ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 8, Issue 8, February 2019

Comparative Analysis of Prediction Models of MOVIE SUCCESS Rate

Ansari Sana Fatima, Shruti Pimple
Department of Information Technology, Sardar Patel Institute of Technology, Maharashtra, India

Abstract: ***Cinemas in today's world are the most popular means of entertainment. Millions of people watch movies all over the world not only for the means of entertainment but also to get stress free and escape from the anxiety and troubles of life. Usually in the movie Industry there is a lot of investment therefore making predictions is on the map. What makes a movie successful? What different criteria can let a movie enter into the list of the top grossing films? This different alternative questions to a mind before making our investment on any film. Thus prior knowledge is required for such predictions whether they will be hit or failure. In this project our aim is to develop a model that predicts the success rate of the movies whether it i s a hit, flop or super hit depending on different parameters, whereas budget is an important parameter Depending on these parameters the success rate of the movie is predicted. Taking different factors into consideration the success level is being vaccinated. These factors in our project can be considered by the movie maker to decide their financial roadmap and also evaluating their comfort zone on taking risk.***

Keywords: **Machine-Learning, Bollywood, Movies, Revenue, Prediction-Model.**

I. INTRODUCTION

"Bollywood," an Indian cinema with its distinctive beauty, flavor and magic, has become a key medium for mass communication ever since it was created. Both entertainments are paired with the exchange of ideas. A typical Indian film contains all of the spices and life. Due to the fact that films such as producers and financiers have different points of view, it is a paid industry and an easy source of income for actors. Similarly in India there are numerous streams of film including comedy films and make money, than the parallel cinema which aims on sensitizing people on different social issues like this the mentality or mindsets of the people which influences them to know that the movie will be hit or gets flop.

We strive to build in our project a predictive model that will give the film success rate. We consider the parameters of star cast, genre, director, and budget. We use various algorithms for the project KNN, Linear Regression, Naive Bayes, Support Vector Machine (SVM) Linear model and logistic regression, in order to predict the film's film set.

For eg, if the film was made with a budget of 50 crores and 40 crores were received in Indian box offices, it would be like losing the film even if it won worldwide.

The prediction made is categorized into three flops, hit and super hit. Normally the net income defines the terms whether the movie is hit, flop, super hit.

Flop: There are a lot of points depending on which flop is declared. Using different parameters gives different results. Many points are calculated for the same such as box office collection, net incomes. If the movie was made on the budget of 50 crores and if the movie has not earned maximum revenue on Indian box office then it will predict losing the movie.

Hit: If the movie earns the profit 20 percent more than the budget than the movie predicted hit

Super Hit: For the movie to be a super hit budget plays a vital role. If the movie earn 50% more than the budget than the movie predicted as the super hit

II. LITERATURE SURVEY

There are many other papers which worked on the algorithms and gave the prediction on various parameters and gave different accuracy. In our Paper we attempt to work with different algorithms contemplating some of these factors like Budget, star cast, genre, director etc using prediction analysis.

In this research paper machine learning technique SVM ,NLP and neural network is used for the movie prediction based on some released features .Prediction is calculated in two ways one is exact match and other is one way prediction The prediction is done based on the different parameters like Rotten tomatoes , IMDb votes ,number of screen ,budget ,box office Mojo .[1]

In this research paper mathematical order in which movie genres was one factor was developed to predict the movie failure and success rate. Criteria by which success rate is predicted includes cast, director, shoot location, songs, writer, movie's release date and target audience. Each criterion here is given a weight based on which movie success rate is predicted. Data mining techniques are used in this paper. Because of the data mining technique used the paper has less chance of failure this paper has defined the above parameters on the basis of which the success rate is predicted. [2]

In this paper for the better performance they have inquired about different techniques for prediction. From the transmedia storytelling the new factor is added. They have used an ensemble approach for predicting the movie's performance. Cinema Ensemble Model (CEM) is used for prediction from previous research papers. Twitter data and movie sales using a technique called web blog data were used for the prediction of the movie's success. The

Manuscript received: 24 January 2019
Manuscript received in revised form: 19 February 2019
Manuscript accepted: 07 March 2019
Manuscript Available online: 10 March 2019

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 8, Issue 8, February 2019

performance of different machine learning technique were examined used logistic regression, decision tree and neural network to predict the movie success. [3]

In this paper they used machine learning tools for the prediction of the movie before its release. Data is evaluated from different parameters Box office India, wogma, cinemalytics and youtube. Songs are an important part of Bollywood movies so they have designed the music score factor which will help in increasing the accuracy for the movie success prediction that is classified into hit and flop classes. Using a bagging algorithm they have created the further model. [4]

In this paper the data is extracted from the social media the paper says that the content on the social media is somewhere correlated with the box office collection. So using Linear Regression and Support Vector Regression algorithms they have done the box office prediction. They have also used linear and nonlinear regression which depends on the popularity the particular movie gained based on comments and posts of the users. [5]

This study talks about the social media platform which has been used as a factor to evaluate the accuracy of the success of the movie. It tells how more than 1000 movies that are released per year become difficult to predict its success rate. So using different parameters like directors, cast, producer etc they consider these parameters also including the box office collection and social platform to forecast the success rate of the movie. [6]

III. PROPOSED WORK

Our goal is to classify the films as flop, hit or super hit based on the return on the investment of the film made by the domestic box office. .Predictions are made directly after the film is released and are more precise. We use algorithms for machine learning such as linear regression, logistic regression, naive multi-class bays, SVM and k-mean to predict the problem.

Web Scraping

The proposed work of web scraping is to exact the movie detail from imdb (international movie database), omdb (open movie database Api website, and for scrapping the data we will use a python library named Beautiful soup . We have exact 2500 movies details for the dataset.

We explain revenue forecast in this paper with supervised learning as a disclosed estimate. We have a sequence of training sessions (X(1),Y(1),X(2),Y(2)....X(n),Y(n); (i)

The input vector refers to a film and y(i) belongs to R6, a categorical dependent variable that is equivalent to six groups of desirable income. In different ways we can model x I and y I and their relationships.

We used 5 different models to classify the contribution into revenue categories for the research production of this project, which is a consequence of which we have approximated the set of domestic film box offices provided by the vectors.

Cross-validation as a statistic for budget and box office collection, as they are somewhat different on many web pages because we have in the final data set over 65% of films with a less domestic collection than their own budget. Dataset was obtained from various sources with web scraping software. Our assumption of this problem of learning used a large space than the sum of the available sample. This was the question of the dimensionality to solve this issue, since we wanted to project the data into a larger dimension space. If the total number of functionality is high and non-linear, the output will not improve, but it will be good enough if we use a linear kernel.

IV. RESULT

```
## 70-30 Split randomized
split = int(0.7*data.shape[0])
np.random.shuffle(data)
X_train = data[:split, 1:]
Y_train = data[:split, 0]

X_test = data[split:, 1:]
Y_test = data[split:, 0]

print (X_train.shape, Y_train.shape)
print (X_test.shape, Y_test.shape)
```

```
(690, 3) (690,)
(296, 3) (296,)
```

Training Model

```
[ ] lm = LinearRegression()
    lm.fit(X_train2, Y_train2)

LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)

[ ] lm.score(X_test2, Y_test2)

0.8200896795243457

[ ] y_pred2 = lm.predict(X_test2)
    print(y_pred2.shape)
    for i in range(y_pred2.shape[0]):
        print(y_pred2[i], Y_test2[i], X_test2[i][1])

(57,)
78.26305740546934 63.0 50.0
62.99339942552939 52.0 30.0
113.73117714501973 105.0 50.0
91.68842802513223 153.0 25.0
21.36909182991019 29.5 8.0
37.624510565784425 32.5 20.0
100.18499486512462 107.0 40.0
90.70266726919017 76.0 33.0
20.014473981920713 30.0 7.0
193.284795124345 368.0 100.0
95.13537942522203 92.0 45.0
107.94384676100617 149.8 37.0
181.46200854449492 201.0 100.0
12.255621706039108 27.75 10.0
156.09319968474983 131.0 90.0
325.4204781834378 328.0 215.0
100.18499486512462 100.0 40.0
21.737949301965614 28.0 17.0
44.397601705731745 34.0 25.0
12.255621706039108 29.0 10.0
19.028712845988666 29.75 15.0
112.00770144497471 134.9 40.0
106.95808600507212 107.0 45.0
113.73117714501973 106.5 50.0
```

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 8, Issue 8, February 2019

Model fitting

Table 1.Comparative analysis

Model	Accuracy	Recommendation
linear regression	0.82	absolutely good for prediction
Svm linear model	0.52	we can use it for prediction
Naive Bayes	0.43	we can prefer this only when prediction are hit or super hit
kNN	0.51	it gives all the flop prediction dont use
logistic regression	0.55	it is little better than knn but don't use it for prediction

V. CONCLUSION AND FUTURE SCOPE

Prediction of movie success basically depends on many parameters, In our paper we have used some important parameters for accuracy and success prediction, besides this success also depends on some other factors i.e. Connection with the audience, Different Concept, Level of impact and many other we have excluded these parameters. We have done web scraping along with parameters like star cast, genre, year, and budget. Using these parameters accuracy is evaluated using KNN, Linear Regression, Naive Bayes, Logistic Regression and predicted whether the movie will be Flop, Hit and Super Hit. We conclude that in this paper a linear model is giving more accuracy than other models for prediction. In today's generation almost every youth has their own account on each social media platform. Frequently these resources are used for getting updated the information can be the cricket score, or stock market, or about the launch of a new product, etc. getting influenced by these resources taking the star cast, directors, budget etc as parameters will help finding accurate prediction of the movie's success. The prediction will be done using different algorithms and then the accuracy will be analyzed. We are going to predict only Bollywood movies. We can further incorporate our implementation for predicting the success rate of web series, and media.

REFERENCES

[1] Nahid Quader, Md Osman Gani, Dipankar Chaki, Md Haider Ali " A machine learning approach to predict movie box-office success" 2017 20th International Conference of Computer and Information Technology (ICCIT), Bangladesh, pp:1-7, 22-24 December, 2017.

[2] Javaria Ahmad, Prakash Duraisamy, Amr Yousef, Bill Buckles "Movie Success Prediction Using Data Mining" 2017 8th International Conference On Computing, Communication and Networking Technologies (ICCCNT), India, pp: 1-4, 3-5 July 2017.

[3] Kyuhan Lee, Jinsoo Park, Ijoo Kim, Youngseok Choi "Predicting Movie Success with Machine Learning Techniques: ways to improve accuracy", Information Systems Frontiers, pp: 577–588, 2016.

[4] Sameer Ranjan Jaiswal, Divyansh Sharma "Predicting Success of Bollywood Movies Using Machine Learning", 10th Annual ACM India Computer Conference, India, pp: 121-124, Nov 2017.

[5] Ting Liu, Xiao Ding , Yiheng Chen, Haochen Chen, Haochen Chen, Maosheng Guo "Predicting movie Box-office revenues by exploiting large-scale social media content", Multimedia Tools and Applications , 75(3):pp:1-20,Oct 2014.