**MOVIE SUCCESS PREDICTION USING MACHINE LEARNING TECHNIQUES**

1 SWETHA S\*, 2 TEJASWINI M, SONIYA PRIYATHARSHINI G

1,2 Students, Department of Computer science and Engineering, Dr.M.G.R Educational and Research Institute of Technology, Madhuravoyal, chennai-95,TamilNadu, India.

3, Faculty, Department of Computer science and Engineering, Dr.M.G.R Educational and Research Institute of Technology, Madhuravoyal, chennai-95,TamilNadu, India.

### Corresponding Author, Mail ID: \* swethadsai0@gmail.com

***Abstract:***

*Predicting a movie’s success is crucial now as people tend to have less time spending for entertainment in their busy schedule. there are many methods to predict the success of a movie, the goal is to predict the success of a pre-released movie. A predictive model is trained and tested using IMDb and social media sources. The prediction result can be hit, average, and flop so that audience would act according to it. This paper explains the movie success prediction using machine learning techniques such as random forest classifier, logistic regression, decision tree and support vector machine and their accuracy. Also, the paper describes about the topics: (i) sentiment analysis (ii) IMDb (iii) feature selection. The purpose of this paper to describe the various techniques to predict the success and to understand clearly about those techniques.*

**Keywords:**

Movie success prediction, predictive analysis, Machine learning, accuracy, IMDb

**1. Introduction:**

A movie's revenue is determined by a variety of factors, including the cast's acting in the film, the budget for the film's production, film critics' reviews, the film's rating, the release year, and so on. However, by analysing previous films' revenues, we can build a model that can help us predict the expected revenue for a specific film. Such a forecast might be extremely helpful for the

film companies that will be producing the film, allowing them to plan different expenses such as artist compensations, movie advertising, promotions in various cities, and so on.

This project will generate predictive models based on machine learning tools and statistic analytical models to classify movies into hit, average, or flop.

By picking the appropriate components like IMDb ratings, IMDb meta scores, IMDb vote counts, Rotten Tomatoes tomatometer, actors and directors' social fan followings, Wikipedia views, trailer views, etc. are just a few variables that might help forecast a movie's success mainly IMDb sources. Whatever the film industry, IMDb data will be appropriate to use as the dataset of IMDb contains a wide range of multilingual movies for better prediction. Various attempts were made and some of the techniques yielded excellent accuracy rate. Other than machine learning, Data mining, business analytics techniques can also use to predict the success rate of a pre released movie. This paper describes and compare few ML techniques like decision tree, random forest, SVM (support vector machine) and logistic regression. A developing technology called machine learning makes it possible for computers to learn autonomously from historical data. Machine learning employs a variety of methods to create mathematical models and make predictions using knowledge or historical data.

In the context of uncertainty, supervised machine learning creates a model that makes predictions using data. A supervised learning technique trains a model to produce accurate predictions for the reaction to incoming data using a known set of input information and recognized response to the output.

This paper guides you to the ML techniques mentioned above and gives a clear idea how it works. important source for prediction, IMDb is explained briefly along with the process of data extraction and sentiment analysis. also, some of the related works on this topic is discussed further.

**2. Literature Survey:**

A mathematical model was developed by Javaria Ahmad et al. to determine the likelihood that upcoming films will be successful. The purpose of this numerical method was to give investors’ confidence in their investments by making an accurate prediction of success. This analysis made use of simulation data, and it cleaned, integrated, and altered hundreds of records. For each series of analyses, a randomly selected subset of this dataset was used. With the purpose of predicting the success of the movie, several criteria were examined. Budget, cast, director, producer, setting locales, plot writer, movie release date, concurrent movie releases, music, release site, and target demographic were a few of the variables. Their suggested model was an algorithm that used X2 analysis to discover correlations between these distinct attributes. With a degree of freedom of 8, the predicted frequency for the connection between genres and reviews was calculated to be 64.39 [10].

While the correlation and degree of freedom between actors and genre were discovered to be 20.6 and 1 respectively, the correlation and degree of freedom between actors and ratings were found to be 11.57 and 1.

In order to do multi-class classification using a variety of machine learning techniques, Nahid Quader used their own movie dataset [6]. The primary objective of this research is to compare several machine learning techniques.

The Scikit Learn Python module was used to implement most of the algorithms. The seven machine learning tools that they employed are: Multilayer Perceptron Neural Network, Support Vector Machine, Gaussian Naive Bayes, Adaboost, Random Forest, and Logistics Regression [10]. Of all these machine learning techniques, the Multilayer Perceptron Neural Network has the highest accuracy (58.53%). MLP is an excellent prediction model because it can handle very complicated data patterns that other models are unable to recognise.

Every one of these methods forecast a movie's approximate net profit value by analysing historical data from various sources such as IMDb, Rotten Tomatoes, Box Office Mojo, and Meta Critic.

The methodology estimates a film box office profit for all seven approaches based on a few pre-released and post-released elements. Based on our own dataset of 755 movies, this study evaluates the performance of all seven machine learning algorithms. Among the seven algorithms utilised, the Multilayer Perceptron Neural Network produces the best results.

There were many that could not be covered here and these are some of the works that has been done regarding this topic.

**3. Proposed System:**

The initial step in this work is to collect, extract and transform the data that is relevant for the prediction. So, collection of data is the first step for the process. Most of the data used here is of IMDb.

The steps are as follows:

*(i) Dataset Gathering*: IMDb will be utilised to collect the initial dataset. We chose movies that are in English and have reviews that are likewise in English in the hopes of generating a dataset for exact forecasting also deleted movies that did not include any box office data.

*(ii) Data Pre-processing:* Because of their large size and likely origin from various, heterogeneous sources, the data we gathered are very sensitive to noise, missing, and inconsistent data. The primary issue with datasets was incomplete fields. To address the missing field issue, A strategy is used to employ a central tendency scale for the characteristic. As measures of central tendency, mean and median is employed. Then, duplicate items were deleted [2].

*(iii) Feature Selection:* To choose whichever elements from a pre-processed datasets can be included in a model; feature selection is carried out. Trial and error are a key component of this process, although there are formal approaches to address it, such like Pearson correlation and the Variance Inflation Factor (VIF). Trial and error are used to establish the subset of qualities that must be included in the VIF analysis.

*(iv) Data integration and transformation:* Data gathered from distinct sources were combined into a single database. Integrated data are changed in this step to make the regression procedure more efficient. The dataset contains both nominal and numerical properties, but in the regression process, all attributes should be numerical. To convert matching nominal qualities to numerical, a central tendency measure of Box office income is applied.

*(v) Selecting feature subset:* We search for correlations between factors and the target variable to obtain the best feature subset, removing the top-most irrelevant attribute after each iteration and finding the poorest attribute for each iteration.

This is an overall general explanation of how every technique processes the data. After this data collection and transformation, the dataset is trained and tested to obtain the desired accuracy. Generally, accuracy from 70% to 90% are considered the best and attaining 90% is not that easy to achieve but researches are made to overcome that [9].

ALGORITHM:

STEP 1: Acquire the dataset suitable for the prediction.

STEP 2:

* Data pre-processing
* Data cleaning
* Fill in the missing values.

STEP 3: Drawing out important features based on the feature elimination technique.

STEP 4: Apply machine learning algorithms (Decision Trees, Random Forest, Regression, SVM) and calculate the prediction score.

STEP 5: Implement models with scoring tools to carry out the prediction score and feasibility of the following model.

STEP 6: Use the model for business decision-making or forecast using the best applicable instrument.

The Data processing method is depicted in flowchart below:

Figure 1: Flowchart depicting data processing

**4. IMDb:**

IMDb stands for “Internet Movie Database” and it is an online database system containing information of movies, TV shows, Dramas, TV series, podcasts etc… which is considered as a standard platform for reviewing the shows [2]. People leave critical reviews and ratings about the movie which can be helpful for other users who wanted to watch a quality content. It also contains the statistics of the actors, director, production team, media popularity and so on. it is available as a website and the iOS, Android, and Amazon Kindle fire devices. This paper uses IMDb ratings and meta scores mainly for the dataset.

**5. TECHNIQUES OF MOVIE SUCCESS PREDICTION:**

*5.1 Decision Tree:*

A decision tree divides the nodes of a decisional tree into all possible splits, and it chooses the split that produces the most homogeneous sub-nodes.

A movie's success can be determined in a variety of ways, although it is often gauged by its box office performance or critical reception.

A decision tree algorithm would first need to be trained on a dataset of historical movie data, which would contain details like the genre, budget, cast, release date, and other pertinent aspects that may affect a movie's success. The dataset would also contain details about each film's performance, such as box office receipts or critic reviews [9].

The dataset would then be used by the decision tree algorithm to discover patterns and connections between the various elements and the box office success of the films. It would produce a tree-like model that divides the data according to the most important considerations, like the genre or the price range. Each split would separate the data into subsets with related characteristics and forecast the commercial performance of the films in each subset based on prior experience [5].

The decision tree model can be used to forecast the success of new films based on its traits once it has been trained. The algorithm would use the traits of the upcoming film as input and iterate through the decision tree to forecast how successful the film will be.

*5.2 Random Forest:*

Leo Breiman and Adele Cutler are the creators of the widely used machine learning technique known as random forest, which mixes the output of various decision trees to produce a single outcome. Its widespread use is motivated by its adaptability and usability because it can solve classification and regression issues.

By choosing a dataset at random from the pool of available datasets, creating a decision tree, and then voting on the outcome of the predictions, the suggested methodology increases the success rate of movie success [8]. Using parameters like Score, accuracy, precision, recall value, F1 score, mean absolute error, and mean square error, the performance of the proposed methodology is evaluated [4]. the accuracy of Random Forest is higher compared to some techniques like BernoulliNB and GaussianNB.

*5.3 Logistic Regression:*

Based on the examination of independent variables, logistic regression produces a dependent variable as an output that is dichotomous and only accepts the values 0 or 1 [1]. The objective is to forecast the film's commercial viability if it is released. Before selecting variables and factors, a few experiments were conducted on which these judgements were based. We first analyse the issue at hand to determine whether logistic regression or linear regression should be employed. Continuous values, which may include more than one, are the result of a linear regression. In contrast to using a logistic regression to determine the likelihood that a film would be a hit, a linear regression could be applied to determine a film's sales.

To determine if the variables we are considering for a logistic regression are adequate, we do a model fitting test by applying log likelihood. A value is predicted via classification, which produces a few little discrete values. The sigmoid function, which provides us with the chance at a certain moment, is used to predict using logistic regression. Calculations in elementary mathematics are used to get the sigmoid [11].

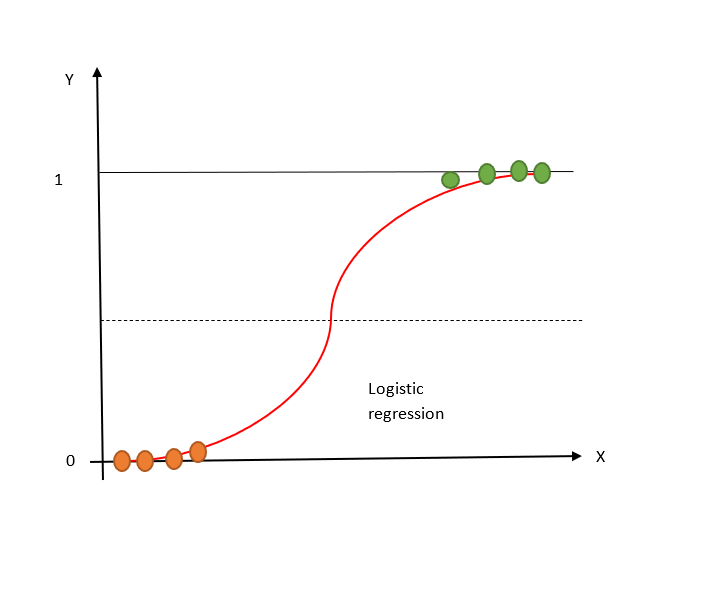
As seen below, h provides the value of the hypothesis, which ranges from 0 to 1, and g(z) computes the sigmoid of z,

Were,

h = sigmoid(X\*theta).

Finding the ideal model to explain the connection between dependent variable and independent factors was the goal of employing logistic regression. It produced coefficients to forecast a signiﬁcant changes of the likelihood that a movie will succeed if it was released.

Smaller training sets have a clear distinction between critic reviews, awards, and genre, while larger sets have a larger gap between cost and gradient. This may cause crucial difference in predicting the success of a movie and to be noted while training the model.

Figure 2: Logistic regression diagram

*5.4 Support Vector Machine:*

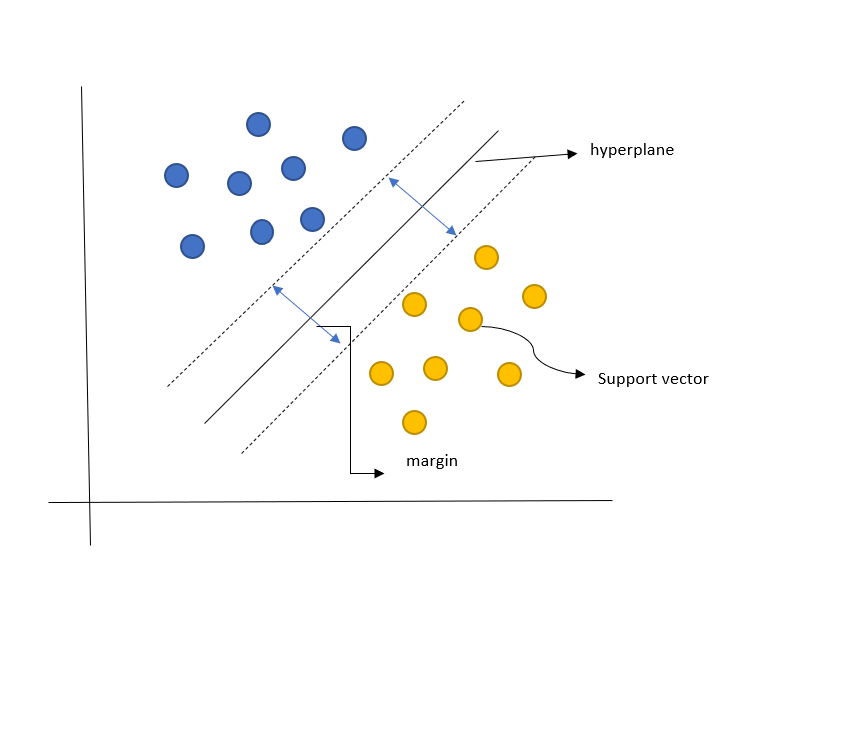
It is a supervised machine learning model which employs classification techniques to solve two-group classification problems. Support vectors are data points that are closer to the hyperplane and have an impact on the hyperplane's position and orientation. By utilising these support vectors, we increase the classifier's margin [12].

An example of a classification job that may be performed using Support Vector Machines (SVM) is forecasting the box office performance of a film.

We would need to initially gather a dataset of movies with different attributes, such as the movie's genre, cast, director, budget, release date, and more, in order to forecast the success of a film using SVM. Each movie in the dataset would also need to be classified as either successful or unsuccessful depending on several factors, such as how well it did at the box office or how well it was received by critics.

We can train an SVM model on a labelled dataset once we have one. By locating a hyperplane that optimises the margin between the two classes, the SVM model will be taught to distinguish between successful and failed films. Every new film may be categorised based on which side of the decision boundary it lies on since the hyperplane serves as a decision boundary.

In order to categorise a new film, its attributes would be extracted and fed into the trained SVM model, which would then provide a prediction about the film's likelihood of success based on the decision boundary it had learnt during training.

Figure 3: SVM representation graph

**6. Sentiment Analysis:**

Sentiment analysis is the process of identifying and extracting the sentiment or emotion expressed in a piece of text using natural language processing (NLP) and machine learning techniques. Finding out whether a text has a good, negative, or neutral sentiment is the aim of sentiment analysis.

Large amounts of text data, including posts on social media, consumer reviews, news stories, and feedback forms, can be analysed using sentiment analysis. Understanding public attitude towards a good, service, or topic is widely employed in a variety of industries, including marketing, customer service, and politics. In machine learning methods, it is employed. it entails using a labelled dataset to train a machine learning model to predict the sentiment of new text.

This approach is useful in this prediction as we can get information and analytics from social media reviews and comments about the particular actor, director, production crew or the trailer of a pre-released movie. This helps in increasing the accuracy rate as it is a real time data that we can get directly from the masses.

**7. Result:**

From analysing, there are some points to be noted before drawing a conclusion. First, selecting an algorithm according your preferences is important. By considering the size of the dataset and sources, we should choose an algorithm that produces higher accuracy. But these ML techniques yield better accuracy than some numerical methods. From above points, we can see that Support Vector Machines have a substantially higher computational complexity than Random Forests. As a result, when the size of the training data is larger, training an SVM will take longer than training a Random Forest. This must be considered when choosing the algorithm. Also, larger dataset will take longer time in logistic regression also causes large gap between cost and gradient. Random Forest and SVM gives higher accuracy than other methods by using these huge sources. The accuracy is found out to be approximately 80%.

**8. Conclusion:**

Machine learning is a powerful AI tool used for various purposes. Use of ML in prediction is indispensable and gives better results in our case, movie success prediction. IMDb is the main database system implied here and the data extraction process paves a clearer way to process the filtered data. The techniques which are used here are SVM, Random Forest, Decision tree and logistic regression. Random Forest works better on larger datasets and model than others. Logistic regression can be used for smaller datasets but the accuracy will be lower. Sentiment analysis is NLP tool which converts emotions to text for processing and it is useful in prediction of the movie success.

**9. Future work:**

This paper explained about Machine Learning techniques to predict success of a movie and in future work, it will be ideal to consider more real time factors such as YouTube reviews, ground data about the movie actors and also data from OTT platforms such as Amazon Prime Video, Hulu, Disney+ for a more accurate result. Wider research should be done as most of the prediction is targeted on US or their regional area movies. Adding more language movies in the dataset would be useful to diverse public.

**References:**

[1] R, Ganesh & Kalaiarasi, S.. (2022). Strange Approach of Movie Rating Prediction Using Logistic Regression Comparing to Gaussian Naive Bayes Algorithm. 10.3233/APC220041.

[2] Vr, Nithin & Pb, Sarath. (2014). Predicting Movie Success Based on IMDB Data.

[3] Olubukola D.A., Stephen O.M., Funmilayo A.K., Ayokunle O., Oyebola A., Oduroye A., Wumi A., Yaw M. (2021), Movie Success Prediction Using Data Mining. British Journal of Computer, Networking and Information Technology 4(2), 22-30. DOI: 10.52589/BJCNITCQOCIREC.

[4] Lee, S., KC, B., & Choeh, J. Y. (2020). Comparing performance of ensemble methods in predicting movie box office revenue. *Heliyon*, *6*(6). <https://doi.org/10.1016/j.heliyon.2020.e04260>

[5] Burgos, Maria & Campanario, Maria Luisa & Lara, Juan & Lizcano, David. (2015). Using Decision Trees to Characterize and Predict Movie Profitability on the US Market. Lecture Notes in Engineering and Computer Science. 1. 274-279.

[6] N. Quader, M. O. Gani, D. Chaki and M. H. Ali, "A machine learning approach to predict movie box-office success," 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2017, pp. 1-7, doi: 10.1109/ICCITECHN.2017.8281839.

[7] Kumar, N. (n.d.). *Performance Analysis of M X N Equalizer Based Minimum Mean Square Error (MMSE) Receiver for MIMO Wireless Channel*. International Journal of Computer Applications - IJCA. https://www.ijcaonline.org/archives/volume183/number44/dissanayake-2021-ijca-921847

[8] Chintu Kumar. MOVIE SUCCESS PREDICTION USING RANDOM FOREST (RF) CLASSIFIER Research Scholar, Sri Satya Sai University of Technology and Medical Sciences, Sehore (M.P.), India

[9] Analyzing and Predicting The Success Of Box Office Collection Of A Movie Using Machine Learning Sanyam Jatale , Rohan Moze , Varsha Khandekar ,Shubham Jain and Sanket Mokate Students, Department of Information Technology Smt. Kashibai Navale College of Engineering, Pune, India. <https://ijarsct.co.in/Paper1392.pdf>

[10] Movie Success Prediction Using Data Mining 1Antara Upadhyay, Nivedita Kamath, Shalin Shanghavi, Tanisha Mandvikar, 5Pranali Wagh B.E Student, 5\*Asst. Professor Department of Information Technology, Shah & Anchor Kutchhi Engineering College, Mumbai, India

<https://ijedr.org/papers/IJEDR1804038.pdf>

[11] Aashya Khanduja Logistic Regression for Predicting Movie’s Success

<https://openreview.net/pdf?id=rylE06GuyX>

[12] Dwi Rantini, Rosyida Inas, dan Santi Wulan Purnami Departemen Statistika, Fakultas Matematika, Komputasi dan Sains Data, Institut Teknologi Sepuluh Nopember Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia. Predicting Popularity of Movie UsingSupport Vector Machines

<https://media.neliti.com/media/publications/323492-predicting-popularity-of-movie-using-sup-4cc89b30.pdf>