A Detailed Review on Predicting Box-Office Success of a Movie

S Jahangeer Sidiq[1], Majid Zaman[2]
[1]*Department of Computer Science, University of Kashmir, J&K, India*
[2]*Directorate of IT&SS, University of Kashmir, J&K, India*
(E-Mail: jasmi800@gmail.com, zamanmajid@gmail.com)

Abstract**—**This paper reviews the state-of-art techniques used for predicting success of movies prior to their release at box-office. Machine learning classification techniques were used for predicting Hit/Flop and regression techniques were used for predicting revenue of a movie. The problem has been studied as both binary and multi-class problem in literature. Not only the metadata of movies was used for prediction task but also social networking platforms such as Twitter, YouTube, Wikipedia and online blogs were used as sources of data. Some of the works in literature used a combination of both types of features thereby increasing the prediction accuracy. The main aim of all these works in literature was to assist all the stakeholders of film industry in decision making process

Keywords—Machine learning, Classification, Regression, box-office, social networks.

I. INTRODUCTION

Success of business is of great interest to financial experts and economists. The historical data has been used for predictive analysis using different machine learning techniques in different domains. Such types of studies have also been performed in predicting success of movies. Large number of studies have been performed in this domain because of the interest of public in entertainment industry and large volumes of data are freely available on web such as IMDB [1].The conventional attributes from IMDB database were used for prediction of movie success in most of the studies. With the advent of social media various events from forums like Twitter and YouTube have been harnessed. Social media websites contributed a lot to the popularity of movies. Anyone can comment, rate, review or share opinions about a movie on internet. Thus social networks play a major role for predicting success of a movie. Researchers believe that classical factors along with social factors are necessary for this purpose. Twitter has gained popularity among all the mediums there by making it a point of focus for researchers by making use of sentiment analysis. Moreover some of the studies in this domain showed that sentiments about movies is not a top factor for determining the movie success prior to its release.

II. LITERATURE REVIEW

The quality of a movie can be determined by its rating and it is a critical indicator whether a movie will be watched by a customer. Hence it is an important research challenge to predict a movie rating before its release or production. Most state-of-art approaches predict rating of movie based on comments, reviews from social networking websites that are generated post-production which is one of the biggest reasons for failure of such approaches. Such approaches are not applicable until the movie is released. In [2] a generative convolutional neural network based regression model is proposed by Ning et .al (2018) for predicting the rating of a movie. This regression model makes use of parameters that are available before the production of a movie such as budget, director, plot information, genres and cast. The model can serve as a tool for all stakeholders in decision making for movie production and can assist users for choosing the movie to watch. The experiments on real datasets demonstrate the effectiveness of this approach.

In [3] Gaikar et .al (2015) used twitter data for prediction of box office collection prior to the movie release. The sentiments were mined from tweets for movie performance prediction. The FIS model is used for predication on Bollywood movies which were released in 2014.The models output has been compared with movie data on movie-based sites. Experiments on movie data confirmed the effectiveness of the FIS model. The predicted results were found approximately same as the actual results. The opening weekend box office collection of a movie prior to its release was predicted using hype factor by making use of tweets. The MSE method was used for predicting accuracy. The proposed model can be very useful for decision making in business for movie distributors. Audience usually watch a movie because of its positive word of mouth. In this work a prediction of box office collection prior to a movie release has been made and it was found that hype factor and prerelease sentiments have a crucial role in the success of a movie.

Authors Lee et .al (2018) in [4] predicted the box-office performance of movies. For improving accuracy of prediction (Cinema Ensemble Model) CEM has been proposed .Based on the theoretical back ground a new attribute namely transmedia storytelling has been introduced. The accuracy of 58.5% has been achieved thereby enhancing performance of the models from past research studies. Both practically and academically our study has numerous good results. This work has uniquely focused on the feature aspect of a predication of movie success. An idea of choosing attribute based on concrete theories. The most of the studies focus on predictive power enhancement using machine learning algorithms for box office movie success prediction. And without overseeing the features that can contribute to the better performance of the model. Moreover identification of which machine learning algorithm is more appropriate for movie domain and build an ensemble based model CEM that has rarely been implemented in previous studies. It was observed that 10% accuracy was increased by CEM in comparison to previous studies.

In [5] the authors Jaiswal et .al (2017) created a model using machine learning algorithms that can predict whether a Bollywood movie will be hit or flop prior to its release. Multiple sources of data were used for carrying out the experiments like BoxOfficeIndia, Cinemalytics, Wogrna and YouTube. The features like music score has been included in the data that is unique feature in Bollywood movies and contributes to high accuracy of prediction. This problem has two types of labels only Hit or Flop. After evaluating several classifiers ensemble approach namely Bagging algorithm was used for creating model. Here factors from different sources were combined to create a model with high predictive accuracy for Bollywood movies. As per the proposed model the collection at a box office is highly dependent on (1) Total number of screens on which a movie is released (2) song quality of a movie also contributes to the movie success. (3)The impact of actor/actress also contributes to the success (4) Time of release of a movie also has an impact on success. The sentiment analysis from social media also contributes to the movie success prediction for English movies. But this is not the case with Bollywood movies under study because of the language barrier that restricts the ability for proper analysis.
Authors Henry et .al (2007) in [6] classified the movies data in to nine classes ranging from flop movie to a blockbuster movie. The results of datasets are presented as average of one-way and bingo predictions. In our model, the forecasting problem is converted into a classification problem .The results of artificial neural networks showed an improvement of 50% on bingo and 90% on one-way.

Zhang & Skiena (2009) in [7] studied correlation of IMDB data and movie news data with movie grosses and models are built using news data, IMDB data and their combination respectively. The predictive power of media was proved using experiments. It was observed experimentally movie news references are correlated highly to movie grosses including sentiment measures. Also prediction of movie gross can be performed by news data, IMDB data or their combination as well. Models obtained by using news data can perform similar to the models created using IMDB data. While as models obtained using both types of data yield best results. KNN classifier and regression were employed for gross movie production. KNN performs better than regression models using the same indicators. Sentiment data are not good predictors for regression based models but are good for KNN models.

Yoo, Kanter & Cummings (2011) in [8] found that using numeric, sentiment based and text based attributes from IMDB linear regression outperformed classification logistic regression for gross revenue prediction. However none of them gives precise results for use in practice.

Oghina et .al (2012) in [9] predicted the movie rating using social media data. The quantitative and qualitative activity indicators were identified from social media and two sets of features both textual and surface were identified. The number of dislikes and likes on YouTube together with textual attributes from tweets lead to the best performance of model.

Jeesha et .al (2018) in [10] used historical data from Bollywood industry for developing a model to predict the box office success of a movie prior to its release. The data set contained a total of 447 movies over a time span of 9 years. The features were chosen from literature and the same features were used for Bollywood dataset of Indian movies. In addition several other features that are unique to Bollywood industry were studied for studying their impact on box office success of a movie. The factors like screen count, budget, genre and release period have significant impact on outcome of prediction. However the features like lead actor, music director, director and sequel were not found to be too significant for prediction as is popularly believed.

Latif & Afzal (2016) in [11] used IMDB data for classification and it was found that impressive results were achieved through simple logistic and logistic regression around 84%.The features which contributed to the information are number of votes for each movie, metascore ,Oscar awards won by movie and the total number of screens on which movie was screened.

Apte et .al (2011) in [12] used k-means clustering for dataset separation between high-release and low-release movies and separating different genres for taking in to account the diversity of movies, the predication accuracy was improved significantly using weighted linear regression and simple regression technique. It is not possible to achieve accuracy of greater than 20% in some of the cases. Moreover some of the genres lack enough data for predications thereby making prediction very difficult.

Liu & Zhao (2015) in [13] proposed a novel method for forecasting movie success by combination of rough set reduction and SVM classifier. It was observed that its accuracy was higher by 10% than that of MLP. So this model is very effective for meeting industrial requirements and can be used by producers and investors of a film for rational decision making for making economic benefits.

Ahmed et .al (2015) in [14] used machine learning classifiers for model creation using classical features from movie data set and features extracted from social media such as Text comments , Tweets and YouTube. And it was experimentally observed that features extracted through social media has more predictive power than those of conventional features. The values of 77% and 61% were obtained for Rating and Income while as conventional attributes predicted with the accuracy of 76.2% and 52% respectively. It was observed that blend of both types of features can outperform the existing approach.

In [15] authors Parimi & Caragea (2013) constructed a graph network in between movies and used for reducing the movie independence assumption that is made my traditional algorithms. A network for movies is used with a transductive algorithm for feature construction. Then a classifier is learned for classifying a new movie with respect to box office collection. It has been observed experimentally that proposed method improves classification accuracy as compared to previous methods in literature.

Mundra et .al (2019) in [16] aims at predicting a movie's success. Several of the datamining algorithms namely K-nearest neighbor, Support Vector Machine, Random Forest, CART and Linear Discriminant Analysis were used on the preprocessed data sets it was observed that Random Forest performed the best with an accuracy of 93.17% for prediction.

In [17] the neural networks have been used by Sharda & Delen (2006) for predicting the success of movie at box-office before it is released. Using this model we have classified in to nine categories ranging from flop to a blockbuster. Because the model predicts the expected revenue range so it can be used for decision making by distributors, studios and exhibitors. Two performance metrics have been used. Within one class of its actual performance and average percent success rate of classifying exactly. The neural network model proposed outperforms the other models in the literature.

In [18] Quader et .al (2017) used several machine learning methods on movie dataset for classification. Here performance comparisons among different machine learning algorithms has been done. Algorithms namely Multilayer Perceptron, Logistic Regression, Support Vector Machine, Naive Bayes, AdaBoost, Stochastic Gradient Descent (SGD) and Random Forest have been used. The databases like Meta Critic, Rotten Tomatoes, Box Office Mojo and IMDb have been used for predicting net profit value approximately. The output is predicted based on some pre-released and post-released features and the dataset with 755 movies was used. The experimental results show that neural network gives best results.

Subramaniyaswamy et .al (2017) in [19] used publicly available data which is relevant and multiple regression algorithm is used for prediction of movie success, resulting R-value more than 0.88.The result of SVM was 56.52% using multiple variables which is higher than SVM using only one variable.

.

III. CONCLUSION

The main aim of this paper is to study all the solutions present in literature for predicting the success of movie. All the papers reported in this study created the models for prediction using different machine learning techniques and using features from movie metadata databases or extracted from other sources such as online social networks. The feature extraction and generation from social networks required lot of efforts but improved classification accuracy than former approach. And we still feel there is lot more provision for data cleaning/extraction from social networks thereby increasing accuracy of classification models. Moreover not all the machine learning classifiers have been used for model creation especially the ensemble classifiers that can improve classification accuracy.

.

REFERENCES

[1] http://www.imdb.com/

[2] Ning, X., Yac, L., Wang, X., Benatallah, B., Dong, M., & Zhang, S. (2018). Rating prediction via generative convolutional neural networks based regression. Pattern Recognition Letters.

[3] Gaikar, D. D., Marakarkandy, B., & Dasgupta, C. (2015). Using Twitter data to predict the performance of Bollywood movies. Industrial Management & Data Systems, 115(9), 1604-1621.

[4] Lee, K., Park, J., Kim, I., & Choi, Y. (2018). Predicting movie success with machine learning techniques: ways to improve accuracy. Information Systems Frontiers, 20(3), 577-588.

[5] Jaiswal, S. R., & Sharma, D. (2017, November). Predicting Success of Bollywood Movies Using Machine Learning Techniques. In Proceedings of the 10th Annual ACM India Compute Conference on ZZZ (pp. 121-124). ACM.

[6] Henry, M., Sharda, R., & Delen, D. (2007). Using Neural Networks to Forecast Box Office Success. AMCIS 2007 Proceedings, 342.

[7] Zhang, W., & Skiena, S. (2009, September). Improving movie gross prediction through news analysis. In 2009

IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (Vol. 1, pp. 301-304). IEEE.

[8] Yoo, S., Kanter, R., & Cummings, D. (2011). Predicting Movie Revenue from IMDb Data. Stanford University.

[9] Oghina, A., Breuss, M., Tsagkias, M., & De Rijke, M. (2012, April). Predicting imdb movie ratings using social media. In European Conference on Information Retrieval (pp. 503-507). Springer, Berlin, Heidelberg.

[10] Jeesha, K., Sumod, S. D., Premkumar, P., & Chowdhury, S. (2018). Does Story Really Matter In The Movie Industry?: Pre-Production Stage Predictive Models.

[11] Latif, M. H., & Afzal, H. (2016). Prediction of movies popularity using machine learning techniques. International Journal of Computer Science and Network Security (IJCSNS), 16(8), 127.

[12] Apte, N., Forssell, M., & Sidhwa, A. (2011). Predicting Movie Revenue. CS229, Stanford University.

[13] Liu, L., & Zhao, Y. (2015). Research of Box-Office Prediction based on Rough Set and Support Vector Machine. International Journal of Hybrid Information Technology, 9(2), 417-426.

[14] Ahmed, M., Jahangir, M., Afzal, H., Majeed, A., & Siddiqi, I. (2015, December). Using Crowd-source based features from social media and Conventional features to predict the movies popularity. In 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity) (pp. 273-278). IEEE.

[15] Parimi, R., & Caragea, D. (2013, July). Pre-release box-office success prediction for motion pictures. In International Workshop on Machine Learning and Data Mining in Pattern Recognition (pp. 571-585). Springer, Berlin, Heidelberg.

[16] Mundra, S., Dhingra, A., Kapur, A., & Joshi, D. (2019). Prediction of a Movie's Success Using Data Mining Techniques. In Information and Communication Technology for Intelligent Systems (pp. 219-227). Springer, Singapore.

[17] Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. Expert Systems with Applications, 30(2), 243-254.

[18] Quader, N., Gani, M. O., & Chaki, D. (2017, December). Performance evaluation of seven machine learning classification techniques for movie box office success prediction. In 2017 3rd International Conference on Electrical Information and Communication Technology (EICT) (pp. 1-6). IEEE.

[19] Subramaniyaswamy, V., Vaibhav, M. V., Prasad, R. V., & Logesh, R. (2017, December). Predicting movie box office success using multiple regression and SVM. In 2017 International Conference on Intelligent Sustainable Systems (ICISS) (pp. 182-186). IEEE.

Mr. S Jahangeer Sidiq received his master's degree and MPhil from University of Kashmir and is working on Class imbalance