A Seminar Report on Performance Evaluation of Machine Learning Classification Techniques for Movie Box Office Success Prediction

Tasneem Naseem[1], Prof. Mali[2]

[1,2]*BE Computer Engineering, Vishwakarma Institute of Information Technology*

Abstract- **With the rapid increase of the multi-billion dollar movie industry, the volume of data that is generated on the internet related to movies is growing at a lightning speed. Machine learning methods have been used by researchers to build classification models. In this report, a variety of machine learning algorithms are used on a movie dataset for multi-class classification. The aim of this paper is to compare the various Machine learning techniques and conduct an analysis on each of their performances. In this report, the selected machine learning methods are Support Vector Machine (SVM), Logistic Regression, Multilayer Perceptron Neural Network and Gaussian Naive Bayes. All these techniques predict an approximate value of the net profit of a movie by examining the historical data collected from varied sources like IMDb, Rotten Tomatoes, Box Office Mojo and Meta Critic. The system predicts box office success based on features that are categorized as pre-released features and post-released features. The performance assessment of these four machine learning methods is done on a movie dataset that contains 755 movies. Among all these techniques, Multilayer Perceptron Neural Network gives a better outcome than the rest.**

Index Terms- Movie industry, box office success prediction, machine learning, classification techniques analysis, performance evaluation.

1. INTRODUCTION

1.1 Background

Predicting the box office success of a movie is complex and before analysing the movies, there is a need to define with success in correspondence with a movie. The definition of success in relation with a movie is relative, some movies are successful based on their worldwide net profit, where some may not have generated a big income but are successful based on good critics' review, popularity and ratings.

1.2 Motivation and Social Impact

The movie industry is a very large and ever-growing industry. For such a vast industry, it is difficult for business sectors to come to an investment decision. It is therefore important for these investors to have prior knowledge of the success of these movie ventures. This calls for a method of prediction which helps them to make an investment decision.

1.3 Objectives and Outcomes

In this research, the success of the movie is predicated on its profit only. Researchers show that almost 25% of movie revenue comes within the first or second week of its release [1]. Predicting the success of a movie before its release is difficult.

In this research, there are two set of features considered - pre-released features and post-released features. Pre-released features are responsible for predicting box office success of an upcoming movie whereas both the pre-released and post-release will be utilized to predict success after its release. Instead of a binary classification that measures whether a movie is a flop or blockbuster [2] the movies in this research are classified using multi-class classification. The movies are classified into one of five categories ranging from flop to blockbuster. In this research, two types of predictions are calculated: one is an exact match referring to a correct classification and the other is a one away match which takes into consideration either a class up or a class down from the class which is an exact match. The reason for this examination is that sometimes the accuracy of the result becomes low because of the marginal value of the classified classes. For prediction four machine learning algorithms, Logistic Regression, Gaussian Naive Bayes, Support Vector Machine (SVM) and Multilayer Perceptron Neural

Network are implemented. All these techniques are good for binary class classification and some of them provide results for multi-class classification as well. In this research, the dataset overlaps different classes at various points, hence making it very complex. However, the more complex a data pattern is the consistency with which Multilayer Perceptron Neural Network provides efficient result increases. Considering all the features, from 755 movies, Multilayer Perceptron Neural Networks correctly categorizes 442 movies in their respective classes. If we consider one away prediction the number of correctly categorized movies increases to 677. One way prediction is when the difference between predicted class and the targeted class is 1.From all the listed techniques MLP Neural Network (58.5%) and SVM (55.3%) work better than others.

1.4 Mathematical Model of Problem Solved

1.4.1 Logistic Regression

Logistic function:

σ (t) =1/1+e-t

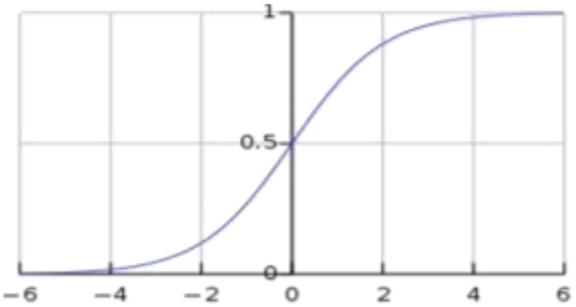

Figure 1. the standard logistic function

1.4.2 Multilayer Perceptron Neural Network

The activation functions are given by

$$y(v_i) = \tanh(v_i) \;\text{ and }\; y(v_i) = (1 + e^{-v_i})^{-1}$$

1.4.3 Gaussian Naive Bayes

The probability distribution of given a class C_k can be calculated by adding into the Normal distribution equation as follows,

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\, e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

2. LITERATURE SURVEY

2.1 Existing Techniques

Many research works previously have made movies' box office success prediction based on data available in the online database IMDb [4] - [6]. Few of them have given priority to the net box office earnings [7] - [9]. Most of the research works have used binary classification to predict the success of a movie and have classified it as success or failure. Success of a movie depends on admissible issues like the cast, the storyline, number of shows the movie is scheduled for, number of screens it will be played at worldwide, etc. In some previous works, prediction was made only on the basis of pre-release data [3]. Alternatively, few research works have adopted the approach of using Natural Language Processing for sentiment analysis on the assessment of audiences and critique of movie critics.

2.1.1 Audience, release and movie based classification

M. T. Lash and K. Zhao's [2] organized their classification analysis on released based, movie based and audience based while developing their classification model using machine learning techniques. Their data source was IMDb and Box Office Mojo. Their model contained movies that released only in the USA and excluded movies from all other countries.

2.1.2 Publicity on twitter

Siva Santosh Reddy et al. [10] focused on the propaganda on twitter for prediction of success. Their hypothesis was that the success of a movie majorly determined by its opening weekend revenue and the publicity it received from the audience. Publicity factor, number of screens where the movie was to be played at and the price of the movie tickets were taken under consideration during this evaluation. In this research, no sentiment analysis was performed to check for the positivity or negativity of the tweets posted. Additionally, their model used easy calculations.

2.1.3 Sentiment analysis of social media

Jonas et al. [11] computed the intensity of positive tweets on social media network IMDb's sub forum Oscar Buzz. Their model failed to predict correct outcome when some words were used for negative meaning. In their model they included different awards for casts and directors but neither of these categories were taken into consideration.

2.2 Comparison of Existing Techniques

In some research works, neural networks were used to predict box office success [3], [12]. Some research models were based on sentiment analysis of social media and social networks as well as publicity analysis [10], [11]. In those cases, Natural Language Processing was used to gather sentiment analysis of audience reviews. It also included number of reviews the movie received and their data source was mostly either IMDb sub forum Oscar Buzz, Twitter or YouTube.

2.3 Comparison and Analysis of Results of other researchers

In most of the research works mentioned, audience review and public ratings played an important role in predicting the success of a movie. But audience assessment isn't completely reliable and their reviews can be biased towards a particular actor or actress. Also, they excluded the movie critics' reviews. Furthermore, very few works considered the number of screens during prediction evaluation. Hence, the accuracy of these predictions will be questionable and will not generate suitable report.

For every movie, two types of data accessible through the internet, as discussed above are one pre-released data which includes the budget of the movie, its star casting and directors, etc. The other one is post-release data which is available on online databases such as IMDb, etc. Some researchers scrutinized both types of data but very few used this data in the implementation of their prediction model.

2.4 Technique/Algorithm

Four machine learning methods have been used for the performance analysis of the dataset in question. All these methods will be described in the following section. Most of these algorithms are implemented using python library Scikit Learn.

2.4.1 Logistic Regression

This regression model estimates the association between the categorical dependent variable values and one or more independent variables by evaluating probabilities using a logistic function [14]. In this research multinomial logistic regression is used as the outcomes can vary from class 1 to class 5. It is used for predicting dependent variables that perfectly fit within the limited number of categories, unlike in linear regression where the dependent variable gives a continuous outcome.

2.4.2 Support Vector Machine

This model is a supervised learning model with related learning algorithms which are used for classification and regression analysis on a given dataset. The algorithm is first fed with training examples that belong to a certain classification category, then an SVM training algorithm assigns new examples to one of the categories thereby making this technique a non-probabilistic binary linear classifier. [15]

The model is a representation of points in space, mapped in such a way that the examples are accurately mapped into separate categories. The clearer the gap between these classifications the better the accuracy. New examples are then mapped and categorized depending on which side of the gap they fall.

2.4.3 Gaussian Naive Bayes

Gaussian Naive Bayes comes under the category of Naive Bayes classifiers; this is a family of simple, probabilistic classifiers based on Bayes' Theorem that assumes strong independence between features.

A typical assumption in this algorithm while evaluating continuous data is that the continuous values associated with each class are distributed based on Gaussian distribution. This algorithm doesn't perform very well on the given dataset as it assumes that every feature is independent and the dataset used has dependent features. [13]

2.4.4 Multilayer Perceptron Neural Network

The Multilayer Perceptron is a class of feed forward artificial neural network. Every node except the input node is a neuron that is activated using a nonlinear activation function. It utilizes back propagation which is a supervised learning technique. [16]

Complex data patterns can be handled using MLP which makes it a good prediction model. In this research work Keras with tensorflow and SciKit learn is used to implement the MLP prediction model. The model used has three hidden layers and 15 characteristics. [13] Contrary to other techniques, MLP does not thrive on small datasets which makes it even more appealing to predict large dataset classifications.

3. IMPLEMENTATION

3.1 Flow of Work

The process followed to implement the prediction evaluation on this dataset is as follows:

3.1.1 Programming Language:

A programming language that supports machine learning API's and packages is selected based on requirement. Here, Python is selected and the packages used are Numpy, Scipy, SciKit Learn and Tenserflow.

3.1.2 Problem:

The problem in which you want to implement the algorithm or technique is selected. In this case, the movie dataset on which success prediction is to be computed is selected and all the 15 features of the dataset are defined and their range of values decided.

3.1.3 Algorithm

The algorithm which will be used for analysis is selected. While selection of algorithm its type of class and specific description of implementation should also be considered. Here, the algorithms used are Logistic Regression, Support Vector Machine, Gaussian Naïve Bayes and Multilayer Perceptron Neural Network. Here, all these algorithms are implemented to compare which model works better during prediction evaluation.

3.1.4 Research algorithm

The selected algorithms are thoroughly researched. Their implementation description, mathematical models, whether they work well as binary classifiers or multi-class classifiers, this helps in overcoming roadblocks that will be encountered during implementation of these techniques.

3.1.5 Unit Test

Unit test for each function is written down. A test driven deployment will help to understand the features and purpose of various algorithms. In this research a 10 fold cross validation is used for each technique and the final result is obtained.

3.1.6 Experimentation

The dataset is finally tested and its result is calculated. Here, the result includes exact class match percentage and one away class match. The percentage of accuracy of all these algorithms is compared and observation as to which technique has performed better is made. Here MLP Neural Network provides best result. It provides 58.5% accuracy for exact class prediction and 89.67% accuracy for one away class prediction.

3.1.7 Optimization

The algorithm can be further optimized by including more features in the dataset. Here, the political and economic factors haven't been considered. The efficiency of the result can be increased by including such factors during analysis.

3.1.8 Specialization

Once the result is obtained, specific methods can be used to enhance its optimization capability. Here, the precision, recall and f1 scores are calculated using true positive, true negative, false positive and false negative values to provide a better comparison of these algorithms

3.2 Data collection and Data sets

The dataset used contains 755 movies released during the time period 2012-2015. The data resources for the following research are IMDb, Meta Critic, Rotten Tomatoes and Box Office Mojo. Initially the dataset contained 3183 movies but were excluded as most of the required features, including the budget of the movie, were missing. After exclusion of such movies, the dataset reduced to 800 movies. Among these, budgets of certain movies were available but had other features that have been considered during evaluation, missing. After eliminating these movies the final dataset was a collection of 755 movies. Table I depicts the features of the dataset. Both pre-release and post-release have been utilized in the prediction model. A total of 15 features have been proposed in this model. Among these features, cast star power, no of screens, director star power, MPAA, budget and release month are pre-release features whereas features like IMDb Rating, Tomato Rating, Tomato Meter Audience Meter, Audience Rating, User Review, Metascore, Critics Review and IMDb votes are all post-release features.

In the evaluation of this dataset, MPAA rating which stands for Motion Picture Association of America, the Critics' Meter, Critics' Reviews, Audience Rating and Audience Review supplied by Rotten Tomatoes, IMDb Rating and Metascore from MetaCritic have been inspected. IMDb votes which is the number of viewers that reviewed a particular movie are also counted. The number of reviews is multiplied with their corresponding sentiment value and this collective attribute is used as a feature for the audience review of IMDb and the audience and critic reviews of Rotten Tomatoes. The star power of the cast and director have been calculated by summation of overall income of all the movies done by the specific cast or director during their career. Similarly the star power of a movie, is the addition of overall income of all the actors/actresses and directors of the movie.

TABLE I. Dataset with Description of All Features [13]

Features	Type	Description/Range of possible values
MDb Rating	Float	0 to 10
Tomato Meter	Integer	0 to 100
Tomato Rating	Float	0 to 10
Audience Meter	Integer	0 to 100
Audience Rating	Float	0 to 5
Metascore	Integer	0 to 100
MPAA	Integer	Value between 1 to 6 indicating G, PG, PG13, R, NC-17, NR respectively
Cast Star Power	Integer	Addition of all casts' lifetime gross income
User Review	Float	Sentiment value multiplied by no of reviews
Critics Review	Float	Sentiment value multiplied by no of reviews
IMDb Votes	Integer	Number of IMDb votes
Release Month	Integer	Between 1 to 12
Budget	Integer	Budget of a movie
No of Screens	Integer	Number of screens a movie released
Director Star Power	Integer	Addition of directors' lifetime gross income

In addition, the number of screens and release month of the movie has also been examined. The budget also takes inflation rates into account and adjustment is made accordingly. Considering the inflation rate is important, as the value of money is changing every day. The value of $200M ten years ago versus the value of the same amount today is not the same.

Instead of using binary classification to categorize the movies as either "Blockbuster" or "Flop", multi-class classification is used or Table II describes the categorization of the classes used.

TABLE II. Target Class Classification [13]

Target Class	Range(USD)
1	Profit <= 0.5M (Flop)
2	0.5M < Profit <= 1M
3	1M < Profit <= 40M
4	40M < Profit <= 150M
5	Profit > 150M (Blockbuster)

3.3 Results Obtained

All these machine learning techniques work well as classifiers. Logistic Regression and Naive Bayes are good for prediction on small datasets, like the one used in this research whereas Support Vector Machine and Multilayer Perceptron Neural Networks work better for dataset with complex pattern recognition. Among these four methods, Multilayer Perceptron Neural Networks obtains the best result. From the movie dataset of 755 movies, the aforementioned machine learning technique is accurately able to categorize 442 movies and 677 movies if one away prediction is included.

Pre-release features that are available online before the release of the movie and post-release features that are available after the premier weekend of the movie release have been examined in the prediction model. In Figure 2 and Figure 3 we can compare the differences in performance of various machine learning methods taking into consideration both one away prediction and the exact prediction. Figure 2 shows prediction based on pre-release and post-release features whereas Figure 3 shows prediction based only on pre-release features.

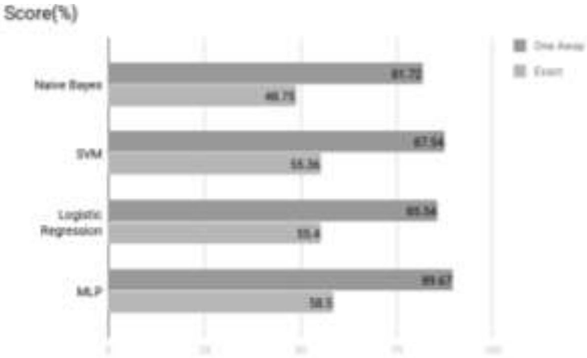

Figure 2. Performance comparison of different methods for all features. [13]

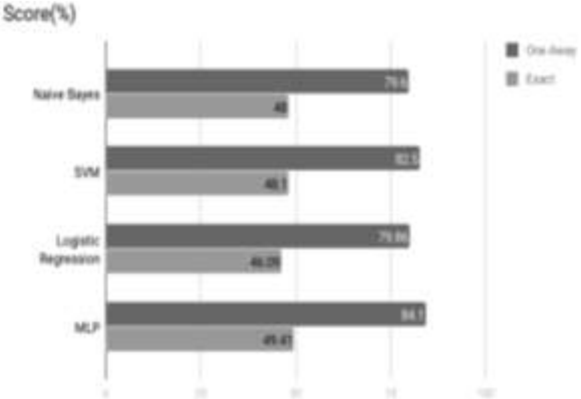

Figure 3. Performance comparison of different machine learning methods for pre-released features. [13]

While considering both set of features, MLP provides the best result as it is able to recognize complex patterns in data. It yields 58.53% accuracy for exact classes and 89.76 for one away classes.

4. RESULTS AND DISCUSSION

4.1. Discussion on Obtained Result

Logistic Regression and Naive Bayes are good for binary classification but with dataset that involve complex data pattern recognition Support Vector Machine and Multilayer Perceptron Neural Networks work better. As mentioned in the previous chapter Multilayer Perceptron Neural Networks produces a result with an accuracy of 58.53% for an exact class match and 89.67% for one away match

One way prediction is when the difference between predicted class and the targeted class is 1. For example, a movie that was categorized in class 4, but was predicted in class 5 which means that the movie was classified as a blockbuster hit. That means the classifier prediction was one class more than its true value. For an exact match classification prediction it will be taken as an error, but if we take one away classification prediction, it will be accepted as the correct outcome.

A major difficulty in the dataset model is that there are multiple overlapping data points as shown in Figure 4. This issue makes it difficult for algorithms to learn patterns efficiently.

Other methods like Support Vector Machine and Logistic Regression also perform well. Even though Support Vector Machine is a powerful classifier, the major problem arises the data regions need to be separated. As the data points in this research model overlap it often becomes difficult for Support Vector Machine algorithm to precisely separate the data regions. In this situation, Logistic Regression does a good job as this algorithm can effectively categorize data and works well on small datasets, like the one used in this research. We can observe that Naive Bayes technique under performs in comparison to the other three algorithms.

All these methods are implemented using a 10-fold cross validation which is the most efficient way to test. Table III includes all ten folds' accuracy of all four methods as well as their final accuracy. Figure 4 represents specific classes. Various regions of darkness represent specific classes and all the data points within the respective classes. Here, it can be observed that there is a mixture of one set of coloured region with the other coloured region. This is due to data point overlapping and indicate classification errors in the model. The X axis represents the budget of the movie while the Y axis depicts the number of screens.

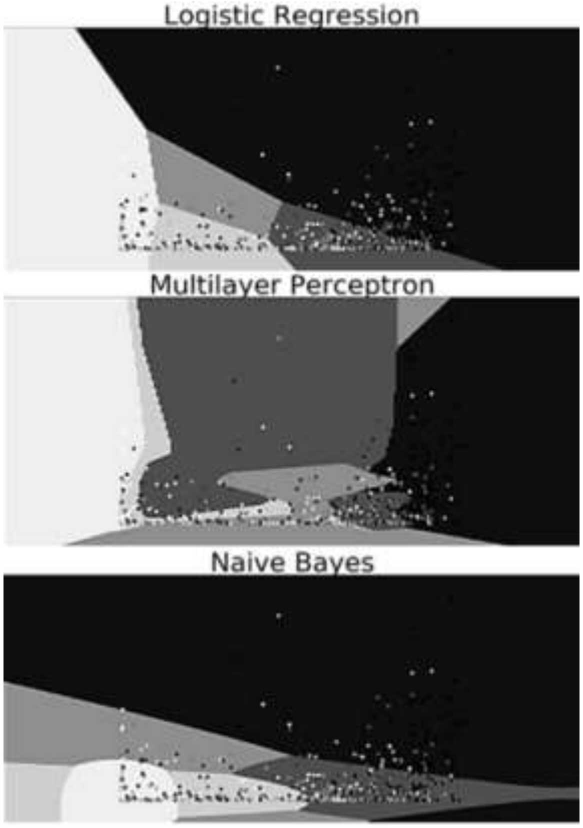

Figure 4. Data overlapping in different methods. [13]

TABLE III. Accuracy (IN %) Of Each Fold for Four Algorithms (All Features) [13]

Algorithms	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Final
MLP	58.45	63.93	62.33	55.94	67.17	54.67	51.86	57.1	63.52	49.62	58.33
SVM	53.24	46.75	62.34	59.74	50	57.34	53.34	54.67	54.05	61.12	55.26
LR	54.54	50.64	53.24	62.34	55.26	54.67	50.67	52	59.45	61.12	55.4
NB	37.67	51.94	51.04	52.84	55.26	49.34	40	48	50	51.38	48.75

The points overlap on each other due to which some data points become invisible. Different data points overlap distinct classes and classifiers are not able to properly determine the true classes of these data points. This problem is represented using a 2D graph as it is easy to understand.

The graph is made against budget and number of screens as these are the important constituents of success prediction of the research model and are available before the actual release of the movie, even then other features can be used. In Figure 4 dark shaded regions are different for different methods as their classification calculations are distinct. Each colour in Figure 4 represents a distinct class and colours of the data points are similar to that of the class to which they belong. For instance if the black region represents class 5 then all the black points belong to class 5. Again if you find some white points in the black shaded area, it means that the classifier has not accurately been able to classify these data points.

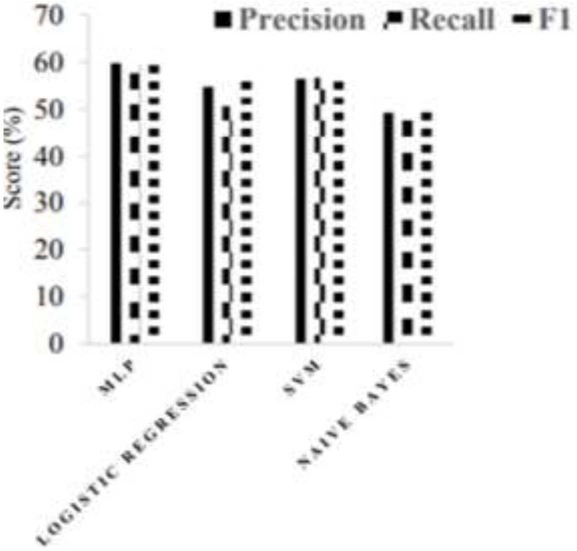

Figure 5. Precision, recall and F1 score for all features

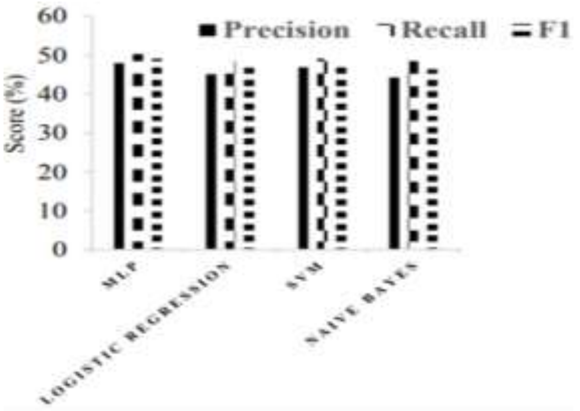

Figure 6. Precision, recall and F1 score for pre-released features.

In the following figures, Figure 5 and Figure 6 provide a visual depiction of the precision, recall and F1 scores of the four algorithms. Figure 5 demonstrates all features and Figure 6 shows scores for pre-release features only. These scores are calculated in a different manner and provide a good differentiation of performance evaluation of the four machine learning techniques. The important terms to understand before proceeding to see how the precision, recall and f1 scores are computed are true positive, false positive, true negative and false negative. These terms are explained in Table IV.

TABLE IV. Important terms to calculate precision, recalls and f1 scores

Value	Predicted Class	Test Class
True positive	True	True
False positive	True	False
False Negative	False	True
False positive	False	False

1. Precision Formula: TP/TP+FP; tells how many selected class are relevant
2. Recall Formula: TP/TP+FN; tells how many relevant classes have been selected
3. F1 Formula: 2*(precision x recall)/precision + recall; harmonic mean of precision and recall.

Among all these techniques, MLP gives highest results. While computing this exact prediction accuracy is taken in preference to one away prediction accuracy.

4.2 Comparison of Results

In this research sequel and genre of movie is excluded is excluded. Foreseeing the success of a sequel is difficult and an assumption that just because a prequel was successful, the sequel will also be successful is inaccurate. The aforementioned research works also excluded the genre and sequel [3]. In the earlier research works either the pre-release data was considered [3], [12] or post-release data was considered [5], [6] for prediction but in this research both pre-release and post-release features have been taken into consideration which provides a higher accurate result.

5. CONCLUSION AND FUTURE WORKS

The box office success of a movie depends on myriad characteristics of movies and not just a particular set of attributes. It also depends on external factors, especially audience and viewers. Critic reviews can also a pivotal factor in movie success prediction. External factors can also include political and economic factors of targeted countries. If the socio-economic condition of a country is unstable, even if the movie has a high budget the number of screens it will be played at will be substantially low as there is a very small audience. So including a country's GDP as a feature can be useful for further analysis. To increase the accuracy of the result it is advisable to include the number of viewers, this number can be obtained by counting the number of tickets sold annually.

ACKNOWLEDGEMENT

Acknowledgement It is matter of great pleasure for me to submit this seminar report on "SEMINAR NAME HERE", as a part of curriculum for "Bachelor in Engineering (Computer Engineering)" of University of Pune, I am thankful to my guide Prof. Guide name, Assistant Professor in Computer Engineering Department for his constant encouragement and able guidance. I am also thankful to Dr.B.S.Karkare, Principal of VIIT Pune, Dr. S.R. Sakhare Head of Computer Department for their valuable support. I take this opportunity to express my deep sense of gratitude towards those, who have helped us in various ways, for preparing my seminar. At the last but not least, I am thankful to my parent, who had encouraged and inspired me with their blessings.

BIBLIOGRAPHY

[1] 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.

[2] M. T. Lash and K. Zhao, "Early Predictions of Movie Success: The Who, What, and When of Profitability," Journal of Management Information Systems, vol. 33, no. 3, pp. 874–903, Feb. 2016.

[3] R. Sharda and E. Meany, "Forecasting gate receipts using neural network and rough sets," in Proceedings of the International DSI Conference, 2000, pp. 1–5.

[4] J. S. Simonoff and I. R. Sparrow, "Predicting Movie Grosses: Winners and Losers, Blockbusters and Sleepers," Chance, vol. 13, no. 3, pp. 15–24, 2000.

[5] A. Chen, "Forecasting gross revenues at the movie box office," Working paper, University of Washington, Seattle, WA, June 2002.

[6] M. S. Sawhney and J. Eliashberg, "A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures," Marketing Science, vol. 15, no. 2, pp. 113–131, 1996.

[7] D. Gregorio, "Prediction of movies box office performance using social media," Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13, 2013.

[8] S. Gopinath, P. K. Chintagunta, and S. Venkataraman, "Blogs, Advertising, and Local-Market Movie Box Office Performance," Management Science, vol. 59, no. 12, pp. 2635–2654, 2013.

[9] M. C. A. Mestyán, T. Yasseri, and J. Kertész, “Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data,” PLoS ONE, vol. 8, no. 8, 2013.

[10] A. Sivasantoshreddy, P. Kasat, and A. Jain, “Box-Office Opening Prediction of Movies based on Hype Analysis through Data Mining,” International Journal of Computer Applications, vol. 56, no. 1, pp. 1–5, 2012.

[11] K. Jonas, N. Stefan, S. Daniel, F. Kai “Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis” University of Cologne, Pohlig Strasse 1, Cologne, Germany.

[12] T. G. Rhee and F. Zulkernine, “Predicting Movie Box Office Profitability: A Neural Network Approach,” 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016.

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

[14] Rodríguez, G. (2007). Lecture Notes on Generalized Linear Models. pp. Chapter 3, page 45

[15] Cortes, Corinna; Vapnik, Vladimir N. (1995). "Support-vector networks". Machine Learning

[16] Rumelhart, David E., Geoffrey E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation". David E. Rumelhart, James L. McClelland, and the PDP research group. (editors), Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundation. MIT Press, 1986.