Webology (ISSN: 1735-188X)
Volume 18, Number 5, 2021

A Movie Success Prediction Based On Public Reviews Using Supervised 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)

Abstract

Movies is the art form that provide entertainment, express emotions, giving social message and teach people something. There are different types of Films Genres available like sci-fi films, family drama films, romantic films, action films, horror films, musical films, adventures films etc. Movies have the power to learn about new culture, experience a different perspective, or open our eyes to a world we know nothing about. By watching movies, people can boost their moods, feeling relax by listening musics and can reduce stress. Some movies are good or bad depending upon the story, music, star-cast, public review etc. There are some common factors which describe movie hit or flop that include, a well written script, music, songs, actors and actresses, photography, and good story. There are some other factors also which include film revenue collection, best story, public reviews etc. By using machine learning approach we can predict the movies success. The movie success totally depends on public reviews and comments. If story is too good then automatically audience will watch the movie and they will do the mouth publicity about the film.

Keywords:- Film, Actor, Actress, Supervised Learning, Reviews, Story

1. Introduction

Every week movie is released in India and huge amount is invested for making movie. To predict the success of movie, there are few characteristic which we need to predict whether the movie is hit or flop. The characteristic are actor, actress, director, music director, and public review those who watch the movie and give their feedback on movie. Public review, critics score, Box office revenue collection are the main feature to predict the movie success. Movie prediction is useful for producer to understand whether his investment made by him is worth it or not. Based on previous data, and experience he can make decision to make his film better. Data mining and machine learning techniques will help to analyse the past data to predict the success of movie. We can categorize the success class such as hit, flop, super hit of the movies. This system helps to find out whether the movie is super hit, hit, flop on the basis of historical data of actor, actress, music director, writer, director, marketing budget and release date

of the new movie. If the movie releases on weekend, new movie will get higher weight age or if the movie releases on week days new movie will get low weight age. The factors such as actor, actress, director, writer, music director and marketing budget historical data of each component are calculated and movie success is predicted. Supervised learning is a subcategory of machine learning and artificial intelligence. It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately. There are other platforms like IMDb (Internet Movie Database) , TMDB, All Web, MovieChat, Rotten Tomatoes , Meta critics etc. where people can share their reviews and ratings about movies and we attempt to use those platform to get the dataset. In this project we use supervised learning to predict the movie success based on dataset.

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 ([5], [6], [7]), initially. Few previous work ([7], [8]), 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.

Hemant Kumar and Santosh Kumar [4] proposed a narrative approach to constructing and using a semantic relation between actor, director, genre, budget, ratings, producer and other certain essential attributes, thus alleviating the hypothesis into forecasting. Distinctively we will comply the formulation and composition of linear regression, Support Vector Machine and decision tree as an augmented amalgamated algorithm to assemble features for classification and predictions. Subsequently, the proposed investigational onsequences which will illustrate the precise and efficient results and prediction thereof. The proposed framework predicts the achievement of a motion picture in light of its gainfulness by utilizing chronicled information from different sources. Utilizing informal community examination and content mining methods, the framework naturally separates a few gatherings of highlights, including "who" are on the best composition (actor and director) what a film is about, "when" a motion picture will be released, and in addition "semi variety" highlights that match "who" with "what", and "when" with "what". Examination comes about with motion pictures amid years' time frame demonstrated that the framework beats benchmark techniques by a substantial edge in anticipating motion picture productivity. Novel highlights we proposed likewise made extraordinary commitments to the expectation. Moreover, to planning a choice emotionally supportive network with reasonable utilities, our investigation of key factors for motion picture productivity may likewise have suggestions for hypothetical research on group execution and the achievement of imaginative work.

Quader et al. [9] 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. [14] 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.

Dipak Gaikar et al. [10] developed a mathematical model for predicting the rating and success classes such as hit, flop and neutral of the movies. In order to do this, we have used a machine learning and data mining algorithm. The algorithm used for classification is k-NN. Popularity factor of various movie parameters like actor, actress, director, writer, budget etc. is collected which helps in the movie success prediction. This project helps the director or producer of the movie to pedecides the parameters such as actor, actress etc. of the movie. This project also helps the user to decide

whether to book ticket in advance or not based on upcoming movie prediction. George H. Chen and Devavrat [11] Shah focus is on nonasymptotic statistical guarantees, which we state in the form of how many training data and what algorithm parameters ensure that a nearest neighbor prediction method achieves a user-specified error tolerance. We begin with the most general of such results for nearest neighbor and related kernel regression and classification in general metric spaces. In such settings in which we assume very little structure, what enables successful prediction is smoothness in the function being estimated for regression, and a low probability of landing near the decision boundary for classification. In practice, these conditions could be difficult to verify empirically for a real dataset. We then cover recent theoretical guarantees on nearest neighbor prediction in the three case studies of time series forecasting, recommending products to people over time, and delineating human organs in medical images by looking at image patches. In these case studies, clustering structure, which is easier to verify in data and more readily interpretable by practitioners, enables successful prediction.

Ladislav Peska and Peter Vojtas [12] described details of our approach to the RecSys Challenge 2014: User Engagement as Evaluation. The challenge was based on a dataset, which contains tweets that are generated when users rate movies on IMDb (using the iOS app in a smartphone). The challenge for participants is to rank such tweets by expected user interaction, which is expressed in terms of retweet and favorite counts. During experiments we have tested several current off-the-shelf prediction techniques and proposed a variant of item biased k-NN algorithm, which better reflects user engagement and nature of the movie domain content-based attributes. Our final solution (placed in the third quartile of the challenge leader board) is an aggregation of several runs of this algorithm and some off-the-shelf predictors.

Sangjae Lee et al. [15] proposed decision trees, k-nearest-neighbors (k-NN), and linear regression using ensemble methods and the prediction performance of decision trees based on random forests, bagging and boosting are compared with that of kNN and linear regression based on bagging and boosting using the sample of 1439 movies. The results indicate that ensemble methods based on decision trees (random forests, bagging, boosting) outperform ensemble methods based on kNN (bagging, boosting) in predicting box office at week 1, 2, 3 after release. Decision trees using ensemble methods provide better prediction performance than ensemble methods based on linear regression analysis in the box office at week 1 after release. This is explained by the results that after comparing the prediction performance between ensemble methods and non-ensemble methods. For decision tree methods, unlike the other methods, the prediction performance of ensemble methods is greater than that of non-ensemble methods. This shows that decision trees using ensemble methods provide better application effectiveness of ensemble methods than k-NN and linear regression analysis.

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.

2.3 Classification of Machine Learning

There are three important types of Machine Learning Techniques such as supervised learning, unsupervised learning and reinforcement learning. We will discuss about supervised learning in details.

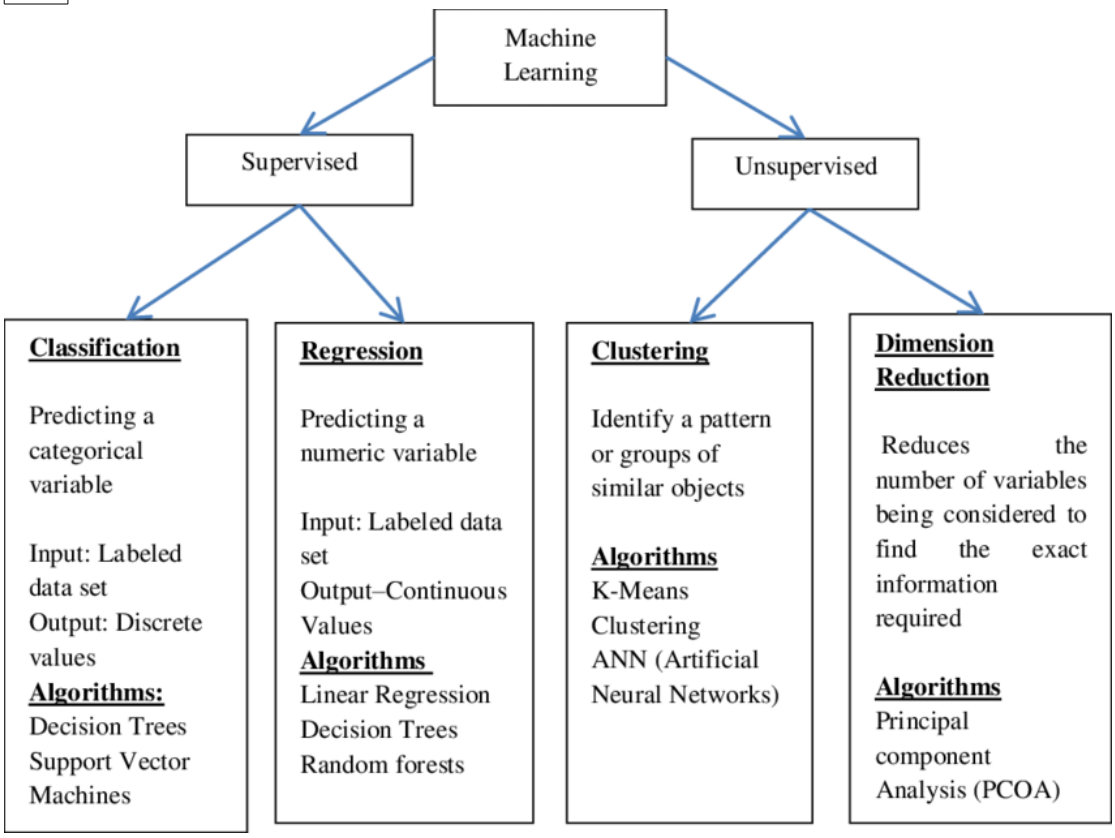

Fig 1: Classification of Machine Learning

3. Supervised Learning

Supervised Learning is the most popular paradigm for performing machine learning operations. It is widely used for data where there is a precise mapping between input-output data. The dataset, in this case, is labeled, meaning that the algorithm identifies the features explicitly and carries out predictions or classification accordingly [1]. As the training period progresses, the algorithm is able to identify the relationships between the two variables such that we can predict a new outcome.

There are two main types of supervised learning problems: they are classification that involves predicting a class label and regression that involves predicting a numerical value.[2]

- Classification: Supervised learning problem that involves predicting a class label.
- Regression: Supervised learning problem that involves predicting a numerical label.

Both classification and regression problems may have one or more input variables and input variables may be any data type, such as numerical or categorical.

4. Unsupervised Learning

Unsupervised machine learning holds the advantage of being able to work with unlabeled data. This means that human labor is not required to make the dataset machine-readable, allowing much larger datasets to be worked on by the program. The model learns through observation and finds structures in the data. Once the model is given a dataset, it automatically finds patterns and relationships in the dataset by creating clusters in it.[13] In supervised learning, the labels allow the algorithm to find the exact nature of the relationship between any two data points. However, unsupervised learning does not have labels to work off of, resulting in the creation of hidden structures. Relationships between data points are perceived by the algorithm in an abstract manner, with no input required from human beings. The creation of these hidden structures is what makes unsupervised learning algorithms versatile. Instead of a defined and set problem statement, unsupervised learning algorithms can adapt to the data by dynamically changing hidden structures.[13] This offers more post-deployment development than supervised learning algorithms. What it cannot do is add labels to the cluster, like it cannot say this a group of apples or mangoes, but it will separate all the apples from mangoes. Suppose we presented images of apples, bananas and mangoes to the model, so what it does, based on some patterns and relationships it creates clusters and divides the dataset into those clusters. Now if a new data is fed to the model, it adds it to one of the created clusters. The example of unsupervised learning is k-mean clustering, principle component analysis, SVD, FP-growth etc. There are many types of unsupervised learning, although there are two main problems that are often encountered by a practitioner: they are clustering that involves finding groups in the data and density estimation that involves summarizing the distribution of data.[13]

- **Clustering:** Unsupervised learning problem that involves finding groups in data.
- **Density Estimation:** Unsupervised learning problem that involves summarizing the distribution of data.

5. Supervised vs Unsupervised Learning

The difference between supervised and unsupervised learning are given below [16].

Parameters	**Supervised Learning**	**Unsupervised Learning**
Input Data	Algorithms are trained using labeled data	Algorithms are used against data that is not labeled
Accuracy	Highly Accurate	Less Accurate
No. of classes	No. of classes is known	No. of classes is not known
Data Analysis	Uses offline analysis	Uses real-time analysis of data

Algorithm used	Random Forest, SVM, Neural Network etc	K-means clustering, Apriori algorithm etc.

6. Conclusion

The intent of this study is to predict the success of movies using supervised learning learning approach. The prediction of movie success can be done using various factors like genre, critics, actor, actress, director, music and box office collection etc. In this paper, we present the classification of machine learning techniques such as supervised and unsupervised learning. To prevent the industry from big loss, we can predict the movie success of previous historical data and by applying the some analysis work and public review we can worked on future improvements of the movie.

7. References

[1] Christopher M. Bishop, "Pattern Recognition and Machine Learning (Information Science and Statistics)", 2006. Page-3

[2] Russel, "Artificial Intelligence: A Modern Approach", January 1, 2015

[3] https://www.edureka.co/blog/what-is-machine-learning/

[4] Hemant Kumar and Santosh Kumar, "Predicting Movie Success or Failure using Linear Regression & SVM over Map-Reduce in Hadoop", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 6, Issue 6, June 2018, pp 6426-6433.

[5] 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.

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

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

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

[8] 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

[9] Nahid Quader 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.

[10] Dipak Gaikar et al., "Movie Success Prediction Using Popularity Factor from Social Media", International Research Journal of Engineering and Technology (IRJET), Volume: 06, Issue: 04, Apr 2019.

[11] George H. Chen, Devavrat Shah, "Explaining the Success of Nearest Neighbor Methods in Prediction", Vol. XX, No. XX, pp 1–250. DOI: XXX.

[12] Ladislav Peska and Peter Vojtas "Biased k-NN Similarity Content Based Prediction of Movie Tweets Popularity", Dateso 2015, pp. 101–110, CEUR-WS.org/Vol-1343.

[13] Hastie et al. “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”, Second Edition (Springer Series in Statistics) , 2016, pp. 28. [online]. Available: https://www.amazon.com/Elements-Statistical-Learning-Prediction-Statistics

[14] 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.

[15] Sangjae Lee et al., “Comparing performance of ensemble methods in predicting movie box office revenue”, Heliyon 6 (2020), pp 1-8

[16] https://www.geeksforgeeks.org/supervised-unsupervised-learning/