A Machine Learning Approach to Predict Movie Box-Office Success

Nahid Quader
Department of Computer Science and Engineering
School of Engineering and Computer Science
BRAC University, Dhaka, Bangladesh
Email: whereisnahidquader@gmail.com

Md. Osman Gani
Department of Computer Science and Engineering
School of Engineering and Computer Science
BRAC University, Dhaka, Bangladesh
Email: usmansujoy33@gmail.com

Dipankar Chaki
Department of Computer Science and Engineering
School of Engineering and Computer Science
BRAC University, Dhaka, Bangladesh
Email: joy.dcj@gmail.com

Md. Haider Ali
Department of Computer Science and Engineering
School of Engineering and Computer Science
BRAC University, Dhaka, Bangladesh
Email: haider@bracu.ac.bd

***Abstract*— 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. Moreover, we 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.**

Keywords— movie industry; machine learning; support vector machine; neural network; sentiment analysis

I. INTRODUCTION

The movie industry is a massive sector for investment but larger business sectors have more complexity, and it is hard to choose how to invest. Furthermore, significant investments come with more significant 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 considerable amount of data available on the internet, which makes it an exciting field for data analysis. Predicting a movie's box office success is a very complicated task to do. The definition of success of a film is relative, some videos 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 consider a movie's box office success based on its profit only. Researchers show that almost 25% of movie revenue comes within the first or second week of its release [2]. So it is hard to predict a movie's box-office success before its release date. In our proposed model, we use two types of features called pre-released features and post-released features. Only pre-released features are considered to predict the success of an upcoming movie. After releasing a film, both pre-released and post-released features will be available for further speculation. There are six pre-released features and nine post-released features. Instead of predicting only flop or blockbuster movies [3], we instead choose to classify a film based on its box office profit into one of five categories ranging from flop to blockbuster.

In this research, we have calculated two types of prediction, one is an exact match which refers correct classifications, and the other is, one away which means taking consideration of one class up or down from a particular type along with the exact match [4]. One away prediction implies a difference between predicted class and target class is 1. For instance, the movie "Interstellar" is a blockbuster movie. According to our classification, class 5 means blockbuster movie. So the target class for "Interstellar" is class 5. If our model predicts class 4 instead of class 5, it will be considered as one away prediction which is very close to exact prediction. From 755 movies 314 movies are predicted wrong, and among these 314 movies, 233 movies are anticipated only one class way from target class that is an excellent prediction comparing to other researchers who used multiclass classification. The reason behind considering one away is, the percentage of accuracy can be changed through changing the margin of target classes. For prediction, several machine learning algorithms are available such as Naive Bayes, Random Forest, and Logistic Regression, etc. These classifiers are good enough for binary classification, and some of them can be used for multiclass classification. However, when data pattern is very intricate, neural network and SVM consistently produced better result [5]. We have applied SVM and Neural network on our dataset for prediction. With all features in

978-1-5386-1150-0/17/$31.00 ©2017 IEEE

consideration, from 755 movies neural network correctly predicts 441 movies. If we consider one away prediction, the number of accurately classified movies become 674. Performance of SVM with different kernels is shown in section IV. In our case, the linear kernel gives 48.44% and 56.16% exact accuracy with 88.11% and 88.67% one away prediction accuracy for pre-released and all features respectively. Here neural network produces 48.41% and 58.41% exact match and 84.1% and 89.27% one away for pre-released features and all features respectively, which is an excellent prediction score. Section two of this paper describes previous works related to movie box office success forecasting; section three contains data description and research methodology; section four explains experimental results and evaluation. Finally, section five concludes this research with some general remarks.

II. Literature Review

In early days, many people prioritized gross box office revenue [6-8]. Few previous works portended gross of a movie depending on stochastic and regression models by using IMDb data [9-11]. Some of them categorized either success or flop based on movies' revenue and applied binary classifications for the forecast. The measurement of success of a movie does not solely depend on revenue. The success of movies relies on a numerous issue like actors/actresses, director, time of release, background story, etc. Further, few people made a prediction model with some pre-released data which were used as their features [4]. Few papers adopted many applications of NLP for sentiment analysis and gathered movie reviews for their test domain [12-13]. Again most of them considered audiences' reviews while they did not take the number of screens and movie critics' reviews in the account. Besides, audiences' reviews can be biased as a fan of actor/actress may fail to give an unbiased opinion.

M. T. Lash and K. Zhao's main contributions were, firstly they developed a decision support system using machine learning techniques, text mining and social network analysis to predict movie profitability [3]. Their research included several features such as dynamic network features, plot topic distributions meaning the match between "what" and "who" along with "what" and "when" and finally star power measurement based on their movies profits. They analyzed movie success in three categories: audience-based, released based and movie based. Their hypothesis based on the more optimistic, positive, or excited the audiences were about a movie, the more likely it was to have a higher income. Similarly, a movie with more pessimistic and cynical receptions from the public might attract fewer people to fill seats. They retrieved data from different types of media such as reviews and comments from Twitter, YouTube, blogs, articles and the sentiment of reviews or comments had been used as a way for assessing audience's excitement towards a movie. Their data were collected from both Box-Office Mojo and IMDb. They focused on the movies released in the USA and excluded all foreign movies from their experiment.

Sivasantoshreddy et al. tried to predict a movie box-office opening gross using hype analysis [14]. Their paper was focusing on twitter data for hype analysis. The main logic behind hype analysis was a success of a movie heavily depending on its opening weekend income and also how much hype it got among people before release. At first, they found the number of tweets about a movie by using web crawler. These tweets were collected on hour basis. There are three factors for hype measurement. The first factor was to calculate "No. of relevant tweets per second." The second factor was "Find the number of distinct users who posted the tweets." The third factor was "Calculate the reach of a tweet." For calculating the reach of a tweet, they counted the followers of a particular user. Their analysis based on hype factor, number of screens where the movies were released and the average price of all tickets per show. The model had simple calculations, and they just counted the number of tweets related to a movie, but they did not use any language processing to know if the tweet was positive or negative. On the other hand, a neural network had been used in the prediction of financial success of a box office movie before releasing the movie in theaters [15]. They converted their model into a classification problem categorized into nine classes. The model was represented with very few features.

In [16], the research tried to improve movie gross profit prediction through news analysis where quantitative news data generated by Lydia (quick text processing system for collecting and analyzing news data). It contained two different models, regression and *k*-nearest neighbor. But they considered only high budget movies. The model failed if the common word used as a name and it could not predict if there were no news about a movie. M.H Latif and H. Afzal used IMDB database only as their primary source, and their data was not clean [17]. Again their data was inconsistent and very noisy as they mentioned. So they used Central Tendency as a standard for filling missing values for different attributes.

Jonas et al. used sentiment and social network analysis for prediction and their hypothesis was based on intensity and positivity analysis of IMDb's subforum Oscar Buzz [18]. They had considered movie critics as the influencer and their predictive perspective. They used a bag of words which gave the wrong result when some words were used for negative means. There was no category award and only concerned with the award for best movie, director, actors/actress and supporting actors/actress. In some cases, success prediction of a movie was made through neural network analysis [4], [19]. Some researchers made a prediction based on social media, social network and hype analysis where they calculated positivity of reviews and number of comments related to a particular movie [20-23]. Moreover, few people had predicted Box Office movies' success based on Twitter tweets and YouTube comments. In both cases, the accuracy of prediction will be doubtful and will fail to give appropriate result. Again a small domain is not a good idea for measurement. In previous works, most researches were based on attributes that were either available before the release or after the release of a movie. Although some of the researchers had considered both types of attributes, in that case very few attributes were counted. The possibility of having better success in prediction goes higher with more attributes involved.

III. Data description and methodology

This section describes different phases of data preparation along with research methodology. These steps are data

acquisition, data cleaning, feature extraction, data integration, and transformation.

A. Data Acquisition

The first phase is data acquisition. This dataset contains 755 movies released in between 2012 to 2015 shown in figure 1. Recent movies are not selected because movie information is changing every day. Our data sources are IMDb, Rotten Tomatoes, Metacritic and Box Office Mojo. Some features are extracted directly from those websites while some are using different python APIs. Our system works on two scripts: one is a scrapper script to retrieve data from sites another script is to interact with APIs.

B. Data Cleaning

This phase is all about data cleaning. Initially, our dataset had 3183 movies. Then we recognize that there are many movies which do not have all data attributes available. Most of the movies do not have the budget available. We start checking with IMDb. While the budget is unavailable in IMDB, we check other sources for the budget. For some movies, we get the budget from Box-Office Mojo but for most of the movies in our dataset budget was unavailable in all sources. After removing those movies, there are 800 movies left. Among these, few movies do not have most of the features. After eliminating those movies, we finally make our dataset with 755 movies which have all information available. Table I shows the summery of our dataset.

C. Feature Extraction

In this phase, feature extraction is discussed. In previous works, very few features were considered in most of the models. We use both pre-released and post released features in our model. Total 15 features are used in our proposed model. We take in consideration of tomato critics' meter, tomato critics' rating, tomato audiences' score and tomato audiences' rating from Rotten Tomatoes, Meta score of Metacritic and IMDb rating from IMDb for a particular movie. We also count the number of viewers rated the movies. The multiplied value of rating and number of users who rated are used as a single feature.

Motion Picture Association of America (MPAA) is a governing body which rates a movie's suitability for particular audience based on its content. There are total six categories for each of a movie which is R, PG, PG13, G, NC, and NR. Here NR means the movie does not have MPAA rating.

This paper considers star power of actors, actresses, and directors. Celebrity like Brad Pitt, Tom Hanks, Julia Roberts and directors like Christopher Nolan, Quentin Tarantino are well known throughout the movie audiences. Famous artists like them not only make high-quality movies but also increase the probability of a movie to be successful. Even a substantial number of audience will go to the theatres just to watch their performances. So, star power plays a significant role in the movie industry, and many movies become blockbuster only because of stars involved in it. Star power of a single actor/actress is calculated by summing up the income of all movies in which that particular actor/actress has starred in [3]. This paper considers the total gross of all the movies of an actor/actress in their career as their star power. Sum of actor gross for all movies gives us the star power of a single cast member. Sum of star power for all casts in a movie is the star power for that particular movie. Director-star power is also calculated in the same manner.

TABLE I. DATASET SUMMARY OF ALL FEATURES

Features	Type	Mean	Median	Min	Max	Std. Dev	Data Source
IMDb Rating	Float	6.44	6.5	1.5	8.6	0.91828	IMDb
Tomato Meter	Integer	51.9735	52	0	99	28.6999	Rotten Tomatoes
Tomato Rating	Float	5.5404	5.7	0	9.2	1.79299	Rotten Tomatoes
Audience Meter	Integer	57.8558	58	0	94	19.0728	Rotten Tomatoes
Audience Rating	Float	3.4108	3.45	0	4.5	0.53570	Rotten Tomatoes
Metascore	Integer	50	52	0	100	20	Metacritic
MPAA	Integer	3.1284	3	0	4	1.0952	IMDb
Actor/Actress Star Power	Integer	9606441175	7999583961	0	50943162024	7846413921	IMDb
IMDb Review Sentiment value (Multiplied by no of review)	Float	172	91.71049	0	2298	246	IMDb
Rotten Tomato Review Sentiment value (Multiplied by no of review)	Float	75	67.9017	0	264	54	Rotten Tomatoes
IMDb Votes	Integer	110633	55918.5	655	1201640	148731	IMDb
Release Month	Integer	6.6251	7	1	12	3.4222	IMDb
Budget	Integer	41492981	20000000	20000	250000000	51823867	Box Office Mojo, IMDb
Number of Screens	Integer	1884	2275.5	1	4404	1549	Box Office Mojo
Director Star Power	Integer	1132921173	417051548	1441	13913175395	1741929261	IMDb

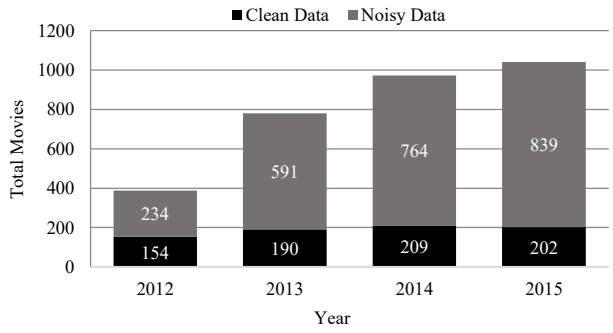

Fig. 1. Number of movies taken from each year

Besides, the release date is also a significant factor in the business of motion picture industry. Movies released before some festival or cultural event have a higher chance to be successful. We consider only release month as one of our features. In figure 2 we can see how the release month effects movie business. Here higher class number means more success. More movies of class 5 (Blockbuster) are released in May, June, November, and December than other months. In our dataset, most of the movies are produced in USA where May and June have summer blockbuster season starts at the USA from late May to late June. November and December are the festival season. Several festivals like Veterans Day, Thanksgiving, and Christmas take place during this period. So we can now relate why release month is essential.

Budget is another pre-released feature. If a movie has a higher budget for making the film it has higher chances to get more popularity for its publicity. So movies with the higher budget have higher chances to income more. In this paper, we calculate both budget and gross income without inflation rate adjustment.

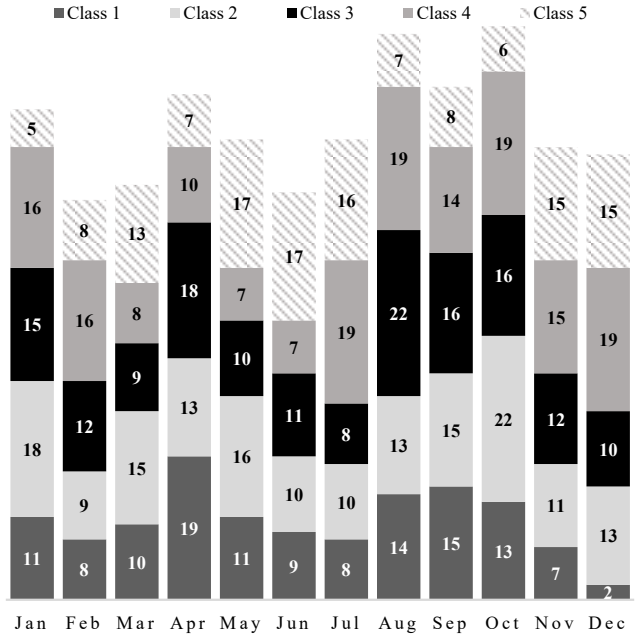

Fig. 2. Relation between target class and month of release

If a movie is released in more number of screens, the maximum number of people can watch the movie, and the maximum number of tickets can be sold. In figure 3 where highly profitable movies belong to the higher class, we can see the number of screens is increasing while class is rising. Class 5 (Blockbuster) includes 107 movies which were released in more than 3000 screens out of total 134 movies.

Further, we take critics reviews into account from Rotten Tomato along with the IMDb's and Rotten Tomato's users' reviews. We also count the number of reviews for each type.

D. Data Integration and Transformation

Table II shows the third phase, data integration, and transformation where we classify our target class into five classes. Rather than giving only two output "flop" or "blockbuster" [3], we make five classifications ranging from

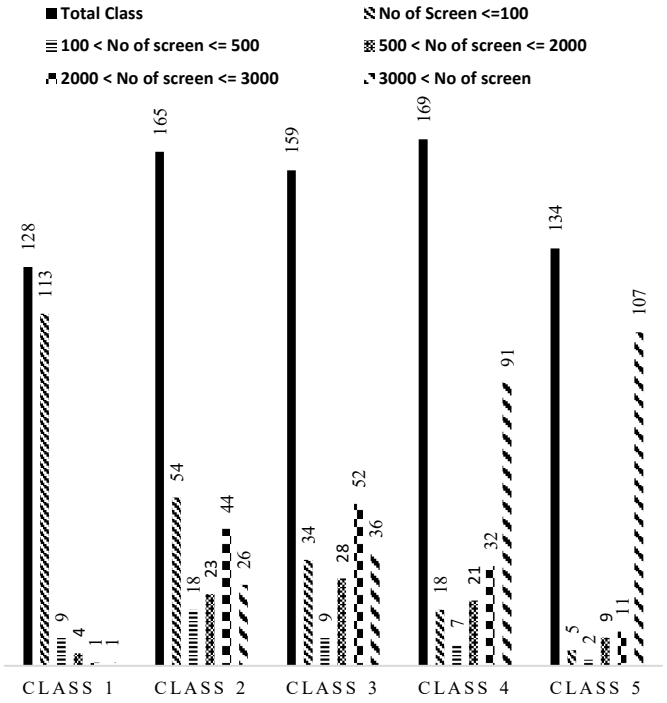

Fig. 3. Relation between target class and no of screen

"Flop" to "Blockbuster." We use multiplied the value of IMDb rating and IMDb votes and counted as one feature. We also consider sentiment values of different reviews for a movie with some reviewers in consideration.

TABLE II. TARGET CLASS CLASSIFICATION

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)

IV. EXPERIMENTAL RESULT AND EVALUATION

This phase is segmented into three parts; first one is sentiment analysis of the collected reviews, next one is about SVM, and the last one describes the neural network.

A. Sentiment Analysis

In total 212,535 reviews from IMDb and 108,464 reviews from Rotten Tomato are inserted in our dataset. Sentiment values of those are calculated with Microsoft Power BI Desktop application. With Power BI Desktop we use Microsoft Azure's cognitive service of Text Analytics API. It gives sentiment value ranging from 0 to 1 where a higher value indicates positive feeling. After getting those sentiment value for each review, we calculate the mean sentiment value for a single movie and multiply it by the number of reviews. And this multiplied value is considered as a feature for all movies.

B. Support Vector Machine (SVM)

We implement 10-fold cross-validation in each of our experiments. In 10-fold Cross-validation, all the elements in our dataset are divided into ten groups. The first group becomes the test data and rest nine groups make the train data for the machine, and we write down its testing accuracy. After testing the first group, the second group becomes the testing data and rest groups make the training data for the machine. In this way all data are tested and mean is calculated from the accuracy of each fold. We use four kernel functions in SVM; the first one is the linear kernel, the second one is Gaussian radial basis kernel then 3-degree polynomial kernel and last one is linear support vector classifier, (linear SVC) kernel. Both linear kernel and linear SVC have same calculation but in different approaches. Kernels are useful for higher dimensional data as in practical life it is hard to calculate when we have more than two or three dimensions. We have 15 features as variables in this model and kernel function can work on infinite dimensional space. For this reason, we choose to apply all those kernel functions in our model. We use Scikit-Learn for the implementation of SVM [24].

In Table III, we can see the exact and one away prediction accuracy for SVM. Different kernels give us slightly different results. Among all those kernels RBF and Linear give a comparatively good result. Confusion Matrix for Linear is shown in figure 5 for all features. When data are overlapping, SVM is unable to make hyperplanes properly. We know 2D plotting is a good way to visualize and understand the vector regions and data relations.

TABLE III. PERFORMANCE COMPARISON OF SVM

Kernel	Exact (Pre Released)	Exact (All)	One Away (Pre Released)	One Away (All)
Linear	48.44%	56.16%	82.11%	88.87%
RBF	49.54%	55.36%	82.25%	87.54%
Polynomial	46.00%	52.58%	83.44%	85.82%
LinearSVC	48.47%	53.64%	82.12%	85.43%

In figure 4, we can see the data points of only two features and classified vector regions. For most of the features, we got same characteristics as this one. We can see that most of the data points are overlapping with each other. That was a problem for SVM to separate hyperplanes properly. Also, regions with different darkness show different vector regions classified by the model as well as the darkness of the data points. If a data point with specific darkness level is found in an area with different darkness level, it indicates classification failure in figure 4.

C. Neural Network Analysis

A Multi-Layer Perceptron neural network (MLP) is used for prediction shown in figure 6. This MLP model is developed using Keras [25], a famous python API for the neural network. Keras sequential model is used to build the model. Scikit-Learn [24] is also used for 10-fold cross-validation. In the proposed model there are three hidden layers, each has sixteen neurons. The input layer has fifteen nodes, and the final layer has five nodes for five outputs.

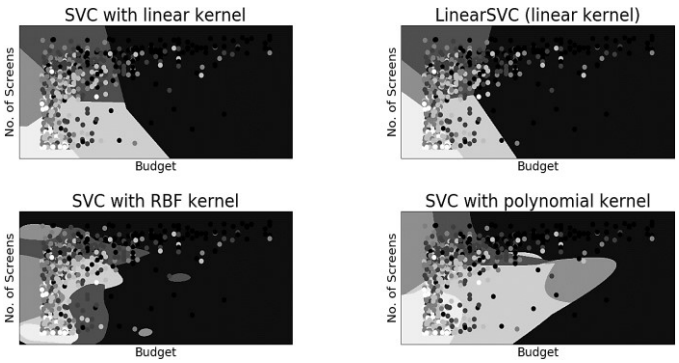

Fig. 4. Data overlapping problem of our dataset

We recognize that three hidden layer MLP architecture gives better result consistently by experimenting a different number of hidden layers. Figure 6 represents our MLP model. For hidden layers, all nodes are not shown in figure 6 to avoid unnecessary complexity [4]. Softmax and ReLu activation functions are used in this model for final layer and hidden layers accordingly. Overfitting is a common problem in Neural Network. For overfitting problem, dropout regulation is inserted after each hidden layer.

TABLE IV. PERFORMANCE COMPARISON OF NEURAL NETWORK

Features	Exact Match	One Away
Pre Released Features	48.41%	84.1%
All Features	58.41%	89.27%

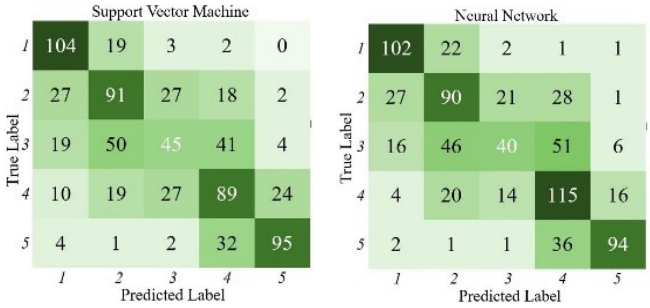

Fig. 5. Confusion Matrices for SVM (Linear Kernel) and MLP (All Features)

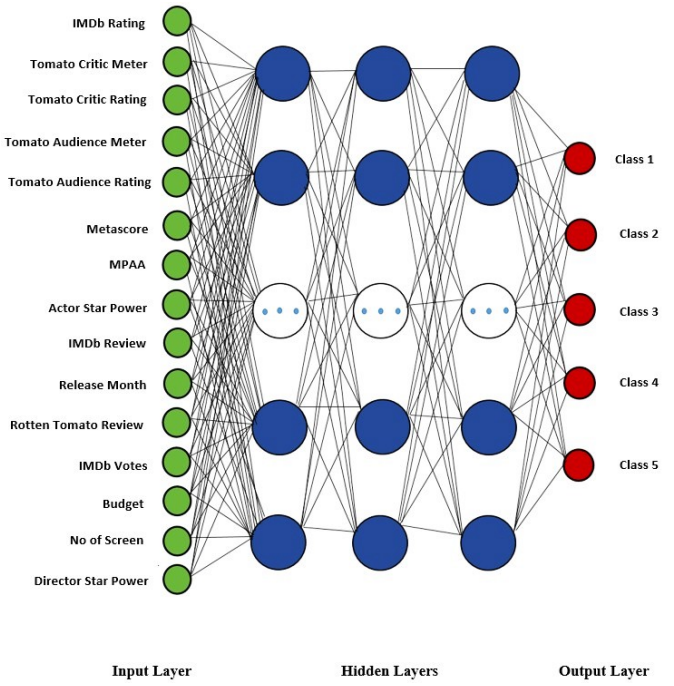

Fig. 6. Multi-Layer Perceptron Neural Network

When our MLP model is trained using only pre-released features, we get 48.41% exact match and 84.1% accuracy if we consider one away prediction shown in Table IV. Similarly, if we consider all features, we get 58.41% exact match and 89.27% one away match. Unless only hit or flop, we make a prediction ranging between flop and blockbuster movie [3], [6-7], [9]. Here in figure 5, two confusion matrices of SVM and MLP are shown with all features in consideration.

In figure 5 there are two confusion matrices. Left matrix is for SVM, and the right one is for the neural network. Both confusion matrices are a 5x5 matrix as we have five classes. There is cell number for both matrices. If we watch all values in a row, we will get the total movies of that particular class. In cell [1, 1], 104 means among 128 movies of class 1, 104 movies are precisely predicted as class 1, 19 movies are predicted as class 2, 3 movies are predicted as class 3, and 2 movies are predicted as class 4.

We can see both pre-released and post-released features takes part in prediction. Among all the pre-released features we extract which feature is most important for forecasting. For our dataset we find out the most crucial pre-released feature is Budget as expected. More budget means more marketing, publicity and recruiting more celebrities, which help a movie to become a blockbuster hit. The second one is the no of screens, which makes sense because more screen means more audiences. The top three features among all six pre-released features are budget, no of screens and release month. In figure 7, feature importance of all pre-released features is visualized.

Neural network and support vector machine both give good results, but neural network produces better prediction accuracy than SVM. For exact prediction neural network achieved 48.41% accuracy which is better than previous research works.

Some research works achieved high accuracy with the only post released features, where some paper used binary classification only [3]. If we reduce the number of target classes, our prediction score will improve. The problem with SVM is, it suffers from separating data points correctly because of data overlapping (shown in figure 4). In that case, SVM is unable to calculate proper hyperplanes as it becomes confused for very frequent data overlapping. In this particular situation, the neural network is performing better for classification and pattern recognition than other machine learning algorithms. Our MLP neural network will play much better if we have more data in our hand. But the final accuracy we reach is already a very good score compared to other papers in this field [4], [15].

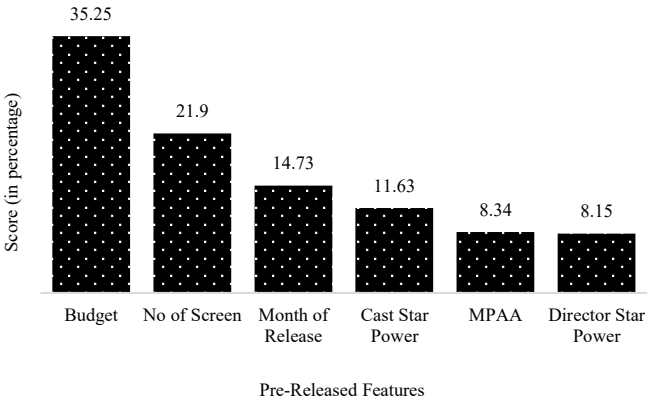

Fig. 7. Feature importance of pre-released features

V. CONCLUSION AND FUTURE WORKS

A movie success does not depend only on those features related to movies. The number of audience plays a vital role for a movie to become successful. Because the whole point is about viewers, the entire industry will make no sense if there is no audience to watch a movie. The number of tickets sold during a specific year can indicate the number of viewers of that year. And the role of movie audience depends on many situations like political conditions and economic stability of a country. A country's GDP rate can be used as a feature to know if there is financial stability during the period when a movie is released. During an economic depression, very few amount of audience will go to the theatres to enjoy movies. So these facts play a vital role in an ultimate success of a movie. So, for future work, we suggest considering these features.

We do not consider genre and sequel to a movie as features. Prediction of a sequel movie is terrible. Some movies gain a handsome amount of profit only for its previous sequel. Some other research papers also avoided sequel [3]. Some research papers considered just pre-released features for prediction [11], [15]. Some others considered mostly post-released data [6-7]. But in our research, we consider both features for future prediction and also prediction after opening weekend. Support Vector Machine (SVM) has accurate prediction rate of 56.16% and 88.87% one away prediction accuracy for all features. While

it gives exact and one away accuracy of 48.44% and 56.16% for pre-released data which is overall a good score. But Neural network gives 58.41% exact prediction with 89.27% one away prediction for all features and 48.41% exact match with 84.1% one away match when only six pre-released features are taken in consideration, which is a very good score.

REFERENCES

[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] 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.

[3] 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.

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

[5] N. Quader, Md. O. Gani and D. Chaki, "Performance Evaluation of Seven Machine Learning Classification Techniques for Movie Box-Office Success Prediction," in *3rd International Conference on Electrical Information and Communication Technology (EICT)*, 2017, pp.1-6.

[6] K. R. Apala, M. Jose, S. Motnam, C.-C. Chan, K. J. Liszka, and F. 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] B. Pang and L. Lee, "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, July 2002, pp. 79–86.

[13] P. Chaovalit and L. Zhou, "Movie review mining: a comparison between supervised and unsupervised classification approaches," in *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, 2005.

[14] 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.

[15] R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks," *Expert Systems with Applications*, vol. 30, no. 2, pp. 243–254, 2006.

[16] W. Zhang and S. Skiena, "Improving Movie Gross Prediction through News Analysis," *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 2009.

[17] M.H Latif, H. Afzal "Prediction of Movies popularity Using Machine Learning Techniques," National University of Sciences and technology, H-12, ISB, Pakistan.

[18] K. Jonas, N. Stefan, S. Daniel, F. Kai "Predicting Movie Success and Academi Awards through Sentiment and Social Network Analysis" University of Cologne, Pohligstrasse 1, Cologne, Germany.

[19] 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.

[20] J. Duan, X. Ding, and T. Liu, "A Gaussian Copula Regression Model for Movie Box-office Revenue Prediction with Social Media," *Communications in Computer and Information Science Social Media Processing*, pp. 28–37, 2015.

[21] L. Doshi, J. Krauss, S. Nann, and P. Gloor, "Predicting Movie Prices Through Dynamic Social Network Analysis," *Procedia - Social and Behavioral Sciences*, vol. 2, no. 4, pp. 6423–6433, 2010.

[22] T. Liu, X. Ding, Y. Chen, H. Chen, and M. Guo, "Predicting movie Box-office revenues by exploiting large-scale social media content," *Multimedia Tools and Applications*, vol. 75, no. 3, pp. 1509–1528, Feb. 2014.

[23] Z. Zhang, B. Li, Z. Deng, J. Chai, Y. Wang, and M. An, "Research on Movie Box Office Forecasting Based on Internet Data," *2015 8th International Symposium on Computational Intelligence and Design (ISCID)*, 2015.

[24] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[25] F. Chollet, Keras (2015), GitHub repository, https://github.com/fchollet/keras. [Accessed: March-2016]