Movie Success Prediction

Kshitij Gupta, Shubham bajpayee, A.Meena Priyadharsini

Abstract*: The film business is a billion-dollar business, and extensive measure of data identified with motion pictures is accessible over the web. In this system we are analyzing the dataset for predicting the success of the movies. For doing this the analysis of the dataset is done in which the chronicled information of every segment, for example, actor, actress, director, music that impacts the achievement or disappointment of a motion picture is given weight age and after that dependent on different parameters we are predicting whether the movie will be a flop, average or superhit. Certain algorithms are used that can help to predict whether the movies will be a flop, average, or superhit. In this model we focus on the attribute selection for predicting success of the movies. A comparative analysis is to be performed so as to find the accurate results among the algorithms used. Few parameters that are important for predicting success of a movie are gross, genres, release date, star powers of actors, actress, directors, and budget etc. In the dataset there are 28 parameters. The task is to find out most relevant parameters. This will be achieved by Feature selection method as shown in figure 1. Feature selection method is present in "sklearn" library of python. Feature selection method includes Decision trees, information gain, gain ratio. Generating heatmap to visualize success of movie in different regions. Various graphs are generated between time vs algorithms and accuracy vs algorithms for analysis.

***Index Terms*: Decision Tree, Regressions, Lasso regressor, support vector regression (SVR), Sentimental Analysis, Metacritic, IMDB.**

I. INTRODUCTION

Film industry largely affects our general public. Distinctive assortments of motion picture are discharged for the current year. Hollywood is seen as the most prepared film industry of the world, and the best similarly as film industry net gain however Indian film is seen as the best film industry in regards to the amount of motion pictures made and the amount of tickets sold.

In spite of the fact that motion pictures are giving different diverse classes however what makes a film effective? These days, as film industry is becoming excessively quick, there are substantial measure of data accessible on the web, which can be commonly utilized for data examination. Motion picture achievement forecast is an exceptionally confused undertaking to do. The meaning of achievement of a film is relative, a few motion pictures are called effective dependent on its overall gross salary, and a few motion pictures may not. Different motion picture database utilizes clicks, audits, web journals, star throws, remarks to anticipate yet connected four data mining procedures to the dataset. The data mining methods that are utilized are Decision Tree regressor, Lasso Regressor, Random forest, Support vector Regression (SVR). The way toward discovering designs in immense data sets which includes different strategies for machine learning, measurements, and database frameworks is called data mining. Data mining is an interdisciplinary subpart of software engineering with a by and large goal to separate data from a data set and change the data into an intelligible structure for further use. Data mining is the investigation venture of the" information revelation in databases" process or KDD.

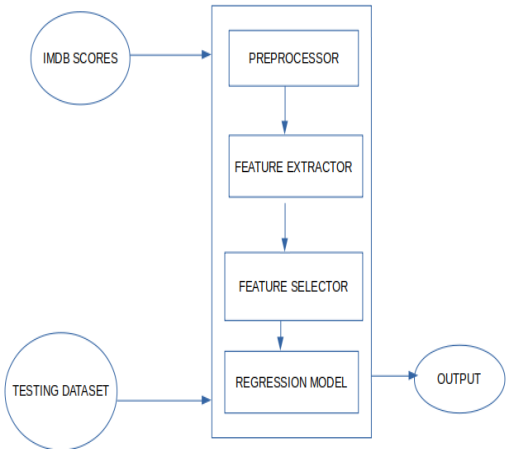

Figure1. Overview of model

We will be using different data mining techniques that are SVR, Random forest, Decision tree regressor and Lasso regressor so as to get the more accurate results as shown in figure 2. The system will include both pre-released and post released movie prediction considering various respective attributes.

II. LITERATURE SURVEY

In the past many researchers have tried to identify features that can be used to predict the success of a movie and have computed correlations between those variables.

K Meenakshi published paper which makes use of k-means clustering and decision tree algorithms and makes use of dataset consisting of 30000 records and the results are calculated from success rate ranging from flop to hit [1]. Decision trees are complex especially in preparing decision trees with large branches, are complex and time consuming.Determining the splitting criterion for each node in a decision tree is complicated task that require more expertise and experience. Decision trees examine only a single field at a time, this may not correspond well with the actual distribution of records in the decision space. Quader, Nahid and Gani, Md and Chaki, Dipankar and Ali, Md proposed a choice emotionally supportive network for motion picture speculation division utilizing machine learning methods. This examination helps financial specialists related with this business for keeping away

Revised Manuscript Received on September 23, 2019

Kshitij Gupta, Department of CSE, SRMIST, Chennai, India.
Shubham bajpayee, Department of CSE, SRMIST, Chennai, India.
A.Meena Priyadharsini, Department of CSE, SRMIST, Chennai, India.

from venture dangers. The framework predicts a rough achievement rate of a film dependent on its benefit by breaking down historical information from various sources like IMDb, Rotten Tomatoes etc. by using Support Vector Machine (SVM), Neural Network and Natural Language Processing [2]. This study only considered pre-released attributes. Prediction of movie is relative and pre and post released attributes both should be considered. Further SVM uses kernel parameters and concept of choice of kernel and Kernel methods are sensitive to over fitting.

Vr, Nithin & Babu Pb, Sarath adopted many applications of machine learning such as linear regression and logistic regression [3]. The accuracy of their model using linear regression was 51%. While using logistic regression, they obtained about 42.2% accuracy. Sequel of movies cannot not be predicted from this model. Predicting only on the basis of one attribute that is gross revenue but success of movie should be relative and therefore cannot be predicted by using only one attribute.

Suhaas Prasad tried to propose a system capable of predicting movie ratings based on user rating histories. A typical filtering strategy is the k-nearest neighbor (kNN) technique in which the user-item inclination is dictated by looking at the evaluations of comparative users or items [4]. The information for the user's appraisals and interpersonal organizations have been given by Flixster, a social motion picture stage where users can rate and audit films with their companions. For the baseline item-based kNN the RMSE comes out to be approximately 0.989. Flixster users may not represent one's true social graphs, better to consider Facebook users.

Darin Im & Minh Thao Nguyen study focuses on relying highly expensive movies every year so as to remain profitable. The Software are used for predicting the success of a movie only when the pre-release details are known [5]. In this paper they used linear gradient descent algorithm. Other than these various attributes such as genres, seasons, MPAA movie ratings, and studio production companies are used. This study determines the profitability which comes out to be approximately 74.2%.

Deniz Demir, Olga Kapralova & Hongze Lai study adopts concepts of logistic regression, support vector machine, and multi-layer perceptron algorithms are used [6]. This project signifies each movie with the help of features, and then use Google search frequencies of these features for predicting the popularity of the movie. The main purpose behind this approach is that the movies that are mostly watched by the people should see the higher search volume of queries relating to the movie. This approach is being known as supervised learning problem, and hence logistic regression is used, SVM and multi-layer perceptron algorithms for this project. One of the main techniques used in this model is dimensionality reduction techniques for selecting the optimum set of features that are performed by one of the experiments. This model shows 72% accuracy.

Arundeep Kaur and AP Gurpinder Kaur proposed a study that uses concepts such as Neural networks and MATLAB algorithms are used [7]. This paper deals with the historical data collection of the parameters under study of various parameters including actors, producer, director, writer, music etc. After this the assign weights to the parameters and develop threshold for various prediction classes. Automation of the process is done using machine learning algorithm and evaluation is done using confusion matrix.

Author	Technique used	Evaluation matrics
K Meenakshi et al 2018 J. Phys.: Conf. Ser. 1000 012100,A data mining technique for analysing and predicting the success of a movie.	Basic concepts like decision tree algorithm and k means clustering are used.	
Quader, Nahid & Gani, Md & Chaki, Dipankar & Ali, Md. (2018). A Machine Learning Approach to Predict Movie Box-Office Success. 10.1109/ICCITECHN.2017.8281839.	Various algorithms like SVM and neural networks are used.	SVM-83.4% accuracy 88.8% accuracy Neural Network-84.1% 89.2% accuracy
Vr, Nithin & Babu Pb, Sarath. (2014). Predicting Movie Success Based on IMDB Data.	Concepts such as linear regression and logistic regression are used.	Logistic regression-42.2%accuracy linear regression-51% accuracy
Darin Im & Minh Thao Nguyen(2011). PREDICTING BOX-OFFICE SUCCESS OF MOVIES IN THE U.S. MARKET	bucketing method by a modified version of k-means clustering and method of Variance Minimization of the Gross	72% accuracy
Suhaas Prasad, Using Social Networks to improve Movie Ratings predictions, Dept.Elect.Eng, Stanford Univ., California, 2010	Algorithms such as KNN(k-nearest neighbours) and flixster data is used.	
Deniz Demir, Olga Kapralova & Hongze Lai December 15, 2012, Predicting IMDB movie ratings using Google Trends	Concepts of logistic regression,support vector machine, and multi layer perceptron algotithms are used.	72% accuracy
Arundeep Kaur, AP Gurpinder Kaur(2013) Predicting Movie Success: Review of Existing Literature .	Neural networks and matlab algorithms are used.	
Jeffrey Ericson & Jesse Grodman, A Predictor for Movie Success, CS229,Stanford University	Locally weighted linear regression and SVM are used.	SVM -70% accuracy Linear regression -88% accuracy

Figure.3. Table 1

Jeffrey Ericson & Jesse Grodman proposed a model that uses Locally weighted linear regression and SVM algorithms [8]. According to this paper, they ran a classification algorithm. This was done by calculating the median and predicting outputs with the use of linear regression. The accuracy of various outputs is calculated with the help of predicted and actual values ranging above or below the median where the prediction comes turns out to be a success. These are performed by holdout cross validation up to 70/30, 70% of data is being trained, and on the remaining 30%, testing is performed. In case of SVM they got a roughly of 70% success rate on each of the rating outputs and for each of the monetary gross outputs they achieved 88% success rate.

III. METHODOLOGIES

This segment portrays distinctive periods of data readiness alongside research technique which includes various data pre-processing techniques.

A. *Data attainment*

Dataset is basically taken from Kaggle, imdb, Metacritic and for post release it is taken from twitter. This dataset contains 5000 records and 28 parameters which includes historical data of each component such as actors, actress, budget, production cast, genres etc. Some features such as likes, retweet count, comments are extracted from twitter API and for this python is used. Few parameters are:

"movie_title"
"movie_imdb_link"
"duration"
"director_name"
"language"
"country"

"budget"
"release date"
"imdb_score"
"actor_1_name
" "gross"
"genres"
"vote_count"

B. *Data cleaning*

This phase focuses on data cleaning. We found out that there are many movies where some attributes are not available, so we removed those movies. For some movies budget is not available on IMDB, but we try to get it from other sources too. If not found we remove those movies from our dataset.

C. *Feature Extraction*

In this phase we will select that attributes out of 28 parameters which can be used to predict success of movie accurately. We will be using both pre-released as well as post released attributes. We take into consideration scores from various movie database websites for a particular movie. We also consider twitter reviews, likes, retweet count for a particular movie. The date is likewise a huge factor in the matter of film industry. Movies released before any occasion seems to have a greater chance to be fruitful.

D. Data integration and transformation

In this module we will classify our target class into three classes flop to superhit. We use algorithm like SVM and neural networks to predict.

Various parameters used in the system for movie success prediction are shown in Table 2.

TYPES	ATTRIBUTES
NOMINAL	ACTORS,DIRECTOR,WRITER,PRODUCTION-HOUSE,GENRE
NUMERIC	BUDGET, IMDB RATING, IMDB VOTES,NO OF RATINGS,META -SCORE, REVIEW ANALYSIS

Figure.4. Table 2

IV. APPROACH

After performing data pre-processing techniques, we implemented backward elimination approach so as to find out the relevant attributes from the dataset which are mentioned in the below table [fig.5].

RELEVANT ATTRIBUTES
1. GROSS
2. NUM_VOTED_USERS
3. NUM_FOR_FOR_REVIEWS
4. BUDGET
5. MOVIE_FACEBOOK_LIKES
6. IMDB_SCORE

Figure.5. Table 3

A. *Random forest*

One of the important supervised machine learning algorithms is random forest. They are an ensemble learning method for classification and regression. They are used without hyper-parameter tuning and produces great result. This algorithm helps in correcting the decision tree habit by overfitting to their training set by building multiple decision trees and merging them together to get accurate prediction and results.

B. *Lasso Regression*

Lasso regression analysis is a shrinkage and variable determination strategy for linear regression models. The objective of lasso regression is to acquire the subset of indicators that limits expectation errors for a quantitative reaction variable. The lasso does this by forcing a requirement on the model parameters that causes regression coefficients for certain factors to recoil toward zero. Factors with a regression coefficient equivalent to zero after the shrinkage procedure are barred from the model. Factors with non-zero regression coefficients factors are most firmly connected with the reaction variable.

C. *Decision Trees*

Decision Trees are basically a type of decision support tool that are based on the flowchart like structure, where each internal referred to as non -leaf node denotes a test on an attribute, one of the ways of display an algorithm that only consists of conditional control statements. Each branch in the decision tree speaks to the result of a test, while each leaf referred to as terminal node which holds a class mark. The node which is the highest one is the root node.

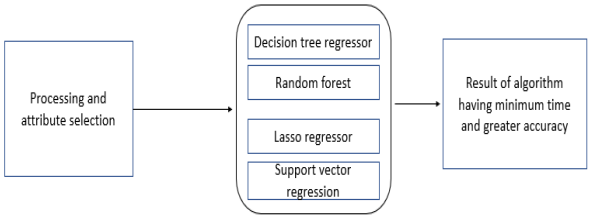

Figure. 6. Algorithm used

V. ANALYSIS

We have also performed an analysis on some attributes such as calculating top 10 actors of movies[fig.7.] and calculated the mean of a particular country to find out the movie success rate[fig.8.]. The graph has also been generated between time taken by each algorithms and algorithms implemented [fig 9]. Another graph between the accuracies and algorithms has been generated [fig 10]. We have also generated a heat map which determines the correlation between various attributes used in the model[fig.11].

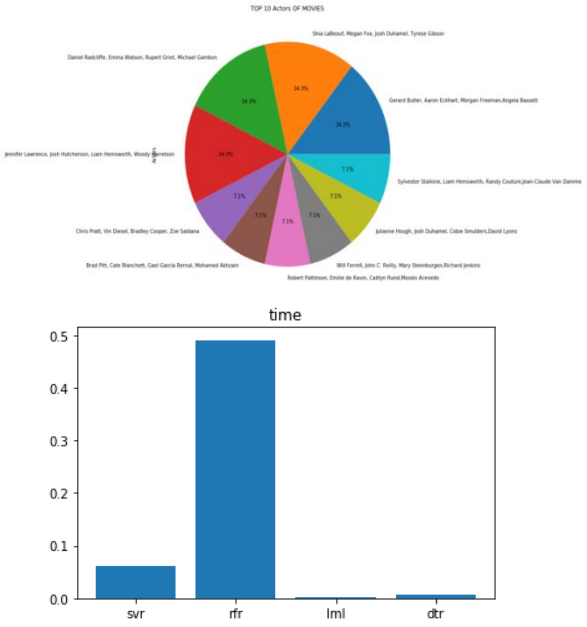

Fig. 7. Actors Analysis

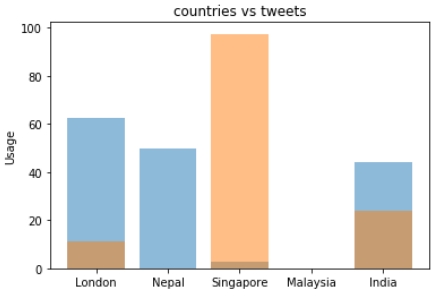

Fig.8 Country vs tweets

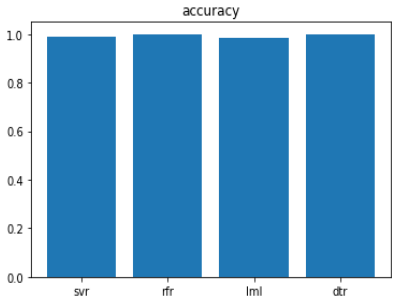

Fig.9. algorithms vs time

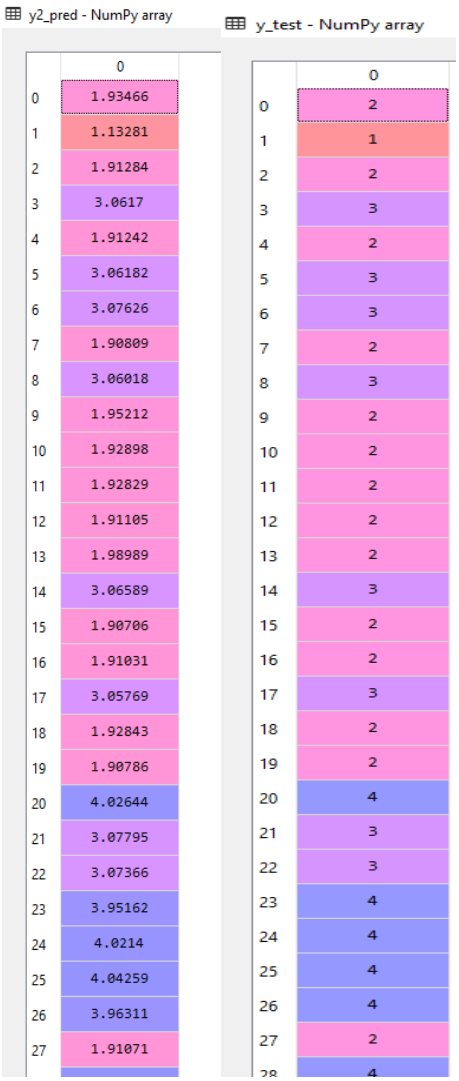

Fig.11. Comparison

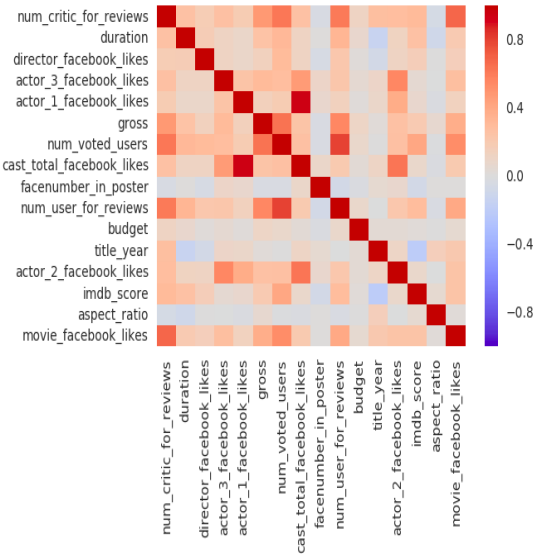

Fig.12. Heatmap

ALGORITHMS USED	ACCURACY
SUPPORT VECTOR REGRESSOR	98.81%
RANDOM FOREST	100.00%
LASSO REGRESSOR	98.45%
DECISION TREE REGRESSOR	100.00%

Fig.13. Table 4

VI. CONCLUSION

Using comparative analysis, we will get the most accurate results whether the movie well be a box office flop, average or superhit using various data mining and machine learning algorithms. A motion picture achievement does not depend just on those features identified with films. The quantity of group of onlookers assumes an essential job for a film to end up fruitful. Since the general purpose is about watchers, the whole business will have neither rhyme nor reason if there is no crowd to watch a motion picture. The quantity of tickets sold amid an explicit year can demonstrate the amount of audience on that year.

A portion of the research papers considered just pre-released highlights for motion picture forecast. Some others thought about by and large post-released data [6-7]. Be that as it may, in our research, we think about the two features for future prediction and furthermore prediction in the wake of opening end of the week.We implemented several algorithms such as Support vector regression (SVR), decision tree regressor and random forest. In case of decision tree regressor, we achieved accuracy of 100%, whereas in case of SVR we achieved a success rate of around 98.81% accurate, in case of random forest we achieved a prediction rate of 100% accurate and in case of Lasso regressor we achieved a success rate of 98.45%. While implementing these algorithms we came up with a conclusion that this is not a classification problem but a regression problem.

REFERENCES

1. K Meenakshi et al 2018 J. Phys.: Conf. Ser. 1000 012100, A data mining technique for analyzing and predicting the success of a movie.
2. Quader, Nahid & Gani, Md & Chaki, Dipankar & Ali, Md. (2018). A Machine Learning Approach to Predict Movie Box-office Success. 10.1109/ICCITECHN.2017.8281839.
3. Vr, Nithin & Babu Pb, Sarath. (2014). Predicting Movie Success Based on IMDB Data.
4. Suhaas Prasad, Using Social Networks to improve Movie Ratings predictions, Dept. Elect. Eng., Stanford Univ., California, 2010.
5. Darin Im & Minh Thao Nguyen(2011). PREDICTING BOX-OFFICE SUCCESS OF MOVIES IN THE U.S. MARKET.
6. Deniz Demir, Olga Kapralova & Hongze Lai December 15, 2012, Predicting IMDB movie ratings using Google Trends
7. Arundeep Kaur, AP Gurpinder Kaur(2013) Predicting Movie Success: Review of Existing Literature .
8. Jeffrey Ericson & Jesse Grodman, A Predictor for Movie Success, CS229,Stanford University.