International Journal of Scientific Research & Engineering Trends
Volume 9, Issue 3, May-Jun-2023, ISSN (Online): 2395-566X

Predicting Movie Success through Ratings Analysis: A Machine Learning Approach

Dipesh Shah , Swarda Mashere, Ayush Kumar, Shamika Chalse , Dr. Rajendra Pawar
School of CET, MITWPU
Maharashtra , India

Abstract- one of the main forms of entertainment worldwide is the movie. This essay focuses on predicting a movie's success rate by performing predictive analysis on the film's numerous elements. The success of films was predicted using machine learning algorithms in this paper, including Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes, Logistic Regression, and ensemble methods. Classifier provided the highest accuracy, according to the data.

Keywords **– Machine Learning, Classification, Prediction, IMDB Data.**

I. INTRODUCTION

Movies are no longer the only source of entertainment; rather, they represent a significant source of international trade and marketing. A strong plot, screenwriting, casting, directing, sound recording, and many other tasks are all part of the filmmaking process. Every year, the film industry creates hundreds of films in a variety of genres, including horror, comedy, thriller, and animation. Each of us is uniquely inspired by cinema. Through this medium, we can learn a lot. A comedy makes us laugh our sorrows away, a science fiction film makes us think large, a biopic makes us realize our aspirations, and so forth.

As a result, the movie business has a major impact on the economy of every nation. Profits from blockbuster hits are enormous, but losses from flops are equally significant. And each directly affects both positive and negative outcomes. There are numerous websites that track films, including Rotten Tomatoes, Metacritique, and the Internet Movie Database (IMDb), which offers details on the directors and budget in addition to user reviews and comments. The most popular movie-related website, Internet Movie Database (IMDB), offers a wealth of information about films, television shows, and other media, including financial statistics, biographies, user ratings, cast and crew lists, reviews, and summaries. It contains a database with 6.5 million movie and episode titles, 10.4 million personalities, and about 83 million registered users.

Every year, thousands of films are released. The global box office brought in $41.7 billion in 2018. The worldwide film business generated $136 billion in box office and home entertainment sales in 2018. With 1,813 feature films produced each year as of 2018, India's film industry is the greatest in terms of production volume. The success of films has been the subject of extensive research. The majority of previous study concentrated on user reviews, with social media sites like YouTube, Twitter, and others serving as the main data sources. Movie qualities including cast, release dates, production companies, and storyline, among others, are a vital information source and can significantly influence the success of a movie's forecast. There is a wealth of information about various movie attributes available on the internet from sources like IMDB, making this a significant use-case for data mining and machine learning. This is because it is quite relevant that successful movie prediction is of great relevance to this multi-Billion Dollar Industry. Producers and filmmakers will be assisted in creating films that are better suited to the tastes of the audience.

II.LITERATURE REVIEW

In a report published in December 2020, Narayana Darapaneni and Christopher Bellarmine [1] discussed utilising machine learning to predict a movie's rating. They used a scale of 1 to 9 to rate the movie's budget in their work. The neural network's accuracy ranged from 79.07% at the lowest end to 84.34% at the best.

The box office income was predicted using a variety of machine learning algorithms by Nikhil Apte, Mats Forssell, and Anahita Sidhwa [2]. They only took into account films that were released after January 1, 1990 because the data prior to that was insufficient. The final dataset has 2510 records in it. Algorithms like linear regression and weighted linear regression were employed. To estimate testing errors, hold-out cross validation was performed. Instead than listening to critics and other people about whether a movie would be successful or not, Rijul Dhir and Anand Raj [2] tried to forecast how successful a movie will be before it arrives at the box office. The suggested research offers a rather effective method for predicting IMDB score on the IMDB Movie Dataset. They also attempted to identify the key elements affecting the IMDB Movie Data score in their investigation. The most accurate model was the random

forest. Another method was forecasting movie success using social media interactions. The dataset's source was the quantity of tweets and exchanges. Using social media channels, Sitaram Asur and Bernardo A. Huberman [3] did a study on future prediction. demonstrating that it is possible to predict a movie's box office take using social media. Twitter was employed as the data source for this study. At the university of Skövde, Karl Persson [4] compared the prediction abilities of Random Forest and support vector machines. He had an 84% success rate with random forest and an 86% success rate using support vector machines. He applied 10-fold cross validation on his results to validate them. Similar research, utilising social media sites like Twitter and YouTube comments, has been presented [5]. A other strategy [6] offers Wikipedia articles as a way to estimate a movie's level of popularity.

Table 1. Summary Of Previous Work.

Study\Method and Results	Validation	Prediction	Success Rate
Predicting Movie Box Office Gross	20% withhold from data set	Movie Revenue	65%
Prediction of Movies popularity Using Machine Learning Techniques	10-fold cross validation	Movie Rating	80%
Predicting movie ratings: A comparative study on random forests and support vector machines	10-fold cross validation	Movie Rating	83%
Predicting the Future with Social Media	Cross validation	Box office revenue	98%

III.METHODLOGY

The working method for this work involves few steps. The methodology is shown in figure 1. The steps are described below.The working method for this work involves few steps. The methodology is shown in figure 1. The steps are described below. The working method for this work involves few steps. The methodology is shown in figure 1. The steps are described below The working method for this work involves few steps. The methodology is shown in figure 1. The steps are described below There are only a few steps in the working process for this work. In Figure 1, the process is displayed. Below is a description of the steps. Data Extraction, Data Preprocessing, Machine Learning Methods, Application, and Comparison of the Results of Different Algorithms

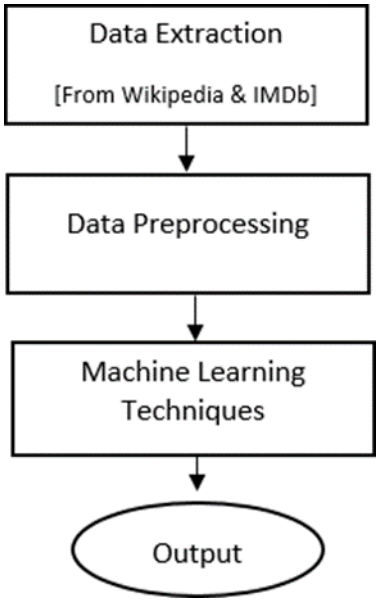

Fig. 1. Working Methodology

1. Data Extraction

Information was taken from the IMDb movie review website. The movie's title, name, year, rating, genre, certificate, runtime, budget, and box office are all extracted from IMDB. The director, screenplay, and cast list for each film are all listed in the cast and crew column of the IMDB data. Second, the IMDb rating for each chosen film is taken directly from the IMDb website. IMDb ratings are growing in popularity every day. IMDb ratings used to be given and trusted. This platform is preferred for this work because of this. IMDb assigns a score of 10 to each film. Table 2 displays the ranges of ratings for each class. In other words, if a movie has an IMDb rating of 3, it will most likely be a flop. The grade will be Average if it is a 7. According to its IMDb rating, each movie in the data set falls into one of four categories: flop, below a typical basis, average, or hit. Consequently, the data set includes real data that was taken from the IMDB website.

Table 2 Class Consideration

Range of Rating	Class
0-3.5	Flop
3.6-5.8	Below Average
5.9-7.4	Average
7.5-10	Hit

2. Data Preprocessing

To preprocess data is to get it ready for classification. Data is handled in accordance with classification standards. In this case, instances with missing attributes are eliminated as part of the data preprocessing. Finally, a data set of 250 films was obtained. Movie, name, year, rating, genre, certificate, runtime, budget, and box office are among the features of the processed data set.

3. Applying Machine Learning Techniques

To create the model, we used five machine learning methods.

Table 3 List Of Classifiers Used

Name of the Classifier
Decision Tree
Naïve Bayes
KNN
Logistic Regression
Ensemble Learning

4. Decision Tree

Although supervised learning is used to handle classification and regression concerns, classification challenges are typically the focus It is a classifier based on a tree structure, with each leaf node denoting the result and inner nodes denoting the dataset's characteristics. A decision tree's Decision Terminal and Leaf Node are its two nodes. It is feasible to make any decision using Decision Nodes, which have a lot of branches, and Leaf Nodes, which reflect the outcomes of the choices made and have no more branches. Based on established criteria, it is an illustration for identifying each alternative response to a question or option. The classifier and regression tree algorithm (CART) builds a tree The only thing a decision tree does is pose a question, and based on the response (Yes/No), it is further subdivided.

5. Naïve Bayes

There are many categorization jobs that can be handled by this probabilistic learning approach. Bayes' Law: The Bayes theorem, sometimes referred to as Bayes' Principle or Bayesian' the law, is a method to assess the likelihood of a hypothesis in light of the information at hand. This is decided by the conditional probability. The Bayes theorem's formula is as follows:

$$P(Q \mid T) = P(T \mid Q)\, P(Q) / P(T) \qquad (1)$$

The chance that an event (T) witnessed would occur given a certain hypothesis (Q) is denoted by the symbol P(Q|T), where P(Q|T) is the posterior probability. Likelihood Probability, or P(T|Q), measures how likely it is that a certain hypothesis is correct in light of the information at hand. The probability of a theory before viewing the evidence is known as priority probability, or P(Q). Marginal probability, or P(T), is the likelihood of an event.

6. KNN

Although classification issues are where the K-NN approach is most frequently utilized, it may also be used for regression. Finding the category or class of a given dataset is made simple by K-NN. Putting the K-NN algorithm into practice: The test set result visualization method consists of the following steps: data pre-processing, K-NN algorithm fitting to the training set, estimation, test accuracy, building of the confusion matrix, and visualization.

7. Logistice Regression:

The categorical variable of dependence is forecasted using a predetermined set of independent factors. When a dependent variable is categorical, the output may be predicted using logistic regression. Therefore, the result has to have a discrete or category value. t provides probabilistic values in the 0–1 range as opposed to a precise value between 0 and 1. There are several possible outcomes, including True, False, Yes, No, 0 or 1. The predicted values are converted to probabilities by applying a mathematical approach called the sigmoid function. Any true value between 0 and 1 is converted into a different value.

The outcome of the logistic regression must fall between the range of 0 and 1, therefore because it cannot exceed this value, it assumes the form of a "S" curve. The logistics function or sigmoid function, respectively, are two other names for the S-form curve. The threshold value notion, which establishes the probability of either 0 or 1, is used in logistic regression. Take values as an example that slope form 0 under the limit threshold to 1 throughout it.
Formula for Logistic Regression The equation for a straight line is, as we well know, is:

$$Y=p0+ p1 * t1 + p2 * t2 + p3 * t3 + \ldots. + pn * tn \qquad (2)$$

7.1 Ensemble Learning:

To address a specific issue, various Machine Learning models (classifiers) are merged. Dedicated ensemble technique for decision tree classifiers. Random Many trees thrive in forests. Unpruned decision tree ensemble Every base classifier sorts a "new" vector of attributes from the starting set of information. Voting is the final step in classifying a new instance. Out of all the trees in the forest, the classification result with the highest votes is chosen by the forest.

IV.RESULTS

We are predicting the success rate of films using data visualization approaches to generate the output.

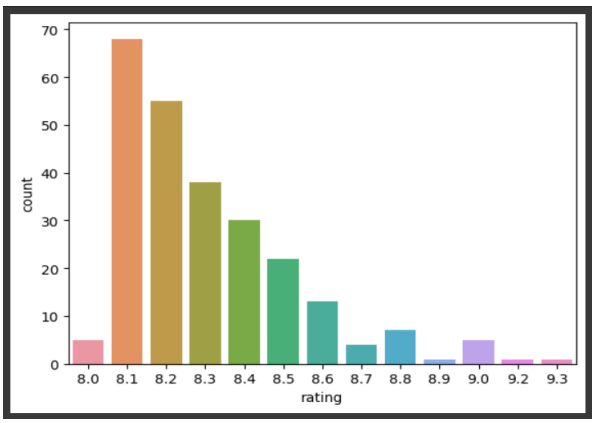

Fig. 2. Pictorial Representation of Count plot.

The observational count is generally displayed using bars in different category-based bins. The graph above displays the movie with the highest rating, 8.1, as well as how the public views that rating.

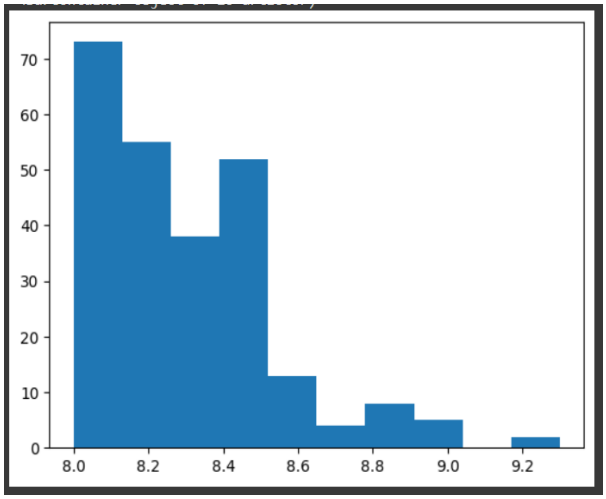

Fig. 3. Pictorial Representation of Histogram.

It provides us with a total of the observations kept in each bin that was created for visualization. Based on audience response, the best and worst possible movie ratings are displayed

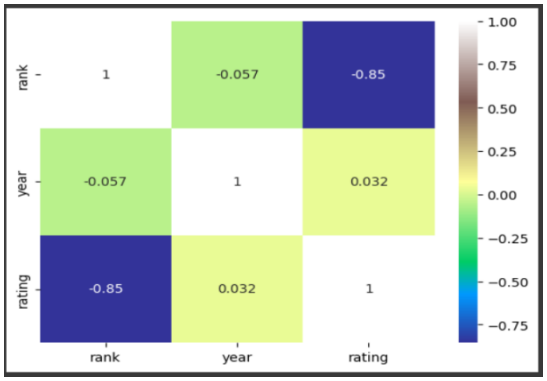

Fig. 4. Pictorial Representation of Heat map.

The correlation matrix of the heat map is used to generate color coding. It demonstrates the degree of correlation between various variables using these coefficients. Darker hues represent a link that is favorable, whereas lighter hues represent a correlation that is negative.

REFERENCES

[1] Pawar, R., Ghumbre, S., & Deshmukh, R. (2018). Developing An Improvised E-Menu Recommendation System For Customer. In Recent Findings In Intelligent Computing Techniques: Proceedings Of The 5th Icacni 2017, Volume 2 (Pp. 333-343). Springer Singapore.

[2] Sign Gesture Interpreter For Better Communication Between A Normal And Deaf Person D Wategaonkar, R Pawar, P Jadhav, T Patole, Rr Jadhav, S Gupta Journal Of Pharmaceutical Negative Results, 5990-6000, 2022

[3] A. Ö. Eren And M. Sert, "Movie Rating Prediction Using Ensemble Learning And Mixed Type Attributes," 2017 25th Signal Processing And Communications Applications Conference

[4] [4] Mohini Gore, Aishwarya Sheth, Samrudhi Abbad, Paryul Jain, Prof. Pooja Mishra, “Imdb Box Office Prediction Using Machine Learning Algorithms”, Ijraset

[5] N. Darapaneni Et Al., "Movie Success Prediction Using Ml," 2020 11th Ieee Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (Uemcon)

[6] B. Çizmeci And Ş. G. Ögüdücü, "Predicting Imdb Ratings Of Pre-Release Movies With Factorization Machines Using Social Media," 3rd International Conference On Computer Science And Engineering (Ubmk)

[7] Aashi Goyal, “Prediction Of Movie Performance Using Machine Learning Algorithms, “Ijraset

[8] Muhammad Hassan Latif and Hammad Afzal, “Prediction of movies popularity using machine learning techniques”, (UEMCON)

[9] M. Marovic, M. Mihokovic, M. Miksa, S. Pribil and A. Tus, "Automatic movie ratings prediction using machine leaming”, International Conference on Communication and Electronics Systems

[10] Ritu Khandelwal, Harshita Virwani,”Comparative Analysis for Prediction of Success Movies”, International Conference on Communication and Electronics Systems

[11] Pawar, R., Ghumbre, S., & Deshmukh, R,” Visual Similarity Using Convolution Neural Network over Textual Similarity in Content- Based Recommender System”, International Journal of Advanced Science and Technology

[12] Pawar, R., Ghumbre, S., & Deshmukh, R,”A Hybrid Approach towards Improving Performance of Recommender System Using Matrix Factorization Techniques”,International Journal of Future Generation Communication and Networking

[13] (ICCES) S. Gogineni and A. Pimpalshende, "Predicting IMDB Movie Rating Using Deep Learning," 2020 5th International Conference on Communication and Electronics Systems

[14] Warda Ruheen Bristi, Zakia Zaman, Nishat Sultana “Predicting_IMDb_Rating_of_Movies_by_Machine_Learning_Techniques”, International Conference on Computing, Communication and Networking Technologies (ICCCNT)