iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** II **Month of publication:** February 2020

DOI: http://doi.org/10.22214/ijraset.2020.2102

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 8 Issue II Feb 2020- Available at www.ijraset.com

Prediction of Movie Performance using Machine Learning Algorithms

Shubham Pawar[1], Saurabh Shinde[2], Aditi Phepale[3], Akshay Sonawane[4], Parnika Shinde[5], Yogesh Deshmukh[6]
[1, 2, 3, 4, 5]*Information Technology (BE), Sanjivani College of Engineering, Kopargaon*
[6]*Assistant Professor, Department of Information Technology, Sanjivani College of Engineering, Kopargaon*

Abstract: ***Movies play an important role in our everyday life. Film industry involves huge investment in terms of money, time, efforts and people. The number of movies produced in the world is growing at an exponential rate and success rate of movie is of utmost importance since billions of dollars are invested in the making of each of these movies. In such a scenario, prior knowledge about the success or failure of a particular movie and what factor affect the movie success will benefit the production houses since these predictions will give them a fair idea of how to go about with the advertising and campaigning, which itself is an expensive affair altogether. So, the prediction of the success of a movie is very essential to the film industry. In this proposed system, we give our detailed analysis of the Internet Movie Database (IMDb) and predict the IMDb score. This database contains categorical and numerical information such as IMDb score, director, gross, budget and so on and so forth. This system proposes a way to predict how successful a movie will be prior to its arrival at the box office instead of listening to critics and others on whether a movie will be successful or not. The proposed system provides a quite efficient approach to predict IMDb score on IMDb Movie Dataset. We will try to unveil the important factors influencing the score of IMDb Movie Data. In the exploratory analysis we found that number of voted users, number of critics for reviews, number of Facebook likes, duration of the movie and gross collection of movie affect the IMDb score strongly. Drama and Biopic movies are best in genres.***
Keywords: Machine Learning, Analysis, Prediction, Naïve Bayes algorithm, Data mining, Rating.

I. INTRODUCTION

Movies play a very important role in our everyday life. Film industry involves huge investment in terms of money, time, effort and people. So, it is vital to predict success and failure of movie. For this purpose, we tend to use Internet Movie Database containing categorical and numerical data. Different algorithms are applied in this dataset to calculate IMDb score. IMDb score is nothing but a measure that predicts success and failure of movie [1] accurately before actual release of the movie. In this work, we use Machine Learning and Data Mining Algorithms for prediction purpose. Data mining extracts patterns and trends from existing data and machines learning determines new algorithms from the previous experiments. Nowadays, hundreds of movies are get produced every year. The movie sector is a massive sector for investment but larger the business sectors, more is the complexity, and it is hard to choose how to invest wisely. Furthermore, significant investments come with more significant risks. As film industry is growing too fast, sufficient amount of data is available on the internet. This makes it an exciting as well as interesting field for data analysis. Predicting a movie's box office success is a very complex but essential task to perform. The definition of success of a movie is not fixed. It is relative. Some movies can be called as successful based on their gross income while some movies can snot be. There are many factors responsible in predicting success of movies. Instead of predicting only hit or flop movies, we can classify a film based on the profit it is making on box office into one of five categories such as super flop, flop, average, hit, super hit. Among them, there exist both good movies and flop ones. Therefore, the main question is how to know quality of movie before actually watching it or how can we choose a best movie to watch and relax on our weekends? Most of the time, we will check the movie score or have a look at its review to make right choice. IMDb website is really a good choice to refer at this time. Due to its popularity, IMDb website provides a great deal of information about movies and their views from movie viewers. The scores that IMDb gives are highly recognized by the audiences which represent the quality of content as well as audience's favor towards it. Therefore, in this project, we will try to unveil some of the important factors influencing the IMDb score and propose an efficient approach to predict success and failure of movie. The data we are using in our project comes from IMDb Movie Dataset on Gaggle.com [2]. It contains 28 variables for 5042 movies and 4906 posters, spanning across 100 years in 66 countries. There are 2399 unique names of directors and thousands of actors/actresses. In recent years, movie ratings are influenced by several factors such as actor, story, release date, likes by viewers, etc. These factors produces proper prediction about ratings for the new movies being released. There also have been numerous semantic analysis techniques to analyse user reviews which were applied to analyse the IMDb movie score. None of the

studies has successfully suggested a model which is good enough to be utilized in the industry. In this project, we use the IMDb dataset to predict whether the movie has made a profound impact on our society or not. Cinema is one among the foremost powerful media for mass communication within the world. Cinema has the capability to influence society each regionally and globally. Many different kinds of movies are produced each year. Some movies portray historical events, some illustrates a culture, whereas some give fantasy, and some do many more. We have a tendency to perform associate exploratory analysis of the data and observe some attention grabbing phenomenon, which also helps us to make our prediction strategy better. Also we will come to know about the features which affect the IMDb score. Our results finally show that we achieve a good prediction accuracy of IMDb score on this dataset.

II. RELATED WORK

Success of a movie primarily depends on the perspectives that how the movie has been justified. In early days, a number of people prioritized gross box office revenue initially. Few previous work portend gross of a movie depending on stochastic and regression models by using IMDb data. Some of them categorized either success or flop based on their revenues and apply binary classifications [3] for forecast. The measure of success of a movie does not completely depend on revenue. Success of movies rely on a numerous issue like actors/actresses, director, time of release, background story etc. Further, some people had created a model to predict success of movie with some pre-released data

A. Background Information

The audience votes on the movie review platforms have become an indicator of the movie success with the spread of Internet usage. IMDb is a very popular and commonly used movie review platforms among the other similar platforms. Therefore, it is very straightforward to predict IMDb movie ratings when it comes to predicting the liking of audiences. In the beginning, the traditional statistical methods were used to predict movie success before the release. However, there was not much success with these methods. Therefore, early studies started to use machine learning methods using metadata (information about the movie like its director, budget, runtime, etc.) and the audience reviews from various movie review platforms such as IMDb, Rottentomatoes, and MovieLens to predict audience's liking, hence the movie success. However, these data sources have become insufficient beside the social media [4], since the number of members and the activities of these members on such movie review platforms are very limited in contrast with the platforms like Twitter, Wikipedia and Facebook as can be seen in Table II. Before a movie is released, its marketing staff and advertisers start to generate information about it on these social media platforms. So, this makes it possible to use the movie data on these platforms, such as page view counts from Wikipedia long before it's release date. Especially word-of-mouth marketing is used heavily by the advertisers on Twitter. Twitter is a huge data source with approximately 650 million tweets per day. Such tweets data and the retweet count of these tweets give important insights about the movie before its release. The success of the predictive model to be used for movie rating prediction highly depends on the selecting and combining the right features. These features could be analyzed by dividing into three main groups such as metadata-based, social media-based and other. Metadata-based features are the features that have direct information about the movie like genre, budget, runtime, and director. Social media-based features are derived from the people's reactions and behaviors on social media like their comments, tweets, and page likes ([4],[5]).

B. Predictions Based on Movie Metadata

The movie metadata features like the genre, runtime, budget, and director have direct information about the movie, so these features are the first features used for the predictive model to make movie success predictions. In their studies, Hsu, Shen and Xie showed that a predictive model could be built by using the linear combinations of the genre, country, runtime, director and actor metadata features to predict IMDb movie rating. Besides the basic models such as linear regression, more complex models could also be generated by using the movie metadata. Sharda and Delen introduced a neural-network model built by using the metadata features to predict movie box office performance in their study [6]. In this study, they converted this regression problem into a classification problem by discretizing the predicted box office values into nine categories. Similar to this study, Ghiassi, Lio, and Moon used neural-networks to classify box office performance using metadata.

C. Predictions Based on Both Movie Metadata and Data from Social Media

According to Bhave, Kulkarni, Biramane, and Kosamkar using both classical and social factors together could lead to more accuracy in predicting movie success.Today, with the widespread of social media platforms, they produce enormous data every day. For example, the people's opinion could be inferred from the Twitter, since tweets directly indicate their thoughts. So, these data have become very valuable for many data mining fields including the predictive analysis of movie success. Asur and Huberman showed that the social media data could affect the box office prediction performance. In their study, it is assumed that the movies which people share positive tweets about will have more audiences.

They proved this assumption modeling with linear regression approach. Many of the studies utilized from one social media source [7], whereas Apala et al. introduced the idea of combining more than one social media sources together. They used both movie metadata like genre and director and sentiments of the YouTube comments of audiences about the official trailer of the movie. They converted this regression problem into a classification problem like Sharda and Delen by splitting the box office performance into three categories, Hit, Neutral and Flop. However, their results were fairly low because of the inadequate data consisting of 35 movies. Liu et al. collected the data from their country-specific social media platform. In their study, the comparison results of linear regression [8] and support vector regression were presented using this specific data. Besides Twitter, Wikipedia could also be considered as a social media data source because of its intense usage and rich content. Mestyan, Yasseri, and Kertesz built a linear regression model to predict movie box office performance using Wikipedia page activities about the movie before it is released. They showed that the prediction results were improved by taking Wikipedia data into account.

D. Predictions Based on Other Sources

There are also other studies that used the data sources different from the common ones like Twitter, YouTube, Facebook, and Wikipedia. Demir, Kapralova, and Lai used Google Trends data in order to predict IMDb movie ratings by converting this problem into a classification problem. In another study, various movie metadata were combined with extracted visual and auditory features from the movie trailers. It was shown that using these combined data was increased the performance of the box office prediction in contrast with using only movie metadata. A very different approach was introduced with the study of Eliashberg, Hui, and Zhang, using only the movie's script and its estimated budget. In their study they aimed to help the producers at the movie making decision step, using the semantics of the scripts based on the kernel-based approach

III. MOTIVATION

Basically we are interested to examine whether there are any trends among films that lead them to become successful at the movie boxoffice, and whether a film's boxoffice success correlates with its ratings ([9], [10]). A useful analysis would help us to predict how well a film does at the boxoffice before it's screening. Now, we don't need to rely on critics or our own assumptions. Basically we want to determine if there is any "Hollywood formula" to make movie successful. There is no any universal way that can claim the greatness of movies. So, the questionaries , How can someone determine the greatness of a movie before it's actual release. Many people rely on critics to measure the quality of a film, while others use their own instincts. But it takes some time to get a sufficient amount of critics' review after release of movie. And human instincts are not always reliable. Analysing the attributes of a movie using machine learning techniques is a relatively unexplored method for predicting its success [11]. Even considering that such information might be of interest not only to the movie sector in the form producers and financiers, but also to academics, service providers and viewers, most of the current work seems to be focused towards user-specific preferences or analysis of movie reviews.

IV. PROBLEM STATEMENT

Proposed system is to predict the success and failure of upcoming movie based on several attributes. The system will use a classification model to find correlation between several attributes like actor, movie rating etc. The system will predict success rating from the correlation between various movie criteria. This model can be used by movie watchers, producers.

V. DATASET

We use IMDb dataset to train and test our system. We take this dataset from kaggle.com. The dataset contains information about several movies such as movie title, actors, directors, genre, sequel, facebook popularity, etc. This data is available in variety of formats. We need to convert it into numerical format in order to apply machine learning algorithm effectively. IMDb is a very good choice to refer at this time. It is very popular as it contains great information about movies and comments from movie-viewers. IMDb scores are highly recognized and recommended by public.

VI. PRAPOSED SYSTEM

The first step is to identify a dataset of movie data which is suitable for analysis. Relevant attributes need to be selected from the movie data. Attributes can be general pre-production information regarding film productions such as movie title, sequel, genre, language and information about writers, actors, and directors. Similarly, the data must include some measure of success, such as user movie ratings. Secondly, the relevant dataset has to be prepared and structured in such a way that the data used is representative of the movie scene at large, as well as suitable for analysis by the relevant machine learning techniques and algorithms. Further, correlation is performed on relevant dataset to find the relationship between al the variable with each other. The important step in training our

system is to apply classification model. There are many classifiers. Lastly, the prediction performance of the relevant machine learning algorithm has to be evaluated on the dataset in order to determine success and failure of movie accurately.

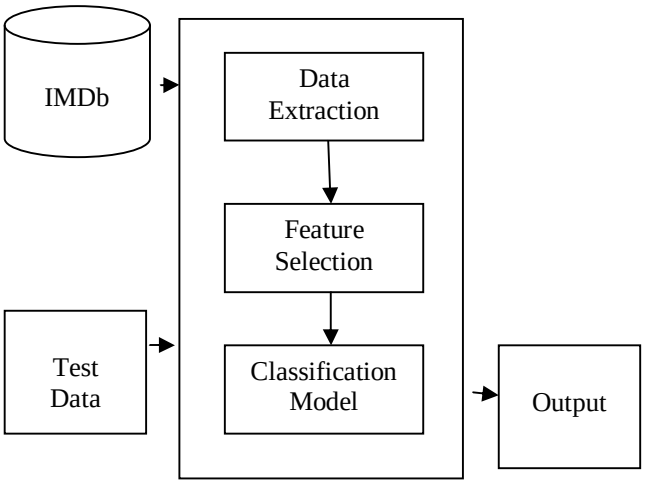


Fig. System Architecture

A. Process Model Tasks

1) *Data Extraction:* Data Extraction is the act or process of retrieving data out of data sources for further data processing or data storage. In our project we are extracting IMDB dataset.
2) *Data Pre-processing:* Data pre-processing is a technique that is used to convert the raw data into clean dataset.
3) *Feature Extraction:* Feature extraction is a general term for methods of constructing combinations of the variables.
4) *Feature Selection:* Feature selection is for filtering irrelevant or redundant features from IMDB dataset.

B. System Design

1) *User Interface Layer:* It basically comprises of interaction of the external user with the system, here the basic manifesto is composed by highlighting the registration module and login module .The login module is used by specific entities those who already have a authenticate account and the registration module is for the new user who wants to sign up to a visualized view of the generated report and select the specific parameters to make future predictions.
2) *Database Layer:* At this layer data required to carry out analysis and to generate a visualized report is stored and retrieved. The data would be like csv file comprises of records of movies, different movie attributes and ratings. The data would be uploaded at real time basis too by the user with the help of user interface by user side.
3) *Movie Performance Analysis System:* At this layer the data manipulation is made. Basically it comprises of correlation, classification and prediction modules.

C. Correlation

The correlation is a very common and most useful statistic technique. It can show whether and how pairs of variables are related with each other. Basically, it gives degree of relationship between two variables. Correlation works only for quantifiable data. Therefore, categorical data should be converted into numerical data in order to apply correlation on it. Correlation matrix is used to find relationship between all the variables in the dataset.

D. Classification

Data classification is the process of sorting and categorizing data into various types, forms or any other distinct parameters. Data classification enables the separation and classification of data according to data set requirements for various objectives. It is mainly a data management process. Which has been used to classify respective data into a particular form so that data pre-processing is made easy and also helps to neglect the features which does not have as much effect or response.

E. Prediction

Prediction refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome. Prediction can be carried out using naïve bayes, Decision Tree, K-Nearest Neighbours (KNN) and AdaBoost.

F. Objectives

1) To predict whether the movie will hit, super hit or flop.
2) To find out review of new movie.
3) To help users to decide whether to book ticket in advance or not.
4) To save money of producer.

VII. CLASSIFICATION ALGORITHMS

A. KNN Algorithm

1) Load the data
2) Initialize K to your chosen number of neighbors
3) For each example in the data
a) Calculate the distance between the query example and the current example from the data.
b) Add the distance and the index of the example to an ordered collection
4) Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
5) Pick the first K entries from the sorted collection
6) Get the labels of the selected K entries
7) If regression, return the mean of the K labels
8) If classification, return the mode of the K labels

B. Naïve Bayes Algorithm

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$P(A|B) = (P(B|A) * P(A)) / P(B)$

Where A and B are events and P(B) not equal to 0.

1) Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as evidence**.**
2) P(A) is the priori of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance (here, it is event B).
3) P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.

Now, with regards to our dataset, we can apply Bayes' theorem in following way:

$P(y|B) = (P(B|y) * P(y)) / P(B)$

Where, y is class variable and X is a dependent feature vector (of size *n*) where:

$X = (x1, x2, x3,\ldots\ldots,xn)$

Now, it's time to put a naive assumption to the Bayes' theorem, which is, independence among the features. So now, we split evidence into the independent parts.

Now, if any two events A and B are independent, then,

Hence, we reach to the result:

$P(y|x1,x2,\ldots,xn)=[P(x1|y)P(x1|y)\ldots.. P(xn|y)P(y)]/[P(x1) P(x2)\ldots. P(xn)]$

So, finally, we are left with the task of calculating P(y) and $P(x_i | y)$.

Please note that P(y) is also called class probability and $P(x_i | y)$ is called conditional probability**.**

The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $P(x_i | y)$.

C. Decision Tree Algorithm

1) Pick the best attribute/feature. The best attribute is one which best splits or separates the data.
2) Ask the relevant question.
3) Follow the answer path.
4) Go to step 1 until you arrive to the answer.

D. SVM Algorithm

1) Prepare and format dataset.
2) Normalize dataset.
3) Select Activating Function.

4) Optimize parameters using search algorithms after cross-validation [8].
5) Train SVM network.
6) Test SVM network.
7) Evaluate model performance

E. Random Forest

1) Select random samples from a given dataset.
2) Construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
3) Perform voting for every predicted result.
4) Select the most voted prediction result as the final prediction result.

VIII. ACKNOWLEDGMENT

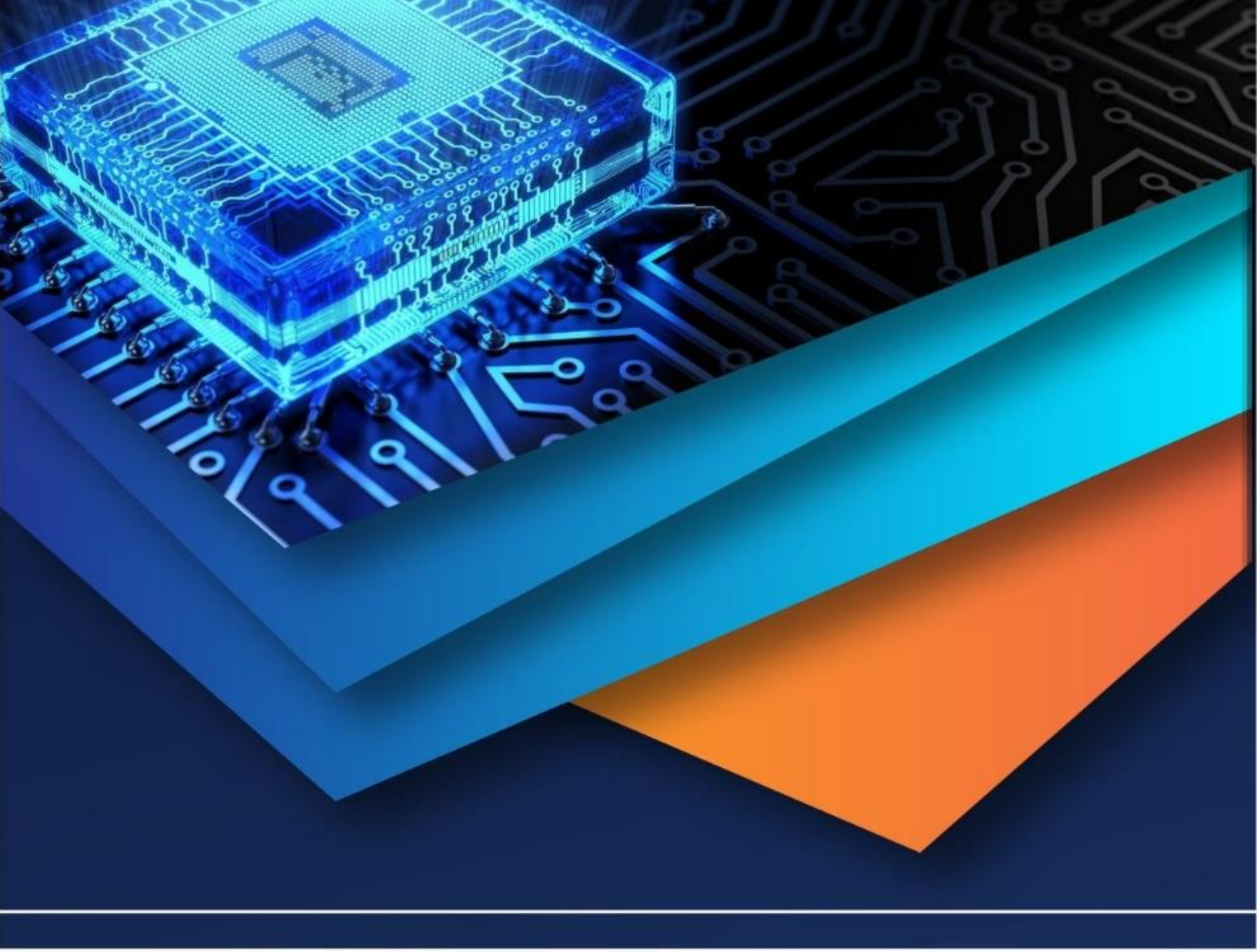
Working on this topic "Movie Success Prediction using Machine Learning Algorithms" is a great learning experience for us. We would like to thank our guide Mr. Y. S. Deshmukh whose interest and guidance helped us to work on this topic as well as he has provided facilities to explore the subject with more enthusiasm.
This experience will always encourage us to do our work perfectly and professionally. We also extend our gratitude to Dr. M. A. Jawale (H.O.D. IT Department). We express our immense pleasure and thankfulness to all the teachers and staff of the Department of In-formation Technology Engineering, Sanjivani College Of Engineering, Kopargaon for their co-operation and support.
Last but not the least, we thank all others, and especially our friends who in one way or another helped us with this research paper.

IX. CONCLUSION

In this system, we present the systematic approach towards some efficient strategies for movie success and failure prediction. The IMDB is an interesting dataset to analyze. A movie success does not only depend on features related to movies. Number of audiences plays a very important role for a movie to become successful. Because the whole industry is about audiences. The whole industry will make no sense, if there are no audience to watch movie. Number of ticket sold during a specific year can indicate the number of audiences of that year.

REFERENCES

[1] Garima Verma and Hemraj Verma's,,"Predicting Bollywood Movies Success Using Machine Learning Technique"IEEE,2019.
[2] Rijul Dhir and Anand Raj,"Movie Success Predections using Machine Learning and their Comparison "IEEE International Conference on Secure Cyber Computing and Communication, 2018.
[3] Ashutosh Kanitar's, "Bollywood Movie Success Prediction using Machine Learning Algorithms" IEEE Third International Conference on Circuits, Control, Communication and Computing, 2018.
[4] Beyza Çizmeci and Sule Gündüz Ögüdücü's," Predicting IMDb Rating of Pre-release Movies with Factorization Machines Using Social Media" IEEE 3rd international Conference on Computer Science and Engineering, 2018.
[5] Steve Shim and Mohammad Pourhomayoun's," Predictng Movie Market Revenues Using Social Media Data" IEEE International Conference on Information Reuse and Integration ,2017
[6] Nahid Quader and Md. Osman Gani," A Machine Learning Approach to Predict Movie Box- office and Information Technology(ICCIT), 22-24 , December, 2017.
[7] Beyza C¸ izmeci and Sule Gunduz Oguducu " Predicting IMDb Ratings of Pre-release Movies with factorization machines using social media" - International Conference of Computer and Information Technology (ICCIT), 2017.
[8] Subramaniyaswamy V., and Viginesh Vaibhav M., "Predicting Movie Box Office Succes using Multiple Regression and SVM" - International Conference of Computer and Information Technology (ICCIT), 2017.
[9] "Global box office revenue 2016 | Statistic." [Online]. Available https://www.statista.com/statistics/259987 /global- box- officerevenue/. [Accessed: 03-Jun-2018] Communication Technology(EICT).
[10] Forbes(2016) " Experts Predict a Drop in Box Office Revenue In 2016 After a Record Year for Hollywood" https://www.forbes.com/sites/simonthompson/2016/01/05/ experts predict-a-drop-in-box-office-- in- Revenue 016-after-a-record-year-forhollywood/#402059897195
[11] M. H Latif, H. Afzal ."Prediction of Movies Popularity Using Machine Learning Techniques" National University of Sciences and technology, H- 12, ISB, vol. 16, no. 8, pp. 127–131, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY