International Journal of Mechanical Engineering

A Review Paper on Bollywood Movie Prediction using Machine Learning

[1]Mukesh Kansari , [2]Dr. Vijayant Verma
[1]Ph.D (Research Scholar), [2]Professor (CSE)
[1,2]Department of Computer Science & Engineering, School of Engineering & IT, MATS University, Arang, Raipur (Chhattisgarh), India

Abstract: Movie industry is a huge sector for investment and become a multi-billion dollar industry but larger business sectors have more complexity and it is hard to choose how to invest. Big investments comes with bigger risks. The CEO of Motion Picture Association of America (MPAA) J. Valenti mentioned that 'No one can tell you how a movie is going to do in the marketplace. Not until the film opens in darkened theatre and sparks fly up between the screen and the audience' [1]. As movie industry is growing too fast day by day, there are now a huge amount of data available on the internet, which makes it an interesting field for data analysis. Predicting a movie success is a very complex task to do. The definition of a movie success is relative, some movies are called successful based on its worldwide gross income, and some movies may not shine in business part but can be called successful for good critics review and popularity.In this paper we considered a movie success based on its profit only. For this type of unpredictable nature of a movie success, it is very confusing decision for investors to make the right choice. Researches says almost 25% of movie revenue comes within the first or second week of its release [2]. So it is hard to predict a move success before its release.

Keywords: Bollywood, Machine Learning, KNN, Decision Tree, Support Vector Machine, Random Forest.

1.Introduction

Movies create a new craze among people especially young people. Not only movie directors and box office officials are concerned with the success of the movie but general people also. People used to talk about these in social medias. Therefore, analysis of social media data about the movies is popular among the data analysts. In this study, we apply machine learning tools to create a model which can predict whether a Bollywood movie will be successful or not, before it is released. The measurement of success of a movie does not solely depend on revenue. Success of movies rely on a numerous issue like actors/actresses, director, time of release, background story etc.In this study, an attempt has been made to develop a model that could predict the success of a Bollywood movie using different supervised machine learning algorithms. Five algorithms viz. Decision Tree (DT), Random Forest (RF), Logistics Regression (LR), K-Nearest Neighbour (KNN), Support vector machine (SVM)are used to develop a prediction model for analyzing movie success. A comparison of the accuracy of the prediction of different models has been done. The industry faces many flops when the movie does not perform well at the Box Office. Our model will try to predict the movie success rate by doing predictive analysis on the many features of the movie. The success of movie also depends upon some ratings like IMDB, Rotten Tomatoes, Meta critics etc. Users gives their ratings and comments on social media and websites about movie. Based on the rating and profit collected, we can predict movie is being hit or flop.

1.1 Motivation

As per data scientists desired to dig deeper into the business side of movies and explore the economics behind what makes a successful movie. Basically, we wanted toexamine whether there are any trends among films that lead them to become successful at the box office, and whether afilm's box office success correlates with its ratings. A useful analysis would help us predict how well a film does at the box office before it screens, without having to rely on criticsor our own instinct. Essentially, we want to determine if there is a "Bollywood formula" to making a success full movie. How can we tell the greatness of a movie before it isreleased in cinema? This question puzzled me for a longtime since there is no universal way to claim the goodnessof movies. Many people rely on critics to gauge the qualityof a film, while others use their instincts. But it takes thetime to obtain a reasonable amount of critics' review after amovie is released. And human instinct sometimes is unreliable. Analyzing the attributes of a movie using machine learning techniques is a relatively unexplored method for predicting its success. Investigating the characteristics of a movie for movie success prediction using machine learning method isa comparatively unexplored method. That information might be of interest not solitary to the movie sector in the form of producers and financiers, nonetheless also to service providers, spectators and statisticians. Nevertheless, current work seems to be focused only towards user-specific preferences or analysis of movie reviews. Being ableto forecast movie success in the form of box office revenue is also thoroughly related to this delinquent and ofgreat interest for the film industry. Such predictions would safeguard a decision support system during thepre-production phases of a movie production.

Similarly, means for movie prediction might also be of value for user commendations; either in the form of all-purpose commendations of not yet released films, or as a supplementary factor in a user-focused commendation system. Such systems might for example be found within services for media streaming, or other analogous services aimed more explicitly towards media unearthing and commendation.

2. Related Work

2.1 Literature Review

The 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 ([2], [3], [4]), initially. Few previous work ([4], [5],[6]), 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 for forecast.

The measurement of success of a movie does not solely depend on revenue. Success of movies rely on a numerous issue like actors/actresses, director, time of release, background story etc. Further few people had made a prediction model with some pre-released data which were used as their features [7]. In most of the case, people considered a very few features. As a result, their models work poorly. However, they ignored participation of audiences on whom success of a movie mostly depends. Although few people adopt many applications of NLP for sentiment analysis ([8], [9]) and gathered movie reviews for their test domain. But the accuracy of prediction lies on how big the test domain is. A small domain is not a good idea for measurement. Again most of them did not take critics reviews in account. Besides, users' reviews can be biased as a fan of actor/actress may fail to give unbiased opinion.

Quader et al. [12] Predicting society's reaction to a new product in the sense of popularity and adaption rate has become an emerging field of data analysis. The motion picture industry is a multi-billion-dollar business, and there is a massive amount of data related to movies is available over the internet. This study proposes a decision support system for movie investment sector using machine learning techniques. This research helps investors associated with this business for avoiding investment risks. The system predicts an approximate success rate of a movie based on its profitability by analyzing historical data from different sources like IMDb, Rotten Tomatoes, Box Office Mojo and Metacritic. Using Support Vector Machine (SVM), Neural Network and Natural Language Processing the system predicts a movie box office profit based on some pre-released features and post-released features. This paper shows Neural Network gives an accuracy of 84.1% for pre-released features and 89.27% for all features while SVM has 83.44% and 88.87% accuracy for pre-released features and all features respectively when one away prediction is considered. Also, they figure out that budget, IMDb votes and no. of screens are the most important features which play a vital role while predicting a movie's box-office success.

Meenakshi et al. [20]In real world prediction models and mechanisms can be used to predict the success of a movie. The proposed work aims to develop a system based upon data mining techniques that may help in predicting the success of a movie in advance thereby reducing certain level of uncertainty. An attempt is made to predict the past as well as the future of movie for the purpose of business certainty or simply a theoretical condition in which decision making [the success of the movie] is without risk, because the decision maker [movie makers and stakeholders] has all the information about the exact outcome of the decision, before he or she makes the decision [release of the movie]. With over two million spectators a day and films exported to over 100 countries, the impact of Bollywood film industry is formidable , together a series of interesting facts and relationships using a variety of data mining techniques.

In particular, concentrate on attributes relevant to the success prediction of movies, such as whether any particular actors or actresses are likely to help a movie to succeed. The paper additionally reports on the techniques used, giving their implementation and utility.

Additionally, found some attention-grabbing facts, such as the budget of a movie isn't any indication of how well-rated it'll be, there's a downward trend within the quality of films overtime, and also the director and actors/actresses involved in the movie.

2.2 Machine Learning Approach

Machine learning algorithms are generally classified as supervised learning, unsupervised learning, and semi-supervised learning. In this paper,decision tree, random forest, logistics regression, SVM, KNN have been applied to develop a Bollywood movie prediction model.

2.2.1 Decision Tree

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving **regression and classification problems** too.The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by **learning simple decision rules** inferred from prior data(training data).In Decision Trees, for predicting a class label for a record we start from the **root** of the tree. We compare the values of the root attribute with the record’s attribute [9]. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

2.2.2 Random Forest

Random forest is a supervised ensemble learning algorithm that is used for both classifications as well as regression problems. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees mean more robust forest. Similarly, the random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method that is better than a single decision tree because it reduces the over-fitting by averaging the result.

2.2.3 Logistic Regression

Logistic Regression comes under Supervised Learning. **Supervised Learning** is when the algorithm learns on a labeled dataset and analyses the training data. These labeled data sets have inputs and expected outputs.

Supervised learning can be further split into classification and regression.

Classification is about predicting a label, by identifying which category an object belongs to based on different parameters.

Regression is about predicting a continuous output, by finding the correlations between dependent and independent variables.

Logistic Regression is a statistical approach and a Machine Learning algorithm that is used for classification problems and is based on the concept of probability. It is used when the dependent variable (target) is categorical.

2.2.4 Support Vector Machine

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes. SVM is used for text classification tasks such as category assignment, detecting spam and sentiment analysis. It is also commonly used for image recognition challenges, performing particularly well in aspect-based recognition and color-based classification. SVM also plays a vital role in many areas of handwritten digit recognition, such as postal automation services. Support Vector Machine (SVM) is part of a group of kernel based methods which are used for pattern classification and regression. A classifier takes an input pattern called feature vector, and determines to which class it belongs to [22].

2.2.5 K-Nearest Neighbour

KNN also known as K-nearest neighbour is a supervised and pattern classification learning algorithm which helps us find which class the new input(test value) belongs to when k nearest neighbours are chosen and distance is calculated between them. It first identifies the k points in the training data that are closest to the test value and calculates the distance between all those categories. The test value will belong to the category whose distance is the least. In pattern recognition, the k-nearest neighbour's algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression [22].

3. Objective

The objective of this work is to predict the movie success rate using machine learning approach (Adaboost classifier) and perform the comparative analysis between the existing method such as SVM, KNN classifier,Random Forest, Decision Tree and Logistic regression etc.

4. Methodology

Data Preprocessing: The machine learning technique plays major role in data Preprocessing because it is the major part where the data would transform or get encoded in a way that all duplicate values would be deleted and also helps in shaping of the dataset.Preprocess the data set and evaluate which attributes are the most useful, by evaluating the correlation between the attributes and the success rate of the machine learning. Using this method, the condition it can achieve a possible success rate when trying to predict the rating and box office revenue.

Apply Algorithm: We used algorithms like Decision Tree, K-Nearest Neighbor, Support Vector Machine, Logistic Regression and Random Forest. We have chosen only these particular algorithms because they are specifically used in classification process and as we have to classify the data based on which can be used for movie success prediction and which could not be used these classification algorithms helped in achieving the desired results.

Split Data: Our dataset was split in 80:20 ratio and was then considered for predicting the test results. The predicted results of data are tested for accuracy using several algorithms mentioned above and the algorithm with most accuracy was finalized for predicting whether the data could be used for movie success prediction

Training and Testing Model: Train the different machine learning model such as Logistic Regression (LR), Support Vector Machine (LSVM), K-Nearest Neighbor (KNN), Decision Tree, Random Forest and predict the success rate.

Result: Analyze the results graphically to compare the results generated by different models.

5. Conclusion

In this paper, I studied to predicting the success of movie using historical data using different algorithms. The success of movie depends on various factors like genre, music, actor, actress, director,critics review, lyrics and most importantly user reviews and comments. Sometimes movies does good business in market but due to poor story movie get flop. The movie success prediction can be calculated using accuracy of different algorithms. We proposed to develop a model for predicting the success of movie being a Flop or Hit, before the movie actually released using machine learning techniques and algorithms.

6. References

[1]B.R.Litman&H.Ahn(1998). Predicting financial success of motion pictures. In B.R.Litman(Ed.),the motion picture mega –industry. Boston , MA: Allyn&bacon publishing ,inc.

[2]S. Gopinath , P.K. Chintaguntha ,and s.venkataraman , "blogs ,advertising, and local market movie box office performance ," management science ,vol.59,no.12 pp. 2635-2654,2013.

[3]mestyan, T.yasseri ,and J.kertesz , "early prediction of movie box office success based on wikipidia activity big data ,"PLoS ONE, vol.8,2013.

[4]J.S.Simonoff and I.R. sparrow , "predicting movie grosses : winners and losers,blockbusters and sleepers," chance vol.13no.3pp.15-24,2020.

[5]A.Chen, "forecasting gross revenues at the movie box office ," working paper ,university of Washington , seattle , WA,june2002.

[6]M.S.Sawhney and J.eliashbarg , "a parsimonious model for forecasting gross box office revenues of motion pictures," marking science ,vol.15,no.2pp. 113-131,1996

[7]R.Sharda and E.meany , "forecasting gate receipts using neural network and rough sets," in proceedings of the international DSI conference ,pp.1-5,2000.

[8]B.pang and L.lee, "thumbs up? Sentiment classification using machine learning techniques,"in proceedings of the conference on empirical methods in natural language processings (EMNLP),Philadelphia ,pp.79-86,july2002.

[9]p.chaovalit and L.zhou, "movie review mining : a comparison between supervised and unsupervised classification approaches," in proceeding of the Hawaii international conference on system science (HICSS),2005.

[10]J.valenti (1978). Motion pictures and their impact on society in the year 2000,speech given at the Midwest research institute , Kansas city , april 25,p.7.

[11]Rijudhir , anand raj "movie success predicition using machine learning algorithms and their comaparison" ,2018 first international conference on secure cyber computing and communication (ICSCCC),978-1-5386-6373-8/18,IEEE.

[12]NahidQuader et al. "A Machine learning Approach to predict movie box-office success",2017 20th international conference of computer and information technology (ICCIT),22-24 december,2017.

[13]Cary D.Butler et al. "predicting movie success using machine learning algorithms",los angeles community choice energy 2017.

[14]kyuhan Leel , jinsoo park ,Iljoo kim and youngseok choi, "predicting movie success with machine learning techniques :ways to improve accuracy" , InfSyst front -2016,DOI 10.1007/s1076-016-9689-z.

[15]Hemraj verma , garima verma "prediction model for bollywood movie success: A comoarative analaysis of performance of supervised machine learning algorithms", the review of socionetwork strategies, 2019.https:doi.org/10.1007/s12626-019-00040-6.

[16]paragAhivale ,omkarAcharya "success prediction of films at box- office using machine learning", international journal for research in applied science &engeneering technology (IJRASET), volume 3 issue IV, April 2015,ISSN:2321-9653.

[17]yoosin kim , mingon kang , seungRyuljeong “Text mining and sentiment Analysis for predicting Box office success”, KSII Transactions on internet And information systems vol.12, NO.8,Aug .2018.

[18]Nithin VR, pranav my,sarathBabuPBz,Lijiya A “predicting movie success Based on IMDb Data”,International Journal of Data Mining Techniques and Applications – 2014,ISSN:2278-2419(pages:365-368).

[19]Sameer Ranjanjaiswal , Divyansh Sharma “predicting success of Bollywood movies using machine learning Techniques” ACM COMPUTE 2017, November 16-17,2017, Bhopal,India.

[20]K Meenakshi, G Maragatham, NehaAgarwal and ishithaghosh “A Data mining Technique for Analyzing and predicting the success of Movie”, IOP Conf.series : Journal of physics :Conf.series 1000(2018) 012100 doi:10.1088/1742-6596/1000/1/012100.

[21]Nikhil Chaudhari , Karthik vardhrajan, shashankshekhar , prabhjitThind and swarnalatha P “A Data mining Approach to Language success Prediction of a feature film”, International Journal of Engineering science & Management Research-2016,ISSN 2349-6193

[22]KomalGothwal, DhiralSankhe ,NiravWaghela , Mitul Sharma, RamanandYadav “Movie success Prediction”, IOSR Journal of Engineering (IOSRJEN) ISSN (e):2250-3021,ISSN(p): 2278-8719.