year 2000 to 2012 through IMDB (Internet Movie Database). They did a comparison between 3 regression models: linear, logistic and Support Vector Machine regression model, and found linear regression to be most suitable one with the accuracy of 51%.

In 2016 Harsh, Anupam and Vineet [2] in their paper work developed a mathematical model of revenue and rating. They studied contrast between linear, non-linear, logistic regression and neural network toolbox. Linear regression came out to be working well for medium and low revenue earning movies. Non-linear was accurate for some values but also deviated for others. Logistic regression gave good prediction for limited range of movies.

In 2017, Sanjai, Abhisht and Geetha [3] talked about the method of classification and fuzzy logic to predict and classify movie to be successful or not. They used normalisation formula, confusion matrix in their work. Their accuracy came out to be around 85%.

In 2015, Parag and Omkar [4] too stressed upon the use of linear and logistic regression in their work. Their project was implemented with 70.4% accuracy.

4. PROPOSED MODEL

This project includes 3 major components viz. Data Gathering, Data Pre-processing and Data Mining. Preparing a dataset by collecting information by referring to social media sites and further refining it by eliminating repeated attributes, we get a dataset that can be used to predict success of a movie. We further use the k-Nearest Neighbor algorithm which is used to cluster data and predict the success class of a movie such as Hit, Flop or Neutral.

4.1 Dataset Description :

The dataset consists of numerous amount of entries which are based on popularity factor of our 5 input parameters. These popularity factors are collected by considering the number of followers of each input parameters from social media platforms. Each row containing 5 input parameters corresponds to a particular movie and the collection of these movies are listed according to the year of their release.

4.2 Data Gathering

The original data is gathered by referring to the number of followers of actors, actresses, directors, and writers on social media sites like Facebook, Twitter, Instagram. We have also gathered IMDB ratings of past movies from 1998 to 2018. Both Hollywood and Bollywood movies are considered. This dataset is updated for every parameter mentioned above and it includes numerical values.

4.3 Pre-processing

The data that we obtained was inconsistent due to its huge size and its origin from multiple heterogeneous sources. To overcome this inconsistency, we applied pre-processing techniques such as data cleaning and data refining before applying data mining techniques.

4.4 Data Mining

The data mining algorithm which we have used for classification of the success class of the movie into Hit, Flop or neutral is k-NN algorithm.

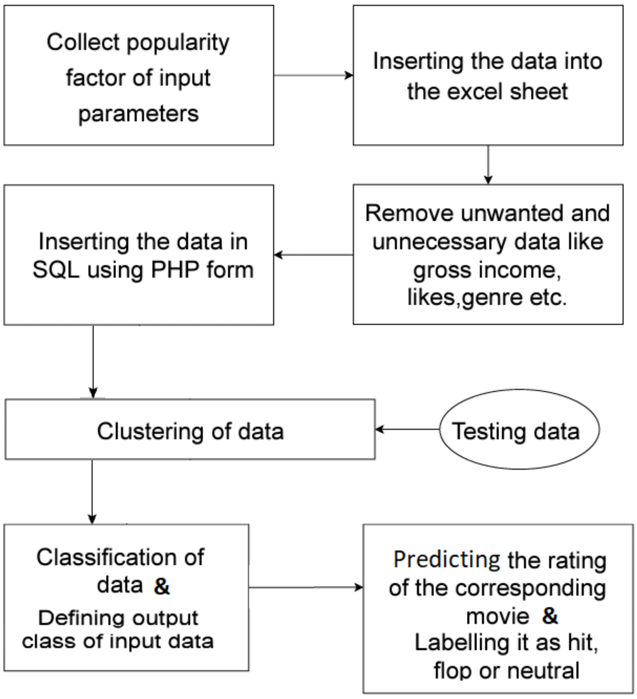
Collect popularity factor of input parameters
Inserting the data into the excel sheet
Remove unwanted and unnecessary data like gross income, likes,genre etc.
Inserting the data in SQL using PHP form
Clustering of data
Testing data
Classification of data &
Defining output class of input data
Predicting the rating of the corresponding movie &
Labelling it as hit, flop or neutral

Fig -2: Proposed model

5. METHODOLOGY

5.1 Overview

This project follows a mathematical model of k-NN algorithm which is used for classification. It searches for "k" nearest neighbors. It performs classification by classifying a case by a majority vote of its neighbours. The case is assigned to a class most common amongst its k nearest neighbours measured using a distance function.

5.2 Mathematical Model

Here we will be depicting and explaining our project in terms of it's mathematical analysis. We will show the formulas used here.

Distance Formula

Various distance calculating methods of neighbors are Euclidean, Manhattan and Minkowski . All these three methods are used in case of continuous variables but in case of categorical variables, Hamming distance must be used. Out of all the methods listed above, we will be using Euclidean distance calculating method in our project.

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad \text{..............................} \quad \text{(i)}$$

5.3 Choosing the k Factor

Data is inspected for choosing the most optimal value for k. In general, a large value of k is selected as it reduces overall noise but nothing is proven. For the validation of k value, Cross-validation is another way to determine a good k value by using an independent dataset retrospectively. The most optimal value for k has ranged from 3-10 historically since it produces better results than 1NN.

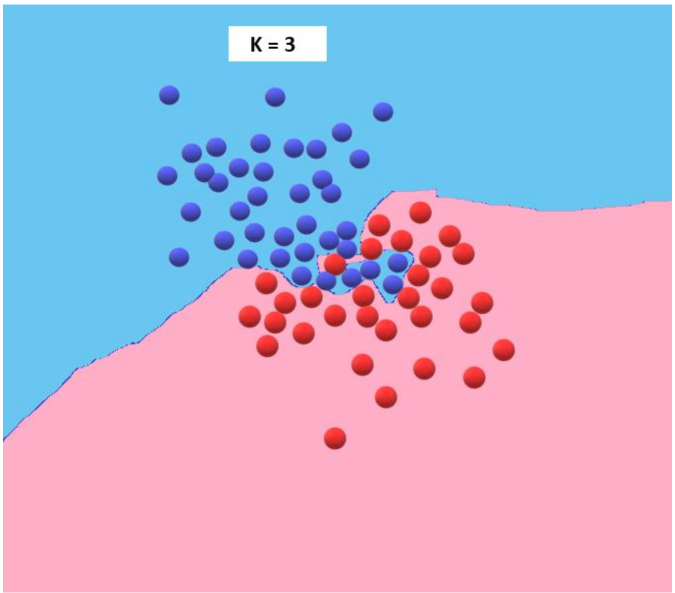

Fig -3: Clustering using value of k=3

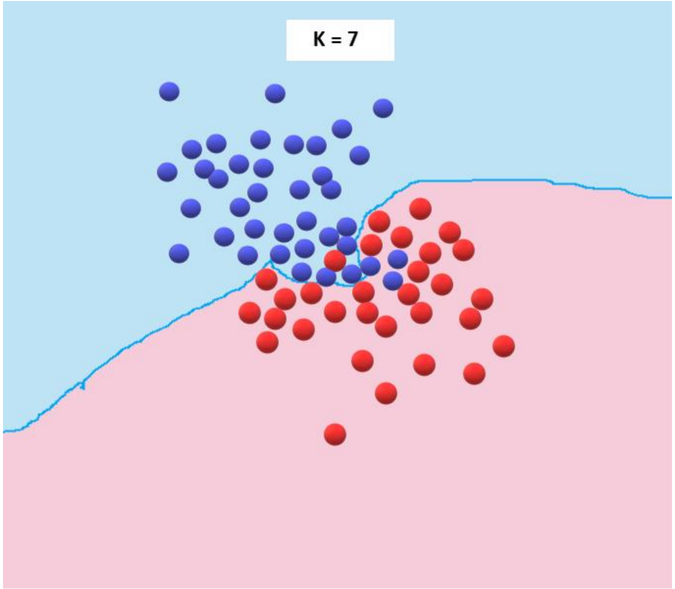

Fig -4: Clustering using value of k=7

In the above shown Figures 1 and 2 respectively, it is seen that the boundary of class becomes smoother with the increasing value of k.

5.1 k-NN (K-Nearest Neighbor)

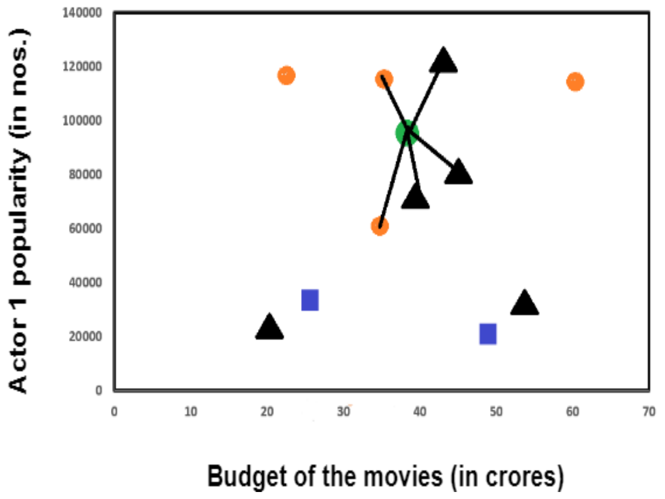

Fig -5: k-NN algorithm

Consider the budget of the movie on the X-axis and actor 1 popularity i.e. the number of followers on Y-axis. We plot the points according to our dataset and then plot a test case and determine the value of "k" i.e. 5 for our case. The k-NN algorithm then chooses five nearest neighbors. We have considered 3 classes i.e. hit, neutral and flop for prediction which are denoted by triangles, circles and squares respectively. Out of the 5 nearest points chosen, 3 have class "hit" and 2 have class "neutral", hence we go with the majority class i.e. the hit class and the predicted class for the respective movie will be hit.

For the purpose of explanation, we have considered a 2-D scale, but in our dataset, we have considered 5 components and hence our actual prediction will be on a 5-D scale.

5.4 Why k-NN

k-NN can be implemented for both classification and regression predictive problems. But, it is used on a large scale in classification problems. It has the following characteristics :

- Ease to implement
- Time Complexity
- Predictive Power

Table -1: Comparison of algorithms

	Logistic Regression	CART	Random Forest	KNN
Ease to implement	2	3	1	3
Time complexity	3	2	1	3
Predictive power	2	2	3	2

6. RESULTS

In this project, we have shown how popularity factor of movie components can be used to predict the success of upcoming movies. For this prediction, we have used various input parameters in our project such as Actor1, Actor2, Director, Writer and Budget. We used popularity factor of each parameter using social media . We know that k-NN is a non-parametric method used for classification, also the datasets of Hollywood and Bollywood used don't have predefined assumption or constraint on their datasets. For the same we found k-NN to be appropriate algorithm to be used for our movie success prediction. The results obtained from our project were observed to be at close proximity to IMDB ratings.

Consider a scenario where we have obtained results of prediction for 10 movies. We compare these ratings with IMDB ratings and obtain the matching accuracy percent. The table 2 shows the predicted ratings, IMDB ratings and matching accuracy percent.

Table-2 : Comparison of predicted ratings and IMDB ratings of movies

Movie no.	Predicted Ratings	IMDB Ratings	Accuracy
1	4.3	5.9	72%
2	5.8	7.2	80.55%
3	6.6	6.9	96%
4	6.1	7.6	80%
5	5.7	7.8	73%
6	5.9	7.2	81.94%
7	5.3	7.2	73.61%
8	7.1	7.7	92.20%
9	4.7	6.8	69.11%
10	5.5	6.9	79.71%

The predicted movie ratings and IMDB ratings are compared and plotted using bar graph as shown in figure 6 and 7.

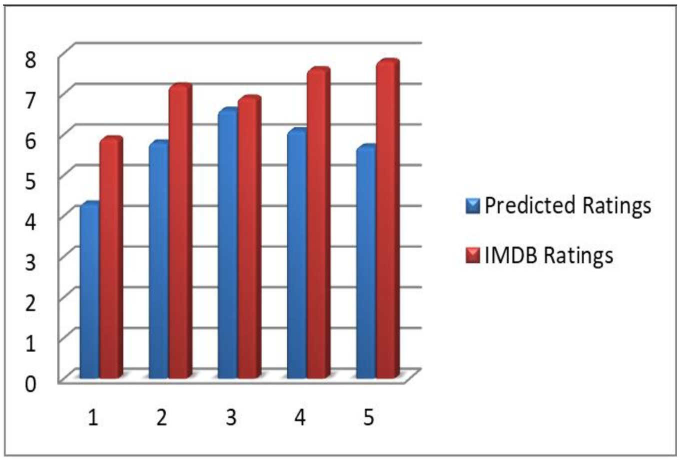

Fig -6: Comparison of results of first 5 movies

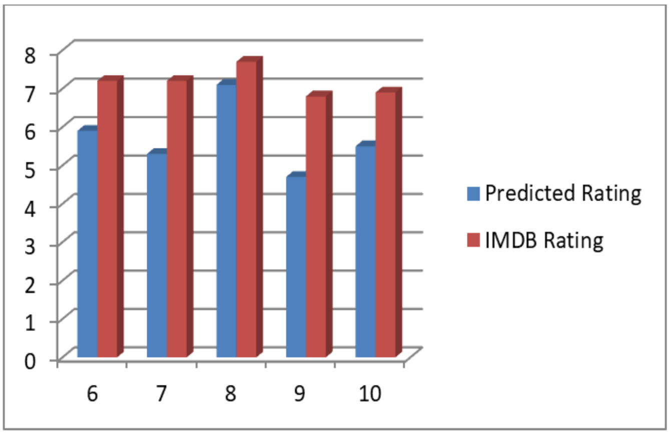

Fig -7: Comparison of results of next 5 movies

3. CONCLUSIONS AND FUTURE SCOPE

In order to predict the success of movies we have used k-NN, a machine learning algorithm. In the data set, the attributes that contributed the most to the information are number of followers on Facebook and IMDB score. This project can be used in online platforms too. Additional features must also be considered to achieve this such as news analysis, movie story analysis and social network sentiment analysis and the information thus obtained could be added to the data set. We can also obtain input from the audience which can be added to the data set through Google forms to improve the result. We can also extend and incorporate our project into a mobile app for future use.

In today's time, we are witnessing lots of digital channels like Netflix, Amazon Prime, Hotstar, ALTBalaji and many more which are showing lots of interesting contents to its users and keeping in mind this turning point in the movie industry we can incorporate our project in a successful prediction of web series and other digital contents. So the users can easily decide to watch which content and stake holders can confidently finance into the correct movie project.

ACKNOWLEDGEMENT

We sincerely thank our project guide Prof. Dipak Gaikar for his interminable support and guidance in carrying out the research work, and also our institute RGIT for providing us with the grants and library material.

REFERENCES

[1] Nithin, V. R., M. Pranav, and P. B. Sarath Babu. "Lijiya: A predicting movie success based on IMDB data." *Int. J. Data Min. Tech* 3 (2014): 365-368.

[2] Harsh Taneja, Anupam Dewan, Vineet Bhardhwaj: "Pre-Release Success Quotient Prediction of Movies," International Journal for Science of Science and Research (IJSR), Vol.05, Issue 10, 2016.

[3] Pramod, Sanjai, Abhisht Joshi, and A. G. Mary. "Prediction of movie success for real world movie dataset." *Int. J. of Advance Res., Ideas and Innovations in Technol* 3, no. 3 (2017).

[4] Parag Ahivale , Omkar Acharya: "Success Prediction of Films at Box Office Using Machine Learning", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 3 Issue IV, April 2015

[5] Meenakshi, K., G. Maragatham, Neha Agarwal, and Ishitha Ghosh. "A Data mining Technique for Analyzing and Predicting the success of Movie." In *Journal of Physics: Conference Series*, vol. 1000, no. 1, p. 012100. IOP Publishing, 2018.

[6] Ahmad, Javaria, Prakash Duraisamy, Amr Yousef, and Bill Buckles. "Movie success prediction using data mining." In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1-4. IEEE, 2017 .

[7] Dipak Damodar Gaikar, Bijith Marakarkandy, and Chandan Dasgupta. "Using Twitter data to predict the performance of Bollywood movies." *Industrial Management & Data Systems* 115, no. 9 (2015): 1604-1621.

[8] Krushikanth R. Apala, Merin Jose, Supreme Motnam, C-C. Chan, Kathy J. Liszka, and Federico de Gregorio. "Prediction of movies box office performance using social media." In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 1209-1214. ACM, 2013.

[9] Babita M. Jangid , Chaitali K. Jadhav , Swati M. Dhokate, Grish M. Jadhav, Prof. G.M. Bhandari: "Survey on Movies Popularity Prediction Project Using Social Media Feature", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 9, September 2016

[10] Merim Babu, B. Muni Archana. "Prediction of Movie Success through the Data Mining" In International Journal of Innovative Research in Technology, Volume 4, Issue 11, April 2018

[11] https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/