# Prediction of Movie Success or Failure Analysis Using Machine Learning Approaches

Dr DINESH KUMAR

*a#, M.Tech student, Department of Computer Science & Engineering, Chouksey Engineering College, Bilaspur- 495004, India.*
*b, Assistant Professor, Department of Computer Science & Engineering, Chouksey Engineering College, Bilaspur- 495004, India*

## Abstract:

As per the worldwide film industry, the income had hit a record $42.5 billion in the year 2019. Nowadays, a massive amount of money is invested in producing and publicizing a movie on different media and digital platforms. In such circumstances, prior information about the movie's success or failure and the factors affecting it will be beneficial for the people involved in the making of that particular movie. So, the prediction of a movie's success is of great importance to the investors and the public. At present, Machine learning algorithms are widely utilized for prediction in areas such as growth in the product demand, railways budgeting, traffic prediction, supply chain management etc. IMDB database consists of numerical and categorical information such as the actors involved, directors, year in which the movie was released, genre(Action, Animation, Comedy Genre, Crime Genre, Drama Genre, Experimental Genre), total runtime of movie, user rating, number of votes etc., total revenue generated, overall Metascore, ongoing trends and so on. In this study, a novel solution is introduced for the forecast of success of a movie in terms of creation houses and viewer recipience. A thorough analysis of IMDB (Internet Movie Database) and a comparative study of Support Vector Machine (SVM), Adaboost, Logistic Regression, Naïve Bayes Classifier (NB classifier) and K-Nearest Neighbours (KNN) on IMDB is presented.

Key terms: SVM (Support Vector Machine), ROC, ML (machine learning), IMDB etc.

## Introduction

A motion picture's income is significantly affected by different reasons such as the film's executive, star cast, budget, ratings, release date, year. Due to these factors, there is no accurate method through which prior knowledge of success can be known. Through investigation of the incomes of previously released movies, a model is designed, which can assist in forecasting the standard income film [1]. Such prediction could be extremely valuable for the film finance experts for putting resources into the making of that film. These experts and the studio could choose a name and ad for the film suitable to the concurring model. This could be useful for cineplexes to earn profit for screening a specific movie [2]. Figure 1 portrays how a motion picture's favourable outcome is influenced.

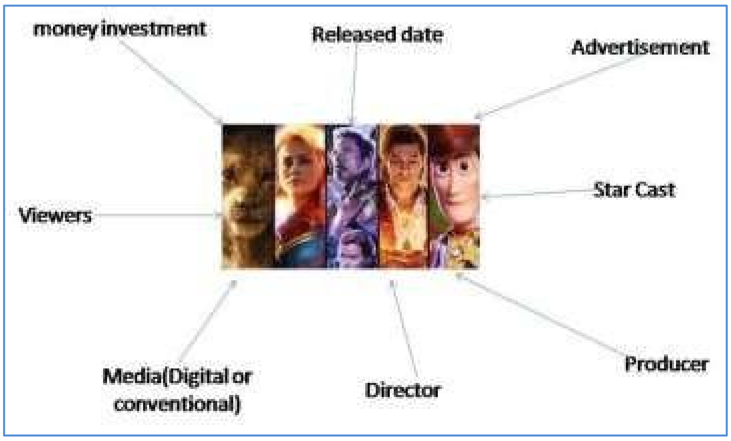

Figure 1. Factors that influenced the success or Failure of movies

These days, online survey framework, (for example, IMDB) has become a significant part of any film business approach. Posting surveys online for items purchase is an approach of communication used by individuals for business insight. This method is applied in the film business as well since online networking contains rich data about individuals' inclinations [4].

In this proposed research, a supportive network for movie speculation has been presented through information mining methods. In this examination, altered science calculations on a dataset which contains various highlights are applied. As per the determined qualities, the respective feature film is categorised into hit, normal or lemon. Through this investigation, an information mining calculation is provided, which gives the most precise outcome for film achievement expectation [5].

**Data Set:**

The data used for this research was obtained from IMDB dataset [3]. The feature films released between 2006 and 2019 in the English language with gross revenue of at least $760,000 is used in this research. Using these search criteria, around 1000 movies is found. The purpose behind setting the timeframe limitation is that we just need to incorporate ongoing motion pictures as it is challenging to analyse motion pictures from various times.

*1. Features of Data Set*

The metadata description that is included in the class (as characterized by IMDB dataset) defined in the following table with example.

Table 1. Data Set Feature

| Name of Feature | Data |
|---|---|
| **Rank** | 1 |
| **Title of movie** | Guardians of the Galaxy- II |
| **Type** | Adventure, Action, Sci-Fi |
| **Explanation** | A gathering of intergalactic lawbreakers are compelled to cooperate to prevent an over the top warrior from assuming responsibility for the universe. |
| **Director** | James Gunn |
| **Actors** | Chris Pratt, Vin Diesel, Bradley Cooper, Zoe Saldana |
| **Year** | 2016 |
| **Movie Time (Minutes)** | 123 |
| **score** | 8.9 |
| **vote** | 755674 |
| **Revenue (Millions)** | 344.13 |
| **Meta score** | 77 |
| **Action** | 1 |
| **Adventure** | 1 |
| **Animation** | 0 |
| **Biography** | 0 |
| **Comedy** | 0 |
| **Crime** | 0 |
| **Drama** | 0 |
| **Family** | 0 |
| **Fantasy** | 0 |
| **History** | 0 |
| **Horror** | 0 |
| **Music** | 0 |
| **Musical** | 0 |
| **Mystery** | 0 |
| **Romance** | 0 |
| **Sci-Fi** | 1 |
| **Sport** | 0 |
| **Thriller** | 0 |
| **War** | 0 |

| | |
|---|---|
| **Western** | 0 |
| **Success** | 1 |

Using IMDB data source can be found in Table 1. From this, the information about actors involved, directors, year of release, movie genre, total runtime of movie, user rating, total votes, total revenue generated by the movie, the overall meta score, age of the viewers, the geographical location of the movie release and other influences such as political movements, ongoing trends etc. are extracted. There has not been much exploration of this. So, the novelty of this study is the use of machine learning algorithm with these number of parameters to predict movies success or failure of new motion pictures. Another distinguishing feature of this analysis is prognosis and comparison of five different algorithms.

## 2. Model and Design

In our research, we have worked with five different algorithms named SVM(Support Vector Machine), Naive Bayes Classifier, Logistic regression, AdaboostandK-neighbour.

### *1. SVM (Support Vector Machine)*

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. Support Vector Machine representation is a portrayal of the models focused in space, mapped with the goal the models of the separate classes are partitioned by a recognizable gap that is as wide as possible. In expansion to the stage straight order, SVMs can productively play out a non-direct categorization, implicitly mapping their contributions to high-dimensional element spaces. Given a lot of preparing models, each set apart as having a place with either of two classifications. An SVM preparing calculation constructs a model that allots new guides to one classification or the other, making it a non-probabilistic parallel straight classifier. For example, to classify apple and orange according to their colour and weight, concentrating on the weight and size(diameter) of apples and oranges. Presently how might a machine utilizing SVM, arrange another natural product as either apple or orange is simply dependent on the information on the size and mass of 20 apples and oranges that were observed and named. [6].

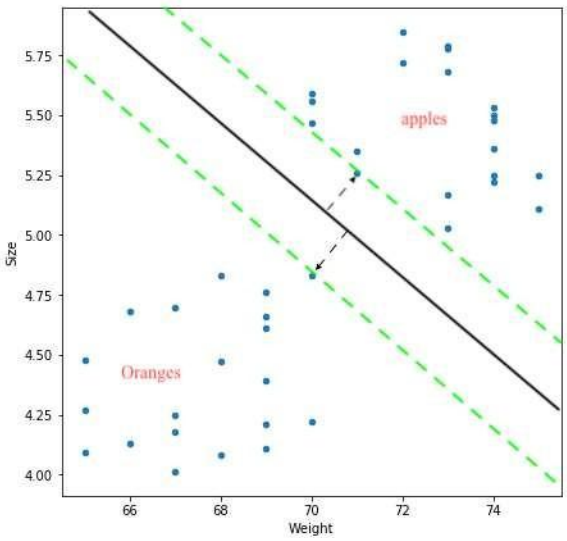

Figure 2. SVM classification of apple and orange using machine learning

2. KNN

KNN (K-Nearest Neighbour) is used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry [7]. Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbours) and summarizing the output variable for those K instances. Three accustoms are taken that is utilized in iterative preparing of a k-NN classifier. This is a novel approach both from the measurable example acknowledgement and the directed neural system learning perspectives. To assess any procedure, two critical viewpoints are considered:

1. Easily interpret the output.
2. Less Complexity or Complexity optimization.
3. Computation time

Here below given an example to find out the class of blue star using KNN algorithm

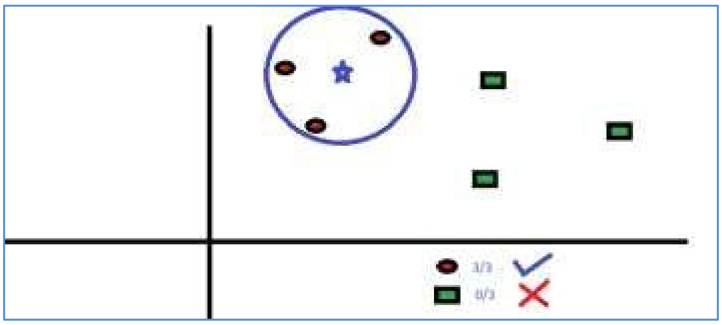

Figure 3. Intend to find out the class of the blue star (BS)

### 3. Naive Bayes Classifier

Bayes' theorem provides a path of computing the probability of information with a given class, which was given prior data. Bayes' Theorem is expressed as:

$$\mathrm{P\ (class/data)} = {}^{(\mathrm{P\ (data/class)}\ *\ \mathrm{P\ (class)})}/_{\mathrm{P\ (data)}}$$

Where P (class / data) is the likelihood of class of given data. A brief introduction to Bayes Theorem for Machine Learning is a grouping calculation for twofold (two-class) and multiclass arrangement issues [8]. Analysis of probabilities for each class is simplified to make their counts tractable. Rather than endeavouring to ascertain the quality of these probabilities, they are expected to be conditionally free given the class. Below is an example for iris bloom includes to anticipating the blossom species given estimations of iris flowers. It is a multiclass grouping issue. The quantity of perceptions for each class is adjusted.

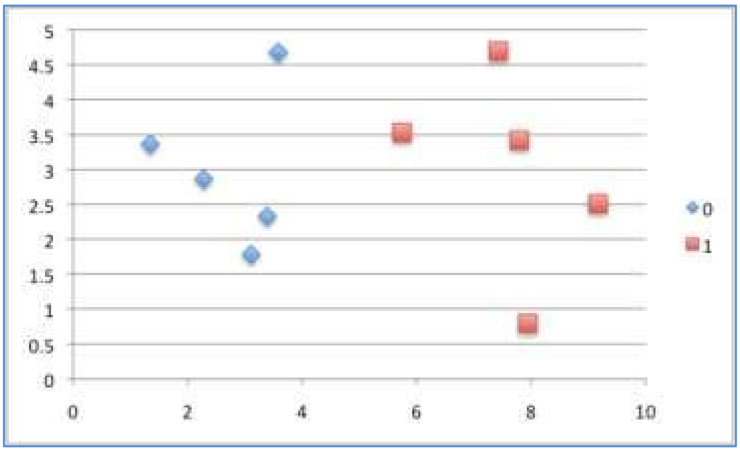

Figure 4. Disperse Plot of Small Contrived Dataset for Testing the Naive Bayes Algorithm for Iris blossom Dataset [8]

### *4. AdaBoost*

As a data science researcher in the consumer industry, boosting calculations are enough for a large portion of the prescient learning errands, at any rate in the present time. They are incredible, adaptable and can be deciphered easily with some ways [10].

F(x)=sign ($\sum^{M}_{=1}$ **ϴ**m fm(x))...................................eq (1)

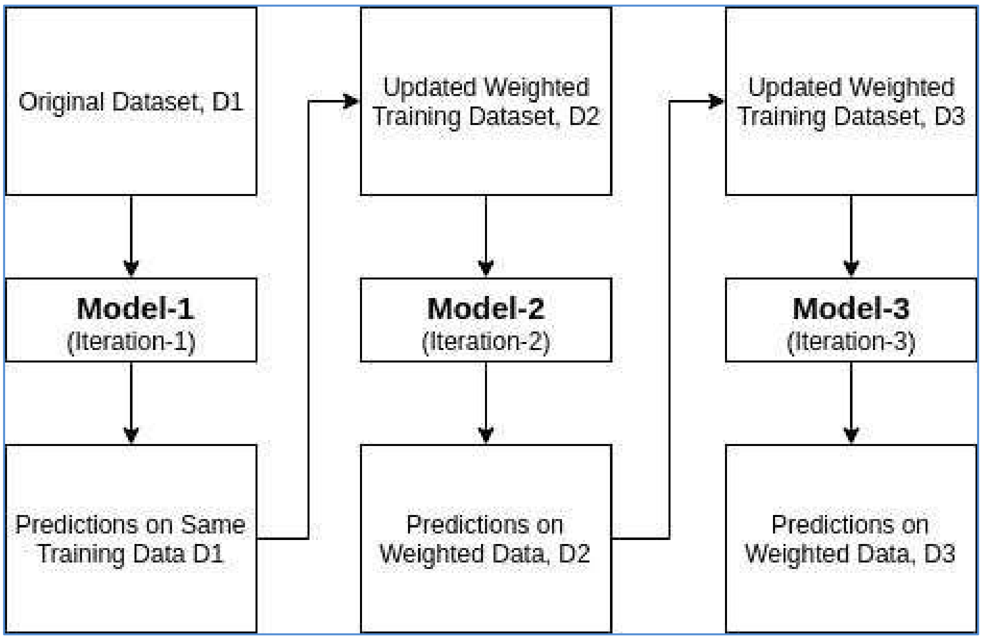
Original Dataset, D1
Model-1
(Iteration-1)
Predictions on Same Training Data D1
Updated Weighted Training Dataset, D2
Model-2
(Iteration-2)
Predictions on Weighted Data, D2
Updated Weighted Training Dataset, D3
Model-3
(Iteration-3)
Predictions on Weighted Data, D3

Figure 5. AdaBoost Decision Tree algorithm architecture

### 5. *Logistic Regression*

Logistic Regression is a popular statistical model used for binary classification for example predictions of the type A or B, or choice of yes or no etc. Logistic regression is also used for multiclass classification, but this study focuses on its simplest application. For instance, a task of foreseeing somebody's gender orientation (Male/Female) based on their Weight and Height.

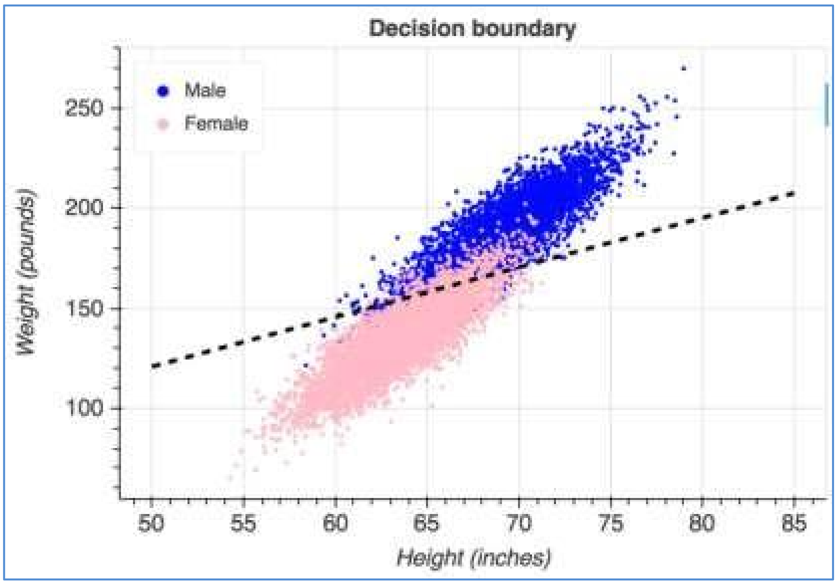

Figure 6. Two parameters- weight and height are used to predicting Gender

## 3. Experiment

### *1. Evaluation metrics*

Assessing machine learning and AI calculation is a fundamental piece of this experimentation. This model may give fulfilling results when assessed utilizing a measurement state precision score, however, show some restraints when assessed against different measurements, for example, logarithmic misfortune or other such aspects. Some major situations are used for grouping accuracy to determine the exhibition of this model; however, it is not sufficient to estimate the efficiency of the model. Therefore, there are various kinds of assessment measurements accessible like Logarithmic Loss, F1 Score, Mean Absolute Error, Mean Squared Error, Area under Curve, Classification Accuracy and Confusion Matrix used for assessment.

After executing all above mentioned in Model and design, we found the following result from the experiment. Receiver Operating Characteristics (ROC) shows the pictorial representation and graph of below table.

Table 2. Result from the experiment. Receiver Operating Characteristics (ROC)

| **Algorithm** | **Confusion Matrix** | **Accuracy** | **Precision** | **AUC** | **Recall is** |
|---|---|---|---|---|---|
| SVM | [[166 7]<br>[ 10 27]] | 0.92 | 0.79 | 0.84 | 0.73 |
| KNN | [[169 4]<br>[ 18 19]] | 0.90 | 0.83 | 0.82 | 0.51 |
| Naive Bayes Classifier | [[147 26]<br>[ 8 29]] | 0.84 | 0.53 | 0.71 | 0.78 |
| AdaBoost | [[172 1]<br>[ 0 37]] | 0.99 | 0.97 | 1.00 | 1.0 |
| Logistic Regression | [[162 11]<br>[ 10 27]] | 0.9 | 0.71 | 0.83 | 0.72 |

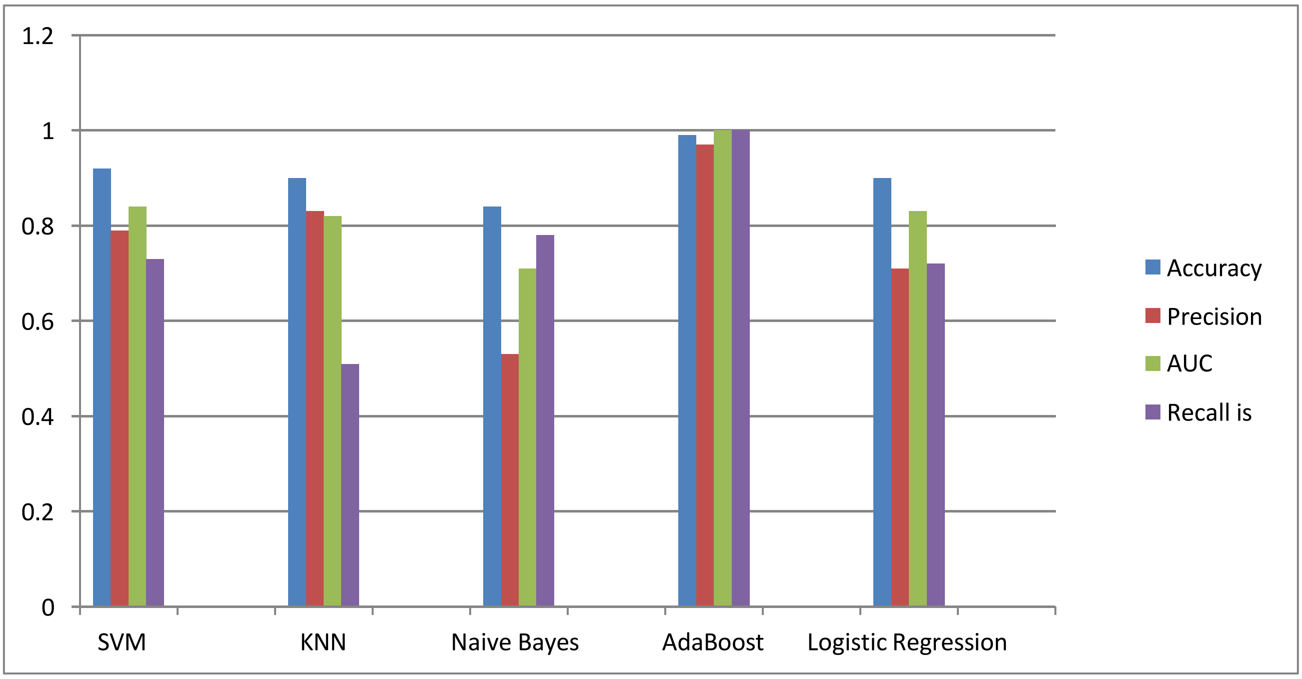

Figure 7. Comparative algorithm performance graph based on Table 1

A beneficiary working trademark bend, or ROC bend, is a graphical plot that represents the demonstrative capacity of a parallel classifier framework as its separation limit is shifted. The ROC bend is made by plotting the genuine positive rate (TPR) against the bogus positive rate (FPR) at different edge settings. The genuine positive rate is otherwise called affectability, review or likelihood of detection in Machine Learning. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups.

*2. Visualization of Receiver Operating Characteristics*

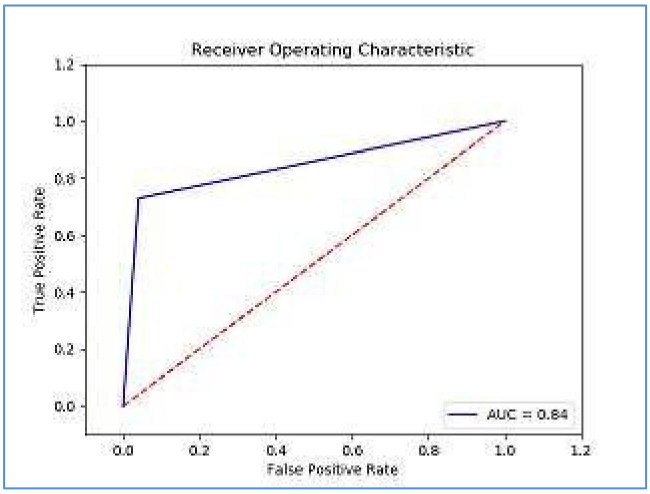

Figure 8.ROC of SVM

SVM calculations gave the precision of 92%, and the AUC value of 0.84, which is acceptable. Despite the precise results obtained by the calculations, it is required to meet to expectation/ criteria.

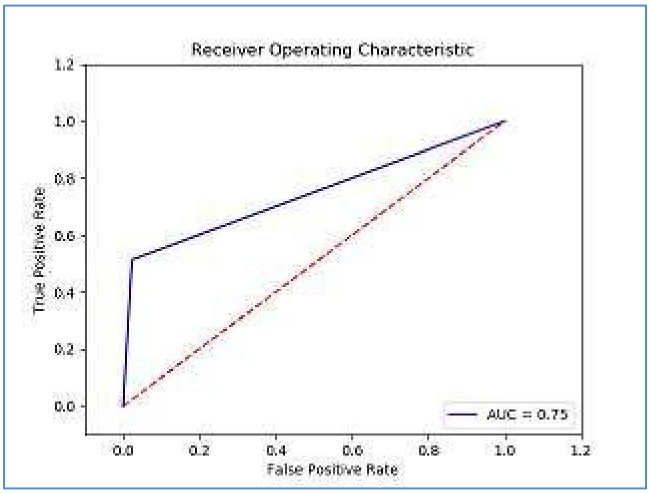

Figure 9. ROC of KNN

KNN calculation is one of the least difficult characterization calculations. Indeed, even with such straightforwardness, it gives profoundly good outcomes. The main contrast KNN classifier has from relapse is the procedure, which utilizes the midpoints of closest neighbours as opposed to casting a ballot from closest neighbours. In light of the above outcomes, it can be surmised that the K-Nearest Neighbour classifier at k = 5 has a decent exactness of 90% and the ROC bend gives an AUC of 0.75.

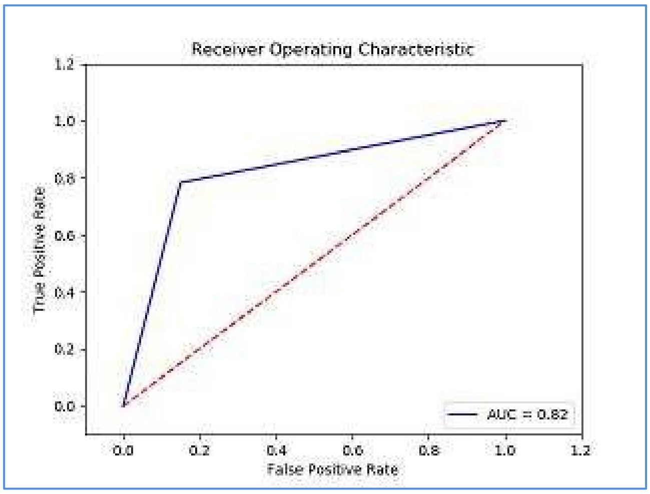

Figure 10. Receiver Operating Characteristics of Naive Bayes Classifier

The precision for Naive Bayes Classifier is 84 %. ROC is useful; however, in contrast with different models which are under the extent of this investigation, NBC appears to a failing to meet expectations for the given dataset. AUC came to be 0.82, which is a conventional figure.

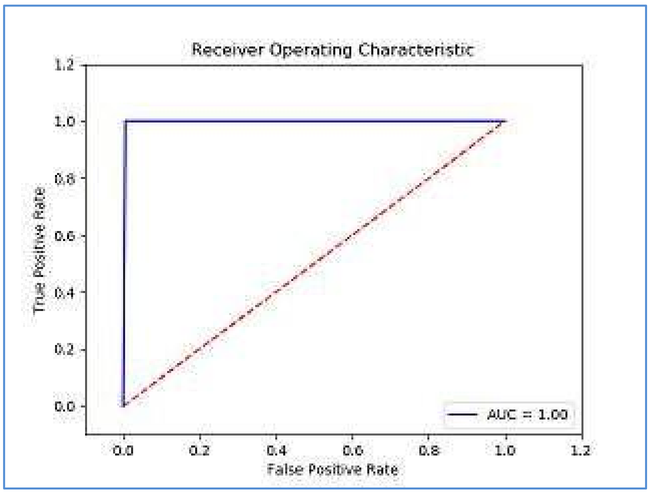

Figure 11. ROC of AdaBoost

The Adaboost calculation is utilized related to Decision tree calculation with a great profundity being equivalent to 2. Choice trees are utilized with Adaboost as they are non-straight while being easy to create. They are likewise quick to group and consequently can be utilized in enormous numbers. From the above outcomes, it tends to be gathered that the Adaboost has an exceptionally high precision of 99%. The ROC bend gives an AUC of 1, which is an ideal score showing an ideal test.

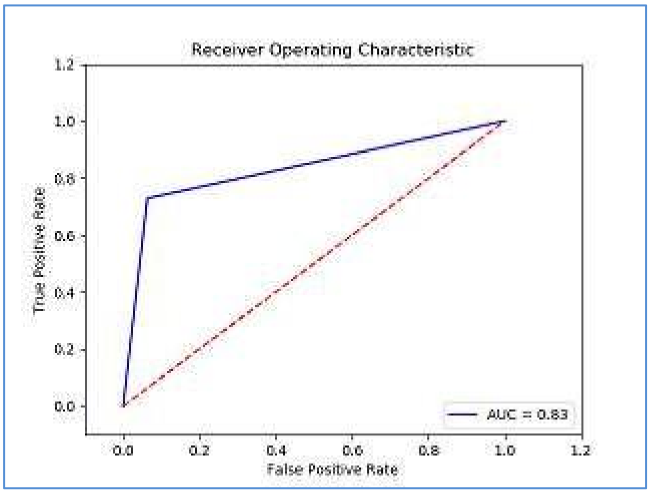

Figure 12. ROC of Logistic Regression

From the above outcomes, it tends to be induced that when twofold qualities are considered as data, the Logistic relapse classifier has a decent precision of 90.0% and the ROC bend gives an AUC of 0.83. The forecasts are very high, and this calculation is truly steady if the dataset has more than one free factor.

## Conclusion

After the trial of the models, it is discovered that the achievement rate for all models is almost the equivalent and Adaboost model had the most elevated precision for this situation for anticipating the film’s success. A greater number of attributes and training set is the key to improving the performance of the model. That is why in future more number of features such as votes, date, time, social networks data analysis, age of viewers, current trends, news break down and location could be considered to the training set in order to get a more accurate result.

## Conflict of Interest

The authors declare that they have no conflict of interest.

## References

[1] Billboard, Available at: https://www.billboard.com/articles/news/8547827/2019-global-box-office-revenue-hit-record-425b-despite-4-percent-dip-in-us, accessed on June 2020.

[2] Statista, Available at: https://www.statista.com/topics/964/film/, accessed on June 2020.

[3]. IMDB Dataset, Available at: https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews, accessed on June 2020.

[4]. Ramesh Sharda and Dursun Delen. Predicting box-office success of motion pictures with neural networks. *Expert. Syst. Appl*. 2006; **30 (2)**, 243–254.

[5]. Radas, S. and Shugan, S. M. Seasonal marketing and timing new product introductions. *J. Mark. Res*. 1998; **35(3)**, 296–315.

[6]. Zhang Y. and Wu L. Classification of Fruits Using Computer Vision and a Multiclass Support Vector Machine. *Sens.* 2012; **12(9)**, 12489-12505.

[7] Laaksonen, J. and Oja, E. Classification with learning k-nearest neighbors. IEEE Volume III. *In*: Proceedings of International Conference on Neural Networks (ICNN'96). 1996, pp. 1480-1483.

[8] R.S. sanjuvigasini et.al. An efficient comparision of data classification algorithm for analysis of iris data sets. *Inter. J. Adv. Eng. Res. Dev.* 2018; **5(2)**.

[9]. Iris Dataset, Available at: https://archive.ics.uci.edu/ml/datasets/Iris, accessed on June 2020.

[10]. Liao Shaowen and Chen Yong. A kind of improved AdaBoost algorithm. *In*: 7th International Conference on Intelligent Computation Technology and Automation. 2014.

[11]. Al-Zuabi I.M., Jafar A and Aljoumaa K. Predicting customer's gender and age depending on mobile phone data. *J. Big Data.* 2019; ***6*(1)**, 18.

[12]. Asur, S., & Huberman, B. A. Predicting the Future with Social Media. IEEE Volume I. *In:* IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology. 2010, pp. 492-499.

[13]. Mestyán, M.; Yasseri, T.; and Kertész, J. Early prediction of movie box office success based on Wikipedia activity big data. *PloS one*. 2013; **8(8)**.

[14]. Oghina A, Breuss M, Tsagkias E and De Rijke M. Predicting IMDB movie ratings using social media. Springer, Berlin, Heidelberg. *In:* 34th European Conference on Information Retrieval. 2012, 503–507.

[15]. Lash, M. T. and Zhao, K. Early predictions of movie success: The who, what, and when of profitability. *J. Manag Inf. Syst*. 2016; 2015, ***33*(3)**, 874-903.