IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, December 2023

Predicting IMDb Rating of Movies by Machine Learning Techniques

Prof. Pathan Sir, Narache Abhiruchi, Kotalwar Shrutika
Department of Computer Engineering
Gramin College of Engineering Vishnupuri, Nanded, Maharashtra, India

Abstract: *Film Industry is not only a industry or a centre of entertainment, rather it is now a centre of global business. All over the world is now excited about a movie's box office success, popularity etc. A huge data is available online about these movies success or popularity. We have used Hollywood movie list from Wikipedia and their rating from IMDb movie rating website to create our data set. Then machine learning classification algorithms are applied of the data set. Lastly an efficient model is developed to predict a movie's IMDb rating. The model gives good classification measures with the data set. Index Terms-Movie Rating, Machine Learning, Prediction, Box Office, IMDb.*

Keywords: Movie Rating, Machine Learning, Prediction, Box Office, IMDb

I. INTRODUCTION

Now a days movies are not the only source of recreation, rather it is one of the major sources of global commerce and marketing. Movies create a new craze among people specially young people. Not only movie directors and box office officials are concerned with the success of movies but general people also. People used to talk about these in social medias. Therefore analysis of social media data about movies is recently popular among the data analysts. Other than this there remains some other scopes like analyzing a director's previous success histories or a actor's previous popularity etc. Again the analysis may be different on different countries. Naturally peoples from all the regions of the world do not react in the similar way. Movies are now available on internet. There are platforms like IMDb (Internet Movie Database) 1, Rotten Tomatoes 2, Metacritics 3 etc. where people can share their reviews about movies. Day by day these platforms are becoming popular since people are getting honest reviews there. So, huge data is available online about reviews and ratings of movies. In this paper this data about movie rating is analyzed to predict rating of movies. There are a good number of studies to predict the movie success rate as people are too much excited about films. Very few studies include the features (director, screen play, actor, actress, genre etc.) of a movie to predict the success rate. So, in this paper, the focus is to predict success rate based on movie's own features. For example, Director, Screenplay Actors/Actress, Country and Genre. Again, IMDb is a popular platform which is growing day by day. Generally people shows much interest if the IMDb rating of a movie if high. Therefore, the ultimate motive is to find out an model that can efficiently predict IMDb rating of a movie.

II. RELATED WORK

Several works are done to predict movie success rate before a movie is released. Many researchers have used various machine learning techniques to predict the success rate. A data about movie attributes rather than social media data is used to analyze and found that logistic regression gives 84% accuracy level. Pramod, Abhisht and Geetha shows that Fuzzy logic gives high accuracy for categorizing predictions. Machine learning approaches are applied on synthetic dataset to build efficient structure for prediction using IMDB is used. Various machine learning algorithms are being used to predict movie success rate. Depending on various movie attributes mathematical models are being implemented to determine movie success. Another study shows the market shares of domestic and international movies in Russia, where the international movies are ahead from the year of 2002 to 2014. This paper also distinguished three factors behind success of a movie. They are, budget, brands like popular actor/actress, directors and viewer's review. Lastly the model concludes that the sanction and budget has comparatively high effect on success of movies. An interesting work about analyzing moviegoers taste along with behavior is done recently. Here at first individuals taste is analyzed to

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, December 2023

make a model then aggregate to predict box office result. Indian film industry is vast and it's impact and influence is huge. So data mining techniques are used to predict whether a Bollywood movie will be a blockbuster one or not . A good number of studies are done based on social media data analysis. The data may be a review or comments or reactions for an event etc.

III. PROBLEM DOMAIN

This paper focuses on finding a model to predict a movie's IMDb rating using the movie's information like studio, director, screenplay, actor, actress, genre and country as input.features. In data set, title of the movie and year of the movie release is also stated. The model is built using machine learning techniques. It is trained and tested by the data set prepared from Wikipedia and IMDb movie rating website. Hence, the model built here is learned and tested by supervised learning technique. The data set created here contains the class attribute "rating" which can be flop, below average, average and hit, is measured by the rating from IMDb website. List of Hollywood movies released in the year of 2018 are fetched from Wikipedia. Then the ratings of each of those movies are collected one by one from IMDb website. Therefore the data set is a real data set and contains data from two different sources. Then a number of popular machine learning classification algorithms are applied on the data set. But as the data set prepared is imbalanced, so techniques are used to make it a balanced one before applying classification algorithms. To get satisfying accuracy measures, resampling without replacement is done with the data set. Lastly, the accuracy measures of different algorithms are compared. The measures include accuracy, kappa statistics, root mean squared error. Therefore, the model can predict movie's rating with high accuracy

IV. PROPOSED METHODOLOGY

The working method for this work involves few steps. The methodology is shown in figure 1. The steps are described below.

- Data Extraction.
- Data Preprocessing.
- Applying Machine Learning Techniques.
- Comparing the results of different algorithms.

An algorithm to develop the efficient model is illustrated in algorithm 1. This algorithm shows every details of our working procedure

Algorithm 1 Algorithm for developing the model

1: Prepare data set
2: Check Minority
3: If needed apply SMOTE algorithm until the minority class becomes equal to the size of it's closest class
4: Classification
5: Accuracy ←− 0
6: while True do
7: Resample Data
8: Call (Classifier)
9: if % of correctly classified Instance >Previous Accuracy Measure then
10: Accuracy ←− % of correctly classified Instance
11: else
12: Break
13: end if
14: end while=0

After completing the preprocessing step, classifiers are used in a repeated style. Each time data is resampled without replacement and classifiers are used. Each time accuracy is recorded. This repeating process is ended when accuracy doesn't increase anymore. Thus, the highest accuracy found from each classifier is recorded. In the algorithm, a while

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, December 2023

loop is used which repeats this process and the loop breaks if accuracy doesn't increase at a step. Therefore the highest accuracy is recorded lastly.

V. EXTERNAL INTERFACE REQUIREMENTS

User Interfaces: The interface of the software will provide options for a relatively easy data input processes textboxes that will be properly labeled. It will also have a user-friendly view of the whole system with simple and easy undertaking of action-driven processes as command buttons are functionally labeled. With all these, target users of this software will relatively find it not difficult to use it. User interface screen will be having various Interfaces like for input and Output. For Input, it will be having option an option for imputing the movie name, movie genre, cast name, pre-production details, storywriter, dop, director, producer names, etc. For Output, It will show the rating of that movie. Again, For Output Screen, it will be generating Graph to describe data graphically.

Software Interfaces: The software interface should be designed to facilitate user interaction, allowing users to perform tasks such as data preprocessing, feature engineering, model training, evaluation, prediction, and recommendation. The interface can be developed using frameworks like Flask, Django, or web-based UI libraries to create an interactive and accessible user experience. The software interface of an IMDb movie rating prediction system using machine learning (ML) typically involves the following components liked Data Collection, Data Preprocessing, Feature Engineering, Model Evaluation and Training etc.

Communication Interfaces: The communication interface of an IMDb movie rating prediction system using machine learning (ML) primarily involves interactions between the user and the system. Here are the key aspects of the communication interface:

1. Input of Movie Data: The user interface should allow users to input movie data for which they want to predict the rating. This can be done through text fields, dropdown menus, or file upload options, depending on the design of the interface. Users can enter movie details such as title, genre, cast members, director, release date, and any other relevant information
2. Feedback and Instructions: The interface should provide clear instructions and feedback to guide users through the prediction process. It should inform users about the required format or specific data fields to be entered. It can display error messages or suggestions for correction if the input is invalid or incomplete.
3. Rating Prediction Output: Once the user inputs the movie data, the system should process the information using the ML model and generate a rating prediction. The predicted rating should be displayed to the user in a prominent and understandable format, such as a numerical value or a visual representation (e.g., star rating).
4. Recommendations: In addition to the rating prediction, the interface can provide movie recommendations based on the predicted rating. It can display a list of recommended movies that have similar characteristics or align with the user's preferences. The recommendations can be presented as a separate section or alongside the rating prediction.
5. Visualizations and Insights: The interface can include visualizations and insights derived from the ML model. This can involve displaying charts, graphs, or other visual representations that showcase the importance of different features in the rating prediction. Users can gain insights into the factors that influence movie ratings and make more informed decisions.

VI. NON-FUNCTIONAL REQUIREMENTS

Performance Requirements:

1. Web server that can handle the load of many users. a. Keep performance with multiple simultaneous connections (logged users).
2. No delayed actions in system a. Quick load time between pressing login button and being logged in. b. General quick load times for any page submission and page loading.
3. Compatibility with web devices.

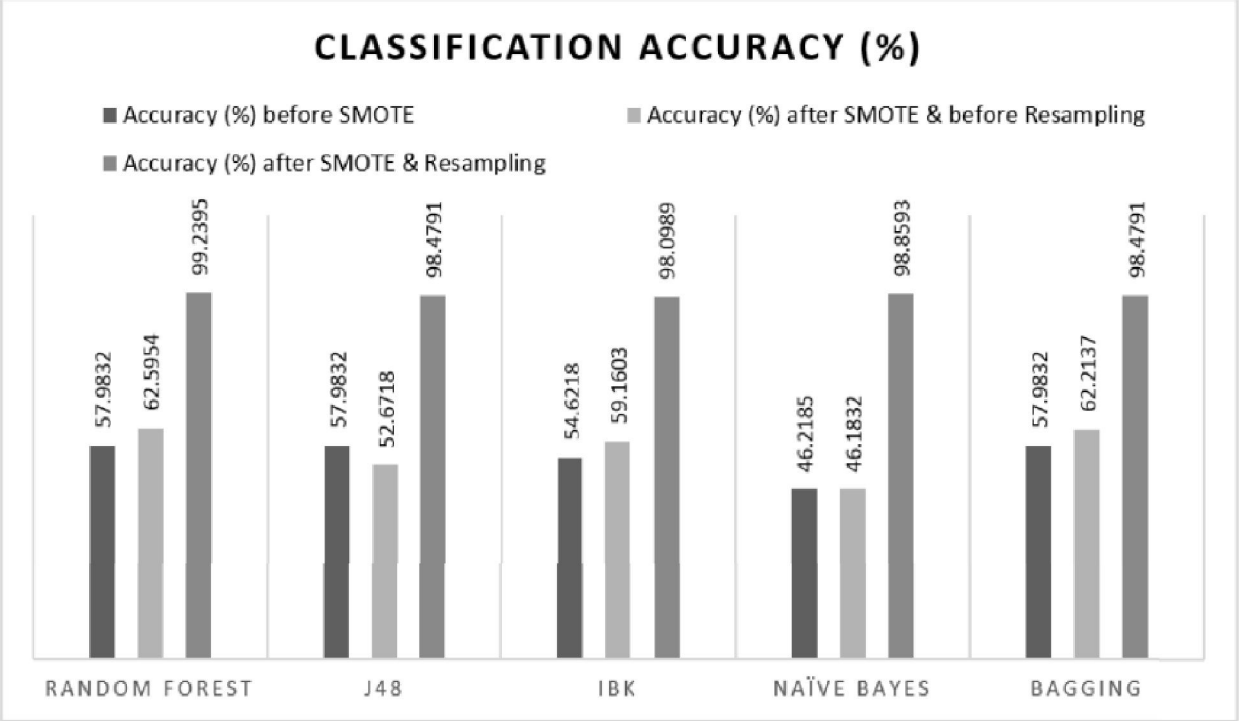
Safety Requirements: Different information is entered into the database such as information about movies which are very old. Such old movies have very low budgets or box office collections due to inflation changes over the years.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, December 2023

Movies which are currently released having less than 5000 votes can also have high IMDB rating for initial few days but the rating declines over the time and such movie data must be handled properly so that our model can be accurate enough.

Security Requirements: The website have respective accounts with password that enables only the organizer/s to login onto the system. Passwords are required so that no one else can access the system or database. Because the participants themselves provide the information entered into the database, there should be very little problems about the information entered. However, the organizer should always triple check every information given. Security systems need database storage just like many other applications.

VII. CLASSIFICATION ACCURACY OF THE CLASSIFIERS



VIII. CONCLUSION

As the business market of film industries are becoming huge day by day, competition here is also growing complex. Therefore, predicting movie's rating is growing complex also. Our model is developed on a real world data set and it is collected from two platforms, Wikipedia and IMDb. The model can also be used to predict some other ratings like Rotten Tomato or Metacritic. Other than films, TV shows, music shows, etc. can be predicted by our model using the features of our model describes some features having more influence on movie success and some other features having less or no influence. According to, budget is having a small positive influence but cast or actor/actress doesn't have any influence on Russian film industry. Thus, our work can be done by a weighted feature classification to reflect these influences. Along with hollywood movie dataset, bollywood or other movie dataset can be used to make the model more efficient. The database can be enriched by including the movies released on recent years. That is, along with 2018, movies from 2017 or 2016 can be included.

REFERENCES

[1] Internet Movie Data Base. URL: https://www.imdb.com
[2] M. H. Latif and H. Afzal, "Prediction of movies popularity using machine learning techniques," International Journal of Computer Science and Network Security (IJCSNS), vol. 16, no. 8, p. 127, 2016. 5.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, December 2023

[3] http://www.statista.com/statistics/259985/global-filmedentertainment-revenue/
[4] https://www.researchgate.net/publication/22253039
[5] E. Frank, I.H. Witten, Data Mining, Morgan Kaufmann Publishers, 2000.
[6] R. Parimi and D. Caragea, "Pre-release box-office success prediction for motion pictures," in International Workshop on Machine Learning and Data Mining in Pattern Recognition. Springer, 2013, pp. 571–585.
[7] 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, 2017, pp. 1-7, doi: 10.1109/ICCITECHN.2017.8281839.
[8]. A. Oghina, M. Breuss, M. Tsagkias, and M. De Rijke, "Predicting imdb movie ratings using social media," in European Conference on Information Retrieval. Springer, 2012, pp. 503–50