2018, IEEE Third International Conference on Circuits, Control, Communication and Computing

Bollywood Movie Success Prediction using Machine Learning Algorithms

Ashutosh Kanitkar
Computer Science Department
Pune Institute of Computer Technology
Pune,India
ashbharadwajkanitkar@gmail.com

***Abstract*—Hindi Film Industry also referred to as Bollywood has now become a multibillion dollar industry and has also surpassed Hollywood in terms of amount of ticket annually sold. With so much money now riding on Bollywood movies it has become imperative to make accurate predictions about success of Bollywood films. Today even if a Bollywood movie does not become a hit at box-office the producer of movie will still make profits through sale of satellite rights and music rights but it is Distributors who suffer losses. Hence it has now become imperative for distributors to purchase distribution rights of movies at reasonable prices such that they can obtain profits from it rather than just break even. This problem is a supervised learning problem and will review regression techniques discussed in the literature for predicting the lifetime net India collections of Bollywood films as well as use classification methods on our Bollywood movie dataset for multiclass classification. An evaluation of all the approaches is proposed in which the accuracy score will be reported.**

Keywords—Bollywood, Movie Success Prediction, lifetime Net India Collection, Regression Techniques, Classification Techniques, IOT, CNN

I. Introduction

Hindi Film Industry also known as Bollywood movie Industry produces largest number of films every year. Hindi Film Industry has now become a multibillion dollar industry with it providing employment to a large number of people such as technicians, actors, producers as well as trade analyst. With so much money now riding on Bollywood movies it has become imperative to predict box office success of movies before the release of movie so that all the people who worked on the movie get appropriate remuneration. The success of movie is relative some movies are called successful based on income of movie while some are called successful based on good review from critics as well as good review from audience. In this paper I consider movie's success based on net box office India collection of movie only.

In this paper I have used 7 machine learning regression algorithm for predicting net India box office success of movie and 6 machine learning classification algorithm for predicting to which class the movie will belong to. For classification instead of considering binary classification of whether movie will be hit or flop I have rather chosen to classify movies into 9 classes of disaster, flop, below average, average, semi hit, hit, super hit, blockbuster and all time blockbuster based on the movies profit. Regression Algorithms implemented are Linear Regression, Polynomial Regression, KNN, Random Forest, Decision Tree, SVM and ANN. Classification Algorithms used for multi class classification is Logistic Regression, KNN, Random Forest, Decision Tree, SVM, Naive Bayes and ANN.

In the next section we discuss different research work related to movie success prediction. In section 3 we briefly discuss about our dataset and its features. In section 4 we discuss about data pre-processing which we performed before passing our data to our algorithms. In section 5 we discuss about our machine learning algorithms. In section 6 we discuss about our experimental results. In section 7 we discuss about future work which can be done using IOT and CNN to improve our model as well as conclusion is stated. Paper is concluded by mentioning some references.

II. Literature survey

Previously some researchers have tried to predict box office success of movie by applying sentimental analysis on review of audience and critics. Sources of data for sentiment analysis were facebook, instagram, twitter, snapchat.Nahid Quader, Md. Osman Gani and Dipankar Chaki [1] performed classification of movies into 5 categories using data from IMDB, RottenTomatoes, MetaCritic and BoxofficeMojo. Their features included both pre released features as well and post released features but their accuracy of pre released features model was only 49.3 percent for multilayered neural network model. Their one away accuracy was 84.1 percent. Travis Ginmu Rhee and Farhana Zulkernine [2] used neural network for predicting whether an movie is an hit or a flop and were able to attain an accuracy of 88.8 percent. The most recent work in literature is by Nidhi and Kaur[3] that tries to classify 111 Bollywood movies into 3 classes and very able to attain an accuracy of 93.1%.A similar study is done by Delen and Shard [4]. They use 834 movies from 1998-2002 and run it through a multilayer perceptron network. They try to classify each movie into one of 9 classes based on its box office revenue using 7 continuous variables and a 10-fold cross validation. Accuracy is measured by the percentage correct classification rate and the 1-away classification rate which resulted in a 36.9% accuracy rate and a 75.1% accuracy rate respectively.

None of above paper considered role that a previous week movie plays in collection of new release. Also none of above paper considered important roles such as amount of competition that the movie faces on release day as well as how many holidays are present in first week of release since if there are more holidays present in first week of movie there will be better chance for movie to have more collections as well .Also Bollywood films are synchronous

with music and none of above mentioned papers considers number of hit songs present in album..

III. Data description

Our dataset contains 250 bollywood movies released between 2014 and 2017.Our data sources are Wikipedia[5], RadioMirchi[6] and BoxOfficeIndia[7].Initially Our dataset contained 550 movies but out of them around 200 movies budget was not available. Out of remaining 350 movies number of hit songs of around 100 movies was not available as well as some other features were not found. We use only pre released features in our dataset as post released features won't be of any help to the producer or distributor during his assessment. There are a total of 32 features in our proposed data model. Conversion rate of a said person is the quotient of division between total net India collections of his movie by budget of that movie. Table below gives description of our dataset.

Feature Name	Type	Description
Genre	String	Genre of film to be released
Lead actor/actress/ supporting actors /director/ producer	String	Lead actor,actress,2 main supporting actors, director and producer of film
Solo release	Boolean	Whether film enjoys a solo release in cinema halls
Holiday release	Integer	No of holiday that are present in first week of film except Sunday
Budget	Integer	Total Budget of Film in Cr Rs.
Screens	Integer	No of screens in which movie is going to be screened in India
Sequel	Integer	Whether film is a part of franchise. If it is second film in franchise then it takes value 1 if 3th film in franchise then value 2 and so on
Runtime	Integer	Total Length of movie in minutes
Hit songs	Integer	Number of Hit songs present in movie album
Remake	Boolean	Whether Bollywood film is a remake of any other film
True Life Story	Boolean	Whether Film is an True Life Story adapted to the Big Screen
Book Adaption	Boolean	Whether Film is an adaption of an novel
Lead Actors/Actress/Supporting Actor(1)/ Supporting Actor(2)/Producer/Director Conversion Rate	Double	Average of conversion rate of last 5 films of lead actor/actress/ supporting actors/director/producer
Lead Actor/Actress/Supporting Actor (1)/Supporting Actor(2)/Director/Producer Average Collection	Double	Average of total net India collection of lead actor/actress/supporting actors/director/producer last 5 films in cr
Last Week Movie Performance	Boolean	It is set to 1 if movie released in last week has collected more than 30 crore Rs in its first week.
No of releases for that particular day	Integer	No of movies scheduled to release with the given movie on that particular day except itself
Big Film Coming	Boolean	Whether an much hyped movie or an movie with an big star is clashing with said movie

IV. Data preprocessing

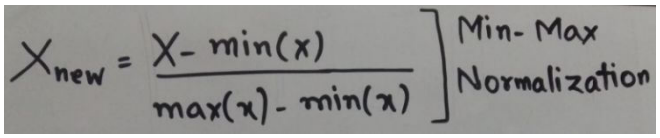
A. Preparing Input Data Items

We have ensured that while calculating the conversion rate we take only those movies of lead actor or actress in which he or she had a lead role and not a supporting role or cameo and it is similar for supporting actor and actress. Also in films where there are actually 2 lead heroes we have considered the one whose average conversion rate feature was higher. Similarly when 2 or more producer/production houses are jointly collaborating on a movie we have considered that production house whose average conversion rate feature was higher.

B. Normalizing the Data

First we convert each actor, actress, supporting actor/actress(1), supporting actor/actress(2) , director, producer/production house into an individual feature taking a Boolean value of whether they are present in movie or not. After conversion of all these String value into their respective nodes taking a Boolean value we now have a feature set of 965 features which will be passed as input to our proposed model. Now As Conversion Rate of Actor in Our Dataset are generally all between 0-5 we do not need to normalize it. On Other Hand all average collection features of lead actor and actress, supporting actor/actress(1) and supporting actor/actress(2) as well as director and

producer/production house have been normalized to have value between 0 and 1 from formula



V. Methods used

For our regression and classification models we have implemented 7 machine learning algorithms for performance assessment of our dataset. We implement artificial neural network using keras library and all other algorithms using Scikit Learn library in python.

A. *Linear Regression*

Linear Regression is one of oldest techniques used for predicting a continuous value. It tries to plot a straight line through given input features. Output variable y is represented as y=f(x) where f(x) =β0 + β1X1 + β2X2 +···+ βpXp[8] where X1..XP are input variables or known as independent variables and B0...Bp are coefficient and y is dependent variable.

B. *Polynomial Regression*

Polynomial Regression is used to add more complexity to our model. It is used when a polynomial curve fits data better than straight line. Output variable y is represented as y= y=f(x^n) where f(x) =β0 + β1X1 + β2X2 +···+ βpXp[8] where X1..XP are input variables or known as independent variables and B0...Bp are coefficient and y is dependent variable.

C. *Logistic Regression*

Logistic Regression is one of oldest techniques used for predicting a discrete value. Logistic Regression gives probability of y belonging to particular class given input features X. Logistic Regression works well for smaller datasets. Logistic regression uses sigmoid function [8] for calculating value of dependent variable.

D. *Artificial Neural Network (ANN)*

It is said that an artificial neural network would be able to recognize any pattern provided it has sufficient data and it has required amount of hidden layers. It is most complex of our machine learning algorithms.

E. *K Nearest Neighbors (KNN)*

KNN tries to predict output variable y by finding distance of input features x to features of all other points in our data set. Then parameter k is used to select k nearest neighbor and for classification problems it takes mode of k neighbors output value and for regression it takes mean of k neighbors output value. One of major disadvantages of KNN algorithm is that sorting of data is required.

F. *Random Forest*

Random Forest is a very powerful machine learning algorithm. It is a kind of Decision Tree. Random Forest helps in reducing over fitting by reducing trees having similar features. This algorithm works well when there are complex relationships between features.

G. *Decision Tree*

Tree based methods for regression and classification problem first divide the predictor space into distinct regions. In order to make a prediction about an observation we first find to which region that observation belongs and then take mean of all observation belonging to that region as output for regression and mode of all observation belonging to that region as output for classification problem. Decision Trees tend to sometimes over fit and hence methods like random forest, bagging and boosting are used to improve predictive accuracy of model.

H. *SVM*

SVM has been shown to perform well in a variety of settings and is considered one of best “out of box classifier”. SVM is generalization of maximal margin classifier [8]. SVM accounts for non linear boundary which maximal margin classifier cannot handle.SVM also uses concept of kernels which are generally used to avoid enlarging of feature space to achieve non linear boundary. We use different kernel functions like Gaussian radial basis function (RBF), linear kernel and polynomial kernel of SVM to have better accuracy.

I. *Naive Bayes*

Naive Bayes is one of simplest algorithm for classification and advantage of Naive Bayes lies in its training and prediction speed. This algorithm assumes that all features are independent.

VI. Results and Discussion

Some of techniques like Linear Regression, Logistic Regression and Naive Bayes are better for recognizing simpler patterns in data while other techniques such as ANN and Random Forest are better for identifying complex pattern in data. For Predicting net India box office collection and class to which movie belongs ANN performs better than others. We show accuracy for net India box office collection in terms of mean absolute error, mean squared error and root mean squared error. We show accuracy for multi class classification in terms of accuracy measures (number of examples that did not match with given examples) as well as precision score , recall score and F1 score are also reported. So precision formula is **TP/TP+FP**[8]. Formula of recall is **TP/TP+FN [**8]. Formula of F1 score is **2*(precision x recall)/precision + recall** [8]. Precision tells us how many of selected classes are relevant while recall tells us how many of relevant

classes are selected [1].Our neural network consist of 6 hidden layer along with an input and output layer. Hidden layer contains 30 units each except for first hidden layer which contains 60 units and activation functions are linear for hidden layer and relu for output layer for regression. ANN for classification consist of only 1 hidden layer containing 5 units and sigmoid activation function and input layer consisting of 965 units having linear function and output layer having 9 units and softmax function. Below Table shows accuracy of various algorithms for regression

Name of Algorithm	Mean Absolute Error	Mean Square Error	Root Mean Square Error
Linear Regression	24.07	1280	35.77
Polynomial Regression	19.34	795.18	28.19
KNN	22.2	1541.19	39.2
Random Forest	20.58	1055.47	32.49
Decision Tree	28.85	2545	50.45
SVM	20.36	1405	37.48
ANN	18.58	775.4	27.84

Results for Classification algorithms are as follows

Name of Algorithm	Accuracy	Precision	Recall	F1-score
Logistic Regression	45.33	0.64	0.45	0.51
KNN	49.33	0.86	0.49	0.62
Random Forest	49.33	0.41	0.43	0.37
Decision Tree	31.77	0.45	0.43	0.39
Naive Bayes	24	0.26	0.24	0.24
SVM	44	0.34	0.44	0.36
ANN	50	0.5	0.5	0.5

VII. Conclusion and Future Scope

As we can see from results Best regression technique for predicting Bollywood movie success is ANN and best technique for classification is KNN even though ANN had 0.7 percent more accuracy, precision and f1 score for KNN was significantly larger. Based on our results we found that the most prominent feature for success of Bollywood movie is presence of actor Aamir Khan followed by actor Prabhas. Also it has been observed from the results that success of movie which was released in previous week, competition from big star's film and more number of releases on that particular day does have an adverse effect on the success of a new movie. Additionally number of hit songs in album of movie and More Number of Holidays in the first week provide a huge positive effect in collection of movie. Hence using above mentioned features increases predictive accuracy.

The average price of ticket and the buzz on social media can also be great new features to consider. Furthermore sentiment of people while watching the trailer (both in case of theatre audience and home viewing audience) could be a pivotal feature to consider as well. People's sentiment about the trailer for theatre audience can be obtained by using IOT where camera sensors can be applied in theatre at multiple places to capture real time video of people while watching the trailer and this video can then be processed using Convolution Neural Network (CNN) frame by frame to find out people's sentiment about each scene of trailer. Similarly for people watching trailer at home, the webcam could be used to capture the person's video while watching the trailer and this video would be processed using CNN frame by frame to find out a person's sentiment about the trailer. After the person's sentiment is calculated, the smart device would then transmit this derived knowledge to secure server from which this knowledge would be retrieved for further use in our ML algorithms.

References

[1] Nahid Quader, Md. Osman Gani and Dipankar Chaki. "Performance Evaluation of Seven Machine Learning Classification Techniques for Movie Box Office Success Prediction" 2017 3rd International Conference on Electrical Information and Communication Technology(EICT).

[2] Travis Ginmu Rhee; Farhana Zulkernine. "Predicting Movie Box Office Profitability: A Neural Network Approach" 2016 15th IEEE International Conference on Machine Learning and application(ICMLA) .

[3] A. & Nidhi, A.P., 2013. "Predicting movie success using Neural Network". International Journal of Science and Research, India,online,vol 2(9),pp 69-71 .

[4] Delen, D. & Sharda, R., 2006. "Predicting box office success of motion pictures with neural network". Expert Systems with Applications,Elsevier, vol 30(2),pp 243-254..

[5] List of Bollywood films of 2014,2015,2016,2017,2018 retreived from https://en.wikipedia.org..

[6] Number of Hit Songs in album generated from https://radiomirchi.com.

[7] Information regarding budget and number of screens of movies from https://boxofficeindia.com

[8] Gareth James &Daniela Witten &Trevor Hastie Robert Tibshirani "An Introduction to Statistical Learning with Applications in R 6th Edition"