2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. Sep 3-4, 2021

2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) | 978-1-6654-1703-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICRITO51393.2021.9596138

Movie Success Prediction Using Naïve Bayes, Logistic Regression and Support Vector Machine

Rachaell Nihalaani
Department of Computer Engineering, Thadomal Shahani Engineering College
Mumbai, India
rachaellnihalaani@gmail.com

Apoorva Shete
Department of Electronics and Telecommunication Engineering, Thadomal Shahani Engineering College
Mumbai, India
shete.apoorva48@gmail.com

Darakshan Khan
Department of Computer Engineering, Thadomal Shahani Engineering College
Mumbai, India
darakshan.khan@thadomal.org

***Abstract*—The entertainment industry is a rapidly growing billion-dollar industry. With new milestones being reached almost every day, this industry has proved itself to be a very profitable business, if done correctly. Since huge investments are involved in the production and making of movies, both in terms of time and money, it would only make sense to try to predict the outcome beforehand. In an attempt to tackle this problem, we have built a model that predicts whether or not a movie can be called a success. The model compares the performance of three machine learning algorithms i.e. Naive Bayes, Logistic Regression, and Support Vector Machine (SVM), over two different datasets, to observe which performs better. We have illustrated the model, as well as its results, findings, and observations in this literature.**

Keywords—Machine Learning, Prediction Model, Movie Success Prediction, Support Vector Machine, Naive Bayes, Logistic Regression

I. Introduction

The entertainment sector is the leading and most influential sector in this day and age. Movies have been known to have an extremely strong hold on an individual's mind, whether it is simply to entertain oneself or to influence and challenge one's thought process. What started as simple means to entertain people, has quickly grown into a billion-dollar business worldwide. In this fast-paced and extremely competitive industry, movies are constantly being produced and released, and naturally, this would require huge investments both in terms of time and money. The prediction of how a movie will fare at the box office could be a game-changer in the movie sector. It would be quite beneficial for shareholders to have access to a model that predicts the success of a movie beforehand, given certain attributes. It would also prove to be highly beneficial to production houses so as to enable them to effectively plan the publicity and promotions, which entails tremendous expenses. Machine learning is crucial as it provides multifarious functions and its implausible versatility and source fixes to tackle convoluted issues quickly and efficiently. In this paper, we have aimed to predict the success of a motion picture as accurately as possible using a few ML algorithms.

This paper discusses and compares three machine learning algorithms, Logistic Regression (LR), Naive Bayes (NB), and Support Vector Machine (SVM). Our research objective is to perform analysis and build a model that predicts whether or not a movie would be considered successful, based on certain parameters. In order to efficiently evaluate our models, we train and test them on two drastically different datasets. The performance of these models will be assessed and compared with the help of certain performance measure indices, such as accuracy. The results and conclusion of this paper will enable future researchers to select the most effective model, out of the three, that gives the best performance for future applications.

The following section reviews the study of related literature. Section 3 describes the two datasets used in this paper. Section 4 provides a brief about the machine learning algorithms used. Section 5 illustrates the methodology followed in building this model. Section 6 illustrates the results and Section 7 draws the conclusion to the paper.

II. Literature Review

Former research accurately predicts movie success, using various algorithms and techniques. In their paper, Nahid Quader et al. [1] illustrates a model to predict roughly how a movie will fare on the basis of profits, past data analysis including release features prior and after the release from different sources and uses SVM, Neural Network, and NLP algorithms, of which Neural Networks give good. This research also illustrates the budget, IMDb votes and screens were crucial aspects to determine the box office success of the movie. Ashutosh Kanitkar, in his research [2], reviewed regression techniques such as Linear Regression, Polynomial Regression, Logistic Regression, Artificial Neural Networks, K Nearest Neighbours, Random Forest, Decision Tree, SVM, and Naive Bayes. This research aimed to envision Bollywood movies' revenues. It also used multiclass classification methods on its dataset. In the paper [3], researchers provided IMDb and predicted the IMDb score - for knowing a movie's success before it comes to the box office using machine learning, rather than being solely dependent on critics and reviews. It used Support Vector Machine, Random Forest, AdaBoost, Gradient Boost and K Nearest Neighbours, out of which Random Forest gave the highest accuracy of 61%. In the paper [4], Prakash Duraisamy et al. developed a model that predicts if the upcoming movies will be a hit or a failure based on multiple features, such as the fame and credibility of the movie cast, the production house, and also the audience. Paper [5] predicts the success rate of movies based on different features selected from the dataset and using various ML algorithms such as Random Forest, DecisionTree, K-NearestNeighbours (KNN), NLP, XGBoost Classifier and Deep Neural Network on the IMDB dataset. This paper concludes that XGBoost was the best classifier for predicting success rates of movies.

This literature review and the developed mathematical model provided us with extremely valuable insight as to how we can go about analyzing our dataset and the features that we must take into consideration in order to build our model.

III. Dataset Description

Data set collection is the first and foremost step. In order to train our model efficiently and achieve the best possible results, we have used two datasets to both train as well as test the chosen algorithms. The first dataset used is the IMDb movies' dataset [6]. This dataset was collected from kaggle.com. It consists of different genres of movies, which can be used for the prediction of the success of moviess. IMDb is a very famous, reliable, and authoritative souce to find reviews and ratings of the latest movies and TV shows. This dataset contains 840 entries and 8 attributes. In Table 1 below, all the attributes along with their descriptions are shown.

The dataset has an uneven distribution of the successes. Out of all the entries, 689 entries are labeled as 0 (failure) and 149 entries are labeled as 1 (success). This distribution of outcomes has been shown below in Fig. 1.

Here in Dataset 1, 75% of the data has been used for training, whereas the rest of the 25% of the dataset has been used for testing our model.

TABLE I. Attributes For IMDB Dataset

Attributes	**What it describes**	**Data type**
Title	Title (name) of the movie	String
Genre	Genre is a stylistic or thematic category for movies based on similarities in the narrative elements.	String
Description	Gives information about the storyline of a movie.	String
Director	Name of movie's director.	String
Actors	Names of the actors in the movie.	String
Year	Year in which the movie was released.	Numeric
Runtime	Runtime is the time between the starting of the movie up to the end of the credits scene.	Numeric
Rating	The IMDb rating of a movie is the average of all user votes.	Numeric
Votes	Users can cast votes, on the scale of 1-10, on any movie in IMDb.	Numeric
Revenue	Revenue is the amount of money raised by the movie.	Numeric
Metascore	Metascore is the weighted average of the scores assigned to critics' movie reviews.	Numeric
Success	Indicates if the movie has been considered a success or not.	Binary

The second dataset used in this paper is obtained from GitHub [7]. It has 5043 entries and 28 attributes. Of these, only a few attributes are filtered and selected. The multiple features include Facebook likes of directors and actors, movie duration, budget and gross collected, IMDb score, etc. There are few duplicate rows, thus the data is not perfect and needs to be cleaned. Here in Dataset 2, 60% of the dataset has been used for training our model, whereas 40% of the data has been used for testing.

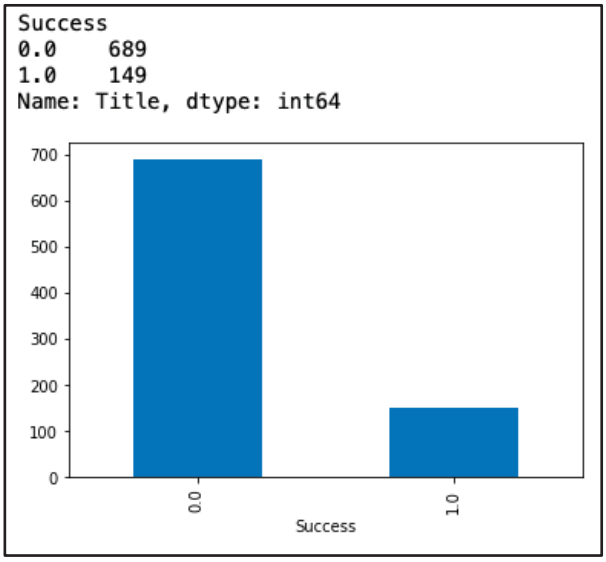

Fig. 1. Data Distribution of Successes and Failures in Dataset 1

IV. Classifiers

A. Logistic Regression

This is a machine learning algorithm that is primarily used for predictive analysis and is built on the probability concept. Logistic Regression is a go-to method for binary classification problems. It is a linear regression model which does not use a linear function, and instead makes use of the sigmoid or logistic function, which is a cost function of higher complexity. Equation (1) shows the definition of the logistic sigmoid function.

$$f(x) = \frac{1}{1+e^{-x}} \tag{1}$$

Sigmoid functions cannot be represented by linear functions as they can take values less than 0 or greater than 1, and there is no possibility of this as per the hypothesis of logistic regression expectation, given by (2).

$$0 \leq h_\theta \leq 1 \tag{2}$$

Logistic regression is most commonly used with a categorical target variable and the data observed has a binary output, i.e. it belongs to either one class or another, i.e. 0 or 1. This algorithm basically gives the conditional probability that y belongs to a particular class given X input features. It has been observed to work well for smaller datasets [8].

B. Support Vector Machine

Support Vector Machine (SVM) is a coherent and simple yet highly preferred machine learning algorithm that can find its use in solving both classifications as well as regression problems. SVM is known for using little computational power to produce considerable accuracy. This algorithm aims to find such a hyperplane that distinctly classifies the data points in an N-dimensional space, where N is the number of features. A hyperplane is a decision boundary that classifies data points such that those on one side of the hyperplane belong to one

class, while those on the other side belong to another class. A number of hyperplanes are possible to separate any two given classes. SVM finds such a plane that the distance between data points of both classes is maximum. This is called maximum margin and can be determined using the data points closest to the hyperplane. Such data points are called support vectors, and they influence the orientation and position of the hyperplane. The idea behind maximizing margin distance is that it adds to the expectation that test data points can be classified more accurately and confidently [9].

C. Naïve Bayes

This well-known classification technique is based on Bayes' theorem. It assumes that predictors are independent. It also assumes that whether or not one feature is present in a class is not at all related to whether another feature is present in it. Bayes' theorem helps to calculate the posterior (updated) probability as shown in (3).

$$P(c|x) = \frac{P(x|c) \times P(c)}{P(x)} \quad (3)$$

Where,

- P(c|x) represents the updates probability of class (c, target) given predictor (x, attributes).
- P(x|c) represents the probability of the predictor given class.
- P(c) is the initial probability of class.
- P(x) is the initial probability of the predictor.

The Naïve Bayes model is easy to build and particularly useful for very large datasets [10].

V. Methodology

This section explains the general process of binary classification and offers a detailed explanation of the flow of the experiment performed. The first step is data collection i.e. to find a dataset suitable for analysis. Two different datasets have been used for training and testing our model, to achieve the best accuracies and analysis. Relevant attributes from both datasets have been selected so that they can be used to analyze the chosen machine learning algorithms, after which their prediction performance is evaluated. Lastly, we acquired and configured a set of suitable tools for the comparison of all the algorithms used.

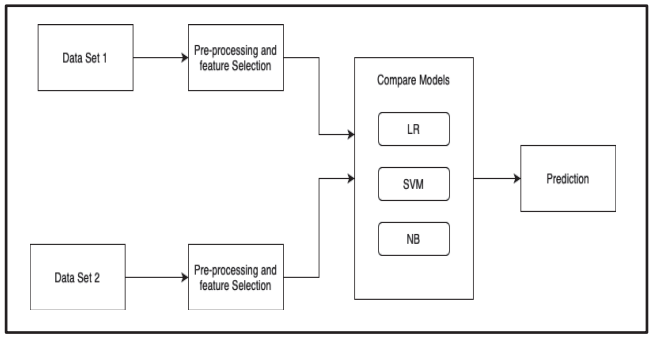

Fig. 2. Flowchart for our experimental process

The two chosen data sets were first pre-processed. The datasets collected are not suitable to be used to train the model right away. Several steps are implemented in order to clean the data so that it is appropriate for model training. The larger dataset [7] contained many duplicate rows which needed to be eliminated. The missing values in the dataset needed to be fixed, this was done by filling None in any NaN category feature and 0.0 or mean in any numeric feature. A few news columns were added to the data set to store some of the features as binary data. After that, a new dataset was defined to store all the numeric features to the cleaned dataset. The X labels, using the larger dataset were set as all the extracted features and data, whereas, the label was set as the target IMDb score. Support Vector Machine, Logistic Regression and Naïve Bayes algorithms are then implemented on this data. Evaluation of the performance of this model has been done on the basis of various evaluation metrics.

The classification outcomes are [11]:

- True Positive (TP): Number of correctly predicted positive classes.
- True Negative (TN): Number of correctly predicted negative classes.
- False Negative (FN): Number of incorrectly predicted negative classes.
- False Positive (FP): Number of incorrectly predicted negative classes.

The evaluation metrics and their formulas are:

$$\text{Accuracy} = \frac{|TP| + |TN|}{|TP|+|TN|+|FN|+|FN|} \quad (4)$$

$$\text{Precision} = \frac{|TP|}{|TP|+|FP|} \quad (5)$$

$$\text{Recall} = \frac{|TP|}{|TP|+|FN|} \quad (6)$$

$$\text{F1 Score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (7)$$

Accuracy (4) is the number of correct classifications divided by the number of all possible cases. It is defined as the percentage of correct predictions for the test data [12]. Precision (5) is the number of true positives divided by the number of predictive positives [13]. Recall or Sensitivity (6) is the number of true positives divided by the number of actual positives. F1 Score (7) is used to convey the balance between precision and recall. These metrics are used in evaluating the performance of classifiers from different perspectives. A flowchart of the methodology followed for the implementation of this model has been shown in Fig. 2.

VI. Results and Analysis

Careful observation of this model on two datasets, Dataset 1 and Dataset 2, and three algorithms, Logistic Regression, Naive Bayes and Support Vector Machine, has given us the following results, as shown in Tables 2 and 3.

TABLE II. Evaluation Metrics For Dataset 1

Algorithm Used	Accuracy	Precision	Recall
Logistic Regression	0.90	0.71	0.72
Support Vector Machine	0.91	0.79	0.72
Naïve Bayes	0.83	0.52	0.78

It is clear from Table 2. that the best accuracy for Dataset 1 is obtained by using the Support Vector Machine algorithm,

followed by Logistic Regression and Naïve Bayes respectively.

TABLE III. EVALUATION METRICS FOR DATASET 2

Algorithm Used	Accuracy	Precision	Recall
Logistic Regression	1.0	1.0	1.0
Support Vector Machine	0.99	0.99	1.0
Naïve Bayes	1.0	1.0	1.0

For the larger dataset, the accuracy has been significantly improved, as shown in Table 3, since there is more data for the model to train itself. Logistic Regression and Naïve Bayes have given the best accuracies, and the accuracy of Support Vector Machine lies just below the previous two algorithms.

An essential part of the exploratory analysis is finding the correlation between the attributes of our dataset. We have used heat maps (Fig.3 and Fig. 4) to visualize attribute correlation, as in such cases, graphics are easier to understand than tabulated data.

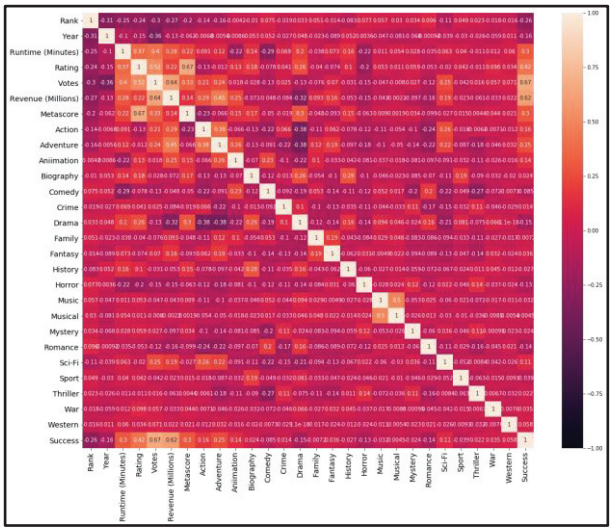

Fig. 3. Correlation Heatmap for Dataset 1

The following observations are made from the heatmap for Dataset 1 shown in Fig. 3:

- The ‘rating’ has a strong positive correlation with the "metascore" and “votes", meaning a higher rated movie has more positive reviews and has a higher number of user votes. This suggests that both users and critics have a considerable influence on the rating of a movie.
- The ‘votes’ have a strong positive correlation with the "revenue", which indicates that a better-voted movie collects more revenue. We might even infer that revenue is influenced by votes in the sense that one may consider votes on a movie before deciding to purchase a ticket.

Similarly, the correlation heatmap for Dataset 2 has been shown in Fig. 4.

It describes the degree of correlation between various attributes whether it is strong or weak which indicates that critics have a large impact on its fame.

The results of both the datasets have been compared in the bar graphs below. Refer Fig. 5 and Fig.6.

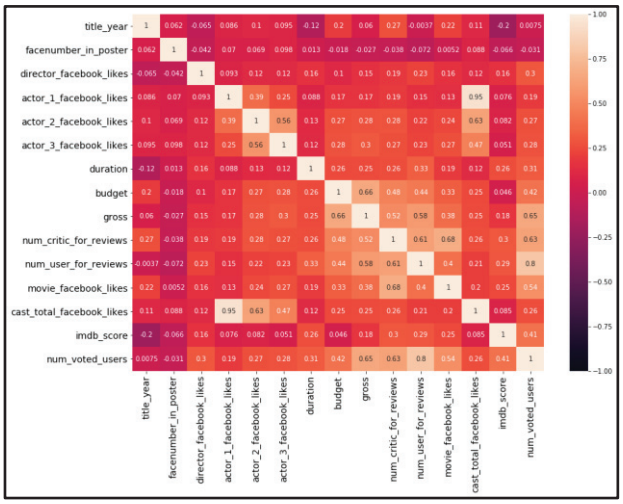

Fig. 4. Correlation Heatmap for Dataset 2

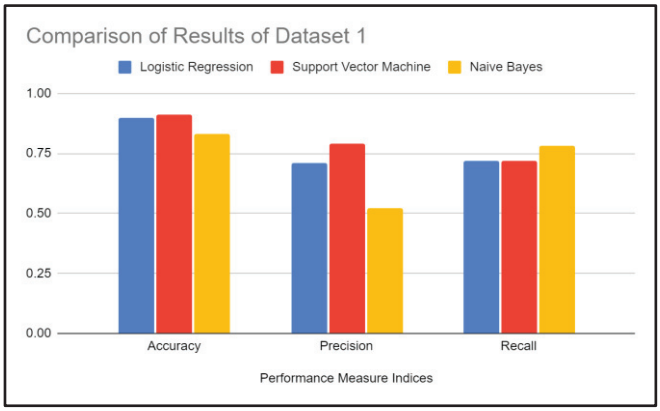

Fig. 5. Comparison for Dataset 2

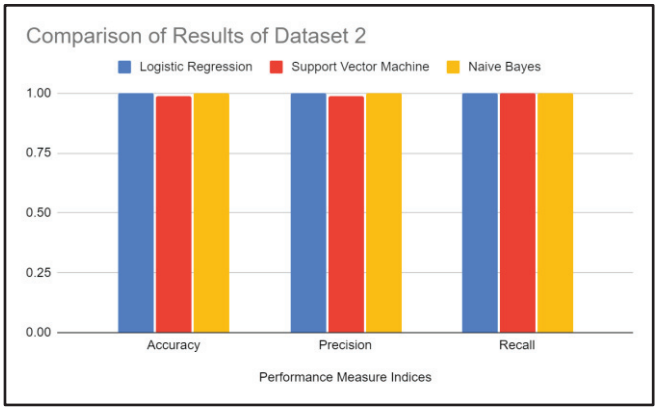

Fig. 6. Comparison for Dataset 2

VII. CONCLUSIONS AND FUTURE SCOPE

In today’s world where movies lie at the heart of the entertainment society, many young and upcoming producers and artists risk a lot of money when they decide to make an idea into a movie. Prediction of a movie’s success rate before it has been released can be of great help to the people who are investing in making motion pictures. In this paper, we have successfully compared a few ML algorithms to achieve the best accuracy for predicting a movie’s success rate. It can be seen from the results, the best algorithm for movie success prediction is Logistic Regression which has **90%** and **100%** accuracies for Dataset 1 and Dataset 2 respectively, giving the highest accuracy for both the datasets used. SVM has proven to be the second-best algorithm, followed by NB.

There is a significant improvement in the success rate of our model as compared to the past research. However, it will give much better accuracy if there is more data used to train this model. In the future, additional data such as social media comments about the movie's plot can also be utilized for better accuracy.

A movie's success rate can be predicted before it has been released. Once the movie is released, the results of how the movie has fared at the box office can be collected and compared, and analyzed. The results obtained from this analysis can be fed to this ML model to give it better training. Thus, the prediction rate of this model can be significantly improved in the future.

ACKNOWLEDGMENT

We would like to express our gratitude to professor Darakshan Khan for her immense support and guidance. We would also like to express our gratitude to our Head of Department, Dr. Tanuja Sarode, and our Principal, Dr. G.T. Thampi.

REFERENCES

[1] 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), 2017, pp. 1-7, doi: 10.1109/ICCITECHN.2017.8281839.

[2] A. Kanitkar, "Bollywood Movie Success Prediction using Machine Learning Algorithms," 2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C), 2018, pp. 1-4, doi: 10.1109/CIMCA.2018.8739693.

[3] R. Dhir and A. Raj, "Movie Success Prediction using Machine Learning Algorithms and their Comparison," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), 2018, pp. 385-390, doi: 10.1109/ICSCCC.2018.8703320.

[4] J. Ahmad, P. Duraisamy, A. Yousef and B. Buckles, "Movie success prediction using data mining," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017, pp. 1-4, doi: 10.1109/ICCCNT.2017.8204173.

[5] N. Darapaneni et al., "Movie Success Prediction Using ML," 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2020, pp. 0869-0874, doi: 10.1109/UEMCON51285.2020.9298145.

[6] "IMDB-Movie-Data.csv, Dataset 1" https://github.com/siddrao/Movie-Success-Prediction/blob/master/AIPro/IMDB-Movie-Data.csv

[7] "Check.csv, Dataset 2" https://github.com/SnirZarchi/Movie-Success-Prediction/blob/master/Dataset/check.csv

[8] " Logistic Regression." towardsdatascience.com. https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

[9] "Support Vector Machine" towardsdatascience.com. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[1] " Naive Bayes Algorithm " analyticsvidhya.com. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[2] "Classification Outcomes" https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative

[3] "Accuracy" https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/

[4] "Precision and recall." wikipedia.org. https://en.wikipedia.org/wiki/Precision_and_recall