International Journal Of Engineering Research And Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 16, Issue 6 (June 2020), PP. 10-14

Movie Success Prediction Using Machine Learning

Ansari Sana Fatima[#1], Shruti Pimple[#2]
[1,2]*Department of Information Technology, Sardar Patel Institute of Technology Email:*

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? These 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 is 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 evaluate their comfort zone on taking risk.*

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

Date of Submission: 20-06-2020 Date of Acceptance: 07-07-2020

I. INTRODUCTION

"Bollywood" an Indian cinema has a charm, flavour and magic of its own. Since its beginning cinema has become a very important platform of mass communication. It Combines both entertainments with communication of ideas. A Typical Indian movie has all the spice and variety of life compressed into it. As there are different views on cinema like the producers and financiers consider it a remunerative business and for actors it is an easy source to earn money. So, similarly there are different streams of cinema in India like comedy movies which entertain and make money, then the parallel cinema which aims on sensitizing people on different social issues like this the mentality or mind-sets of the people which influences them to know that the movie will be hit or gets flop. In our project we aim to create a predictive model which will give the success rate of the movie. Parameters that we are considering are star cast, genre, director, budget. In our project we are using different algorithms KNN, Linear Regression, Naïve Bayes, Support Vector Machine (SVM) Linear Model, Logistic Regression to predict the collection of the movie. We will classify the movie as hit, flop, super hit on the basis of investment made by the movie for example if the movie was made on budget 50 crores and it has collected 40 crores on Indian box office then it will be like losing the movie even if it has earned worldwide. The prediction made is categorized into three flop, hit and super hit. Normally the net income defines the terms whether the movie is hit, flop, super hit.

1. **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.
2. **Hit:** If the movie earns the profit 20 percent more than the budget than the movie predicted hit.
3. **Super Hit:** For the movie to be a super hit budget plays a vital role. If the movie earns 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.

Related Works:

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 performance of different machine learning techniques was examined using 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 Boxofficeindia, wogma, cinematics 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]

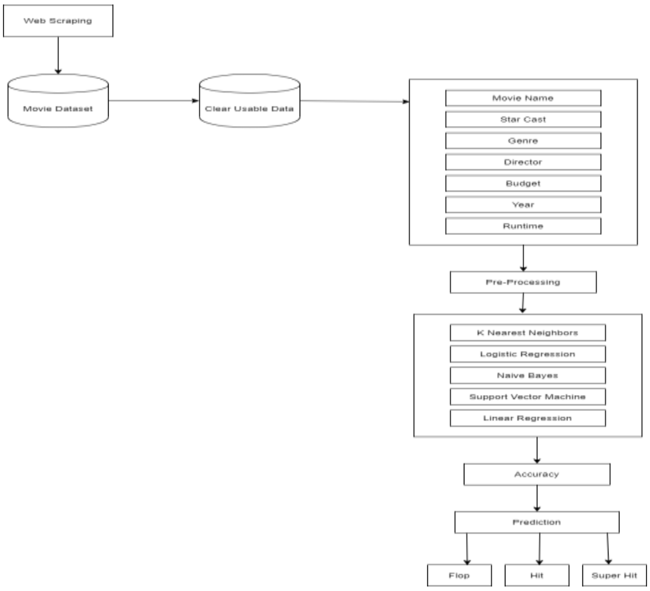


Figure 1: System diagram

III. PROPOSED WORK

Our objective is to classify movies as flop, hit, superhit on the basis of the return over investment of the movie from the domestic box office which the movie has earned. . Predictions are made right after the release of the movie and they are more accurate. In this paper we are using machine learning algorithms such as linear regression, logistic regression, multiclass naive bayes,Svm and k-mean for the prediction.

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

IV. METHODOLOGY

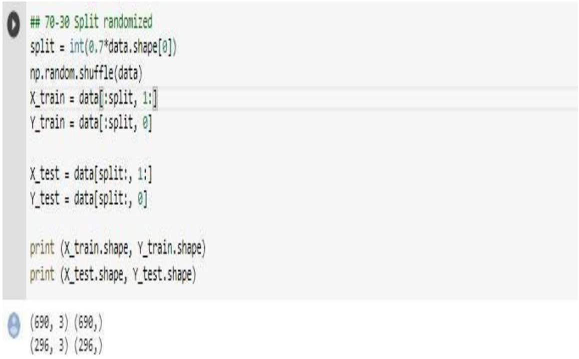
In this paper we illustrate revenue prediction as a discrete prediction using supervised learning. We have set of training (X (1), Y (1), X (2), Y (2)

....X(n), Y(n); where each x(I) corresponds to a vector of input features for an appropriate movie and y(I) belongs to R6 is a categorical dependent variable identical to six desirable revenue categories. We can model x(I)and y(I) and the relationship between them in different ways.

During the research development of this project we exploited 5 varieties of models to classify the input into revenue categories and we used this result from those classifications to approximate the domestic box office collection of the movie produced by the input vector.

Dataset was collected from numerous sources with the help of web scraping. Cross validation of the data as a figure for budget and box office collection because they are very different on multiple websites so in the final dataset we had over 65 percent of the movies which accomplish less domestic collection than their budget. Our assumption for this learning problem used high dimensional space than the sum of the sample we have available. This was the issue of the dimensionality to deal with this problem. We used sum with a linear kernel because we want to project the data into higher dimensional space. If the total number of features is large, nonlinear mapping does not improve performance but if we use a linear kernel it is good enough.

V. RESULTS



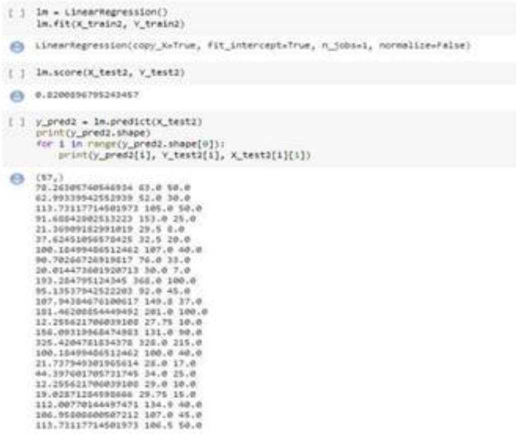
```
## 70-30 Split randomized
split = int(0.7*data.shape[0])
np.random.shuffle(data)
X_train = data[:split, 1:]
Y_train = data[:split, 0]

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,)
```

Figure 2.1: 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)
```

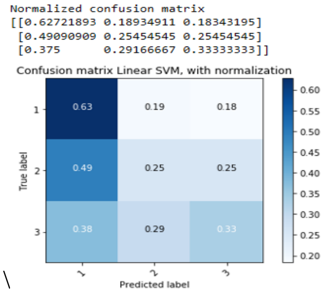
```
0.8200896795243457
```

```
[ ] 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,)
70.26306740546934 63.0 50.0
62.99339942552939 52.0 30.0
113.73117714501973 105.0 50.0
91.68842002513223 153.0 25.0
21.36909182991019 29.5 0.0
37.62451056570425 32.5 20.0
100.18499406512462 107.0 40.0
90.70266726919817 76.0 33.0
20.01447360192071З 30.0 7.0
193.284795124345 368.0 100.0
95.13537942522203 92.0 45.0
107.94384676100617 149.8 37.0
181.46208854449492 201.0 100.0
12.25562170603910 27.75 10.0
156.09313968474983 131.0 90.0
325.42047818З4378 328.0 215.0
100.18499406512462 100.0 40.0
21.73794930196561 28.0 17.0
44.39760170573174 34.0 25.0
12.25562170603910 29.0 10.0
19.02871284598466 29.75 15.0
112.00770144497471 134.9 40.0
106.95800600507212 107.0 45.0
113.73117714501973 106.5 50.0
```

Figure 2.2: Model fitting

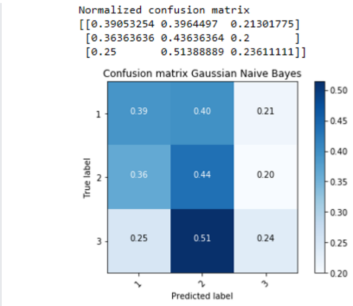
Confusion matrix



```
Normalized confusion matrix
[[0.62721893 0.18934911 0.18343195]
 [0.49090909 0.25454545 0.25454545]
 [0.375      0.29166667 0.33333333]]
```

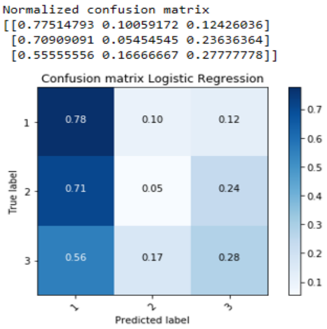
\

Figure 3.1: Linear SVM



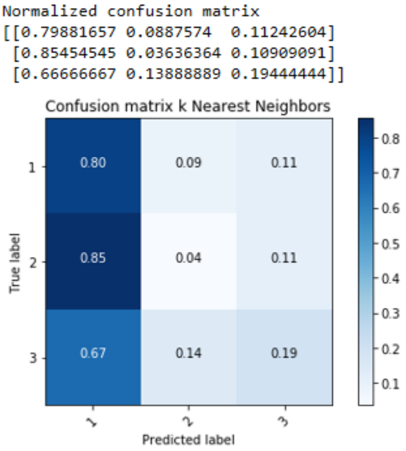
```
Normalized confusion matrix
[[0.39053254 0.3964497  0.21301775]
 [0.36363636 0.43636364 0.2       ]
 [0.25       0.51388889 0.23611111]]
```

Figure 3.2: Naïve Bayes



```
Normalized confusion matrix
[[0.77514793 0.10059172 0.12426036]
 [0.70909091 0.05454545 0.23636364]
 [0.55555556 0.16666667 0.27777778]]
```

Figure 3.3: Logistic Regression



```
Normalized confusion matrix
[[0.79881657 0.0887574  0.11242604]
 [0.85454545 0.03636364 0.10909091]
 [0.66666667 0.13888889 0.19444444]]
```

Figure 3.4: K Nearest Neighbours

Linear Regression gives more accurate accuracy for the prediction

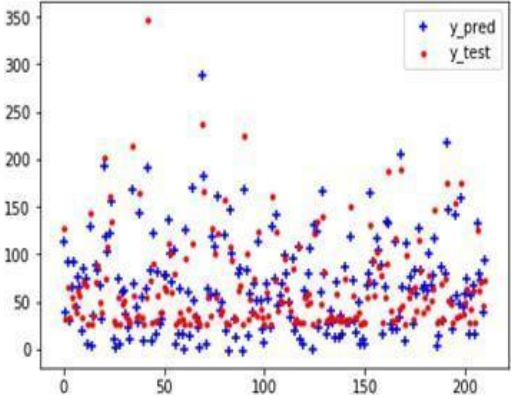


Figure 4: Linear Regression

Model	Accuracy	Recommendation
linear regression	0.82	absolutely good for prediction
Sum 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
in	0.51	it gives all the flop prediction don’t use
logistic regression	0.55	it is little better than knn but don't use it for prediction

Figure 5: Comparatively analysis

VI. CONCLUSION AND 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 things have excluded these parameters. We have done web scraping along with parameters like star cast, genre, year, 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 analysed. 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)

[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)

[3]. Kyuhan Lee,Jinsoo Park,Ijoo Kim, Youngseok Choi “Predicting Movie Success with Machine Learning Techniques: ways to improve accuracy”

[4]. Sameer Ranjan Jaiswal, Divyansh Sharma “ Predicting Success Of Bollywood Movies Using Machine Learning” Proceedings of the 10th Annual ACM India Computer Conference

[5]. Ting Liu, Xiao Ding , Yiheng Chen, Haochen Chen, Haochen Chen, Maosheng Guo “ Predicting movie Box-office revenues by exploiting large-scale social mediacontent”

[6]. Anand Bhave, Himanshu Kulkarni, Vinay Biramane, Pranali Kosamkar “ Role of different factors in predicting movie success” 2015 International Conference On Pervasive Computing(ICPC)

Ansari Sana Fatima. "Movie Success Prediction Using Machine Learning." *International Journal of Engineering Research And Development*, vol. 16(6), 2020, pp 10-14.