International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7, Issue-6, March 2019

Box-Office Revenue Estimation for Telugu Movie Industry using Predictive Analytic Techniques

V. Anantha Natarajan , K SaiHarsha, M Santhosh Kumar

Abstract: ***The film industry is a business cloud for millions in investment and its multiple in revenue. Story rights, production costs, cast remunerations, film promotions etc. charges the production companies. This makes movie analytics inevitably essential for the success of a film and survival of the industry. From the sources like IMDB and Wikipedia; movie related information such as title, budget, synopsis of the story, genre, cast, release date etc. were collected. Analytics were performed on the related data for predicting movie premier collection share, first day share, first week share and overall gross collection to pre-determine the success of the film. Traditional machine learning algorithms and natural language processing techniques were collectively applied to make predictions. These estimations may aid production companies to forecast the make or break chances of the film prior to its release.***

Index Terms: movie analytics, machine learning algorithms, natural language processing, IMDB, and Wikipedia

I. INTRODUCTION

Film industry is a billion-dollar market worldwide. Huge investments from production companies, intense interests of people in watching films and creative works of film makers constantly draws the attention of crowd and heading it in making great profits out of film making. The Indian movie industry is predicted to grow at 11.5% every year, and by the year 2020 it is expected to reach a total gross realisation of Rs. 23.8k crs (roughly around $3.7 billion), based on a report by Bed & Breakfast accommodations aggregator BnBNation [1]. Current gross revenue of it is Rs. 13.8k crs ($2.1 billion). The compound annual growth rate for the last couple of years of the India movie sector in terms of revenue is 10%. The number of films produced in the Indian film industry is the largest in the world. On an average 1.5k – 2.0k films are being produced by the Indian movie industry. But at the same time in terms of revenue the Indian film industry falls behind the film industries of US and Canada where only 0.7k movies are produced yearly on an average.

Revised Manuscript Received on 30 March 2019.
* Correspondence Author
V Anantha Natarajan, Department of CSE, Sree Vidyanikethan Engineering College, Tirupati, India.
K Sai Harsha, Department of CSE, Sree Vidyanikethan Engineering College, Tirupati, India.
M Santhosh Kumar, Department of CSE, Sree Vidyanikethan Engineering College, Tirupati, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

The overall gross revenue of US and Canadian film industry is $11 billion. Despite of producing more number of quality featured films and spending nearly 50% of the production cost for the marketing and promotions, the Indian film industry is experiencing various challenges. This work tries to provide solution to address few of the challenges prevailing in the industry.

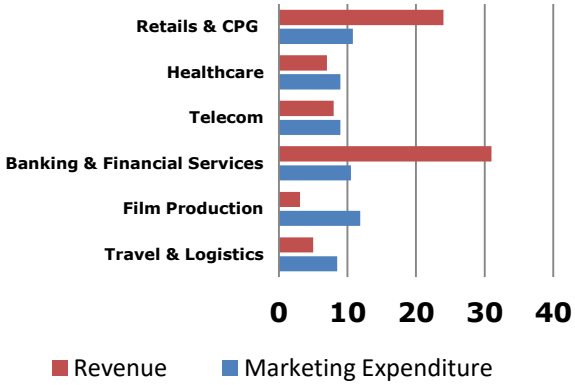


Fig. 1 Comparison of revenue and marketing expenditure of business sectors in Crs.

Nowadays the film production cost includes also the marketing and promotional expenditures which are high as in the case of FMCG and BFS (banking & Financial Services) sector. If business intelligence tools and other sophisticated analytics mechanism were used then the ROMI – Return on Marketing Investment can be increased. At present scenario the hindi film industry is experiencing drop in Box-Office ticket sales even though the average ticket process have went high. In the overall revenue the box-office sales comprises of 74% and the remaining contribution is from mobile & online/ home video rights.

Table 1. Year-wise Film Industry revenue

Revenue (INR Billion)	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019 (predicted)
Domestic Theatrical	62	68.8	85.1	93.4	93.5	99.9	113.6	123.5	133.7	145.1
Overseas Theatrical	6.6	6.9	7.6	8.3	8.6	9.6	10.9	11.9	12.9	13.9
Mobile/ Home video	4.1	4.7	5.4	7	8.4	10.3	12.5	15.4	18.3	21.8
Cable and Satellite Rights	8.3	10.5	12.6	15.2	14.7	15.5	17.6	19.2	20.8	22.5
Ancillary Revenue Streams	2.3	2	1.7	1.4	1.2	1	0.9	0.8	0.7	0.6
Total	83.3	92.9	112.4	125.3	126.4	136.3	155.6	170.7	186.3	204

The role data analytics in movie industry has more benefits and provides educated guesses based on analysing historical data. Analytics will help to make better decision with the available quantitative data and an detailed analysis of the risk involved in the business. AI & Machine learning have entered in to almost all industries including media and entertainment and a notable application of AI & ML is at NetFlix [2]. Beyond predicting the success of the movie the analytics algorithms have the capacity to estimate the overall collection, decide the number of screen to be released, and schedule the movie releases. This paper aims at developing prediction techniques to estimate the box-office collection of movie at different stages. Before initiating the analytics on the movie industry data, the nature and functioning of the industry is reviewed.

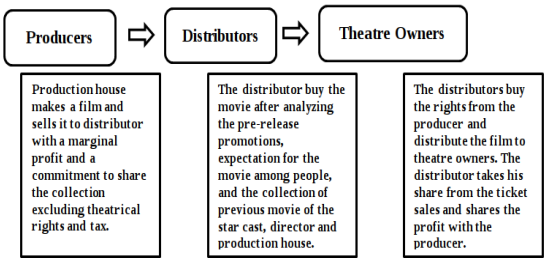


Fig. 3 Process in Film Industry

The Indian movie industry is divided in to 14 circuits based on regional languages and each of the circuit has their own set of distributor and exhibitors' association. The role of each of them is presented in Fig. 3. The Box-office revenue is shared between the producer, distributor, and exhibitors by different revenue sharing scheme. The total cost of a movie is estimated based on the following parameters namely pre-production cost, film production, post production and advertisement expenses. Overall gross collection is the amount obtained from all the theatres in which the movie is release. The amount usually includes tax as the theatre owners charges the audience along with the tax when fixing the price of the ticket. Excluding the tax amount and some portion of collections to the theatre owners, the remaining is termed as a 'share' which is distributed between the producer and distributor depending on their revenue sharing agreements. In real time the predictions are made in different stages of film (one week before release, after first day and first week of release). These analysis will help to decide on the certain factors including number of screens a movie will run on which greatly influence the box-office collections of movie. This work attempts to identify optimal predictive techniques for estimating the parameters related to box-office collection of a movie. The rest of the paper is organized as follows; the section 2 presents a detailed review existing literatures relevant to out chosen problem. The data collection procedure is described in section 3 and in Section 4 the overall methodology of the proposed predictive analytics approach is discussed. The section 5 explains the experiments conducted with collected data and analysis of the performance of various regression techniques in prediction. Finally the section 6 concludes the paper with a conclusion remarks.

II. RELATED WORK

In [3] a novel method was proposed to predict the box-office opening collection of a movie. The opening collection including premiere and first day collection decides the first week collection. Another important which decides the box office performance is the pre-release hype for the movie which is estimated based on the results of sentiment analysis carried in the social media. Mining and analysing opinions expressed by the fans and general audience after the trailer and teaser releases will help to estimate the overall expectations of the movie. Hence in the proposed work the number of views for trailer and teaser in YouTube released by the production company is considered in predicting the box-office revenue. YouTube is most popular video content sharing platform around the globe. For the reason, it became the most preferable choice to the movie producers and studios for attracting viewers through teasers and trailers. Using the insights from YouTube trailers, gross income of movies was predicted [4] by considering opening income, number of views, likes, dislikes and comments. Linear regression model was trained on a dataset of 7988 movie trailers for predicting the movie gross income using the features. This was further extended in our proposed work by considering other attributes in predicting the overall gross revenue and other related revenue elements. By utilizing the publicly available information at the time of announcement, the probability of take over success of a film is predicted [5]. Takeover success prediction was considered as a binary classification problem. By evaluating and analysing many state-of-the-art classifiers, including logistic regression, artificial neural network, support vector machines with different kernels, decision trees, random forest, and Adaboost and validating their effectiveness in takeover success prediction. Best results were obtained from support vector machine with linear kernel and Adaboost with stump weak classifier. The results seemed to be consistent with the general observations. In [6] the suitability of the following machine learning algorithms Multinomial Logistic Regression, Naïve Bayes, and SVM for movie reviews analysis was studied. In another research work data mining tools were used in extraction of interesting patterns to predict the box office success of movie [7]. The data collected from the various social media platforms including twitter, YouTube were used and the prediction is labelled as Hit, Neutral and Flop. Seven machine learning algorithms were analysed to assess their performance on the data set containing information of 755 movies release between year 2015 and 2015 [8]. The machine learning methods analysed include logistic regression, Support vector machine, Random forest, Gaussian naïve bayes, ada boost, stochastic gradient descent, and multi-layer perceptron. In [9] the authors have attempted to increase the accuracy of the prediction model by constructing new features that influences the increase in the accuracy of prediction. They added a unique feature extracted from the story plot of the movie. The authors of the proposed work in [10] have used multinomial regression techniques for predicting the box office collections. The input parameters used in predictions are budget and runtime of the movie. In another research work the authors have used the social media messages such as tweets and Facebook posts to predict on the rating of the movie [11].

With a view of transforming the unstructured customer posts in to actionable intelligence and using the intelligence for analysing trend and predictive analytics the authors of the work presented in [12] have built an interactive environment.

III. DATA COLLECTION

As this paper is more related to the Indian films, datasets available in the internet are not suitable for analysis. Public datasets available in the internet are mostly limited to Hollywood movies. Analyzing regional movies (primarily, Tollywood) and predicting the Overall Gross for a film was the main aim. Due to unavailability of the any regional movie data in public datasets, preparation of an own dataset with 11 attributes (7 features and 4 targets which were considered necessary listed in Table 2.) became imperative. Therefore, data extraction was done majorly from IMDb, Wikipedia and YouTube using Python and a few libraries. Some information like USA-Premiere collections, Day-1 share and Week-1(considered attributes) share are collected from Tollywood data hosting websites. Following are the attributes considered necessary for analysis.

Table 2. Input Features and Targets

Features	Targets
Budget	USA-Premiere
Star-value*	Day-1 collections
Pre-release business	Week-1 collections
Festival-time (0|1)	Overall Gross
Trailer views	
Audio views	
Plot	

Before selecting the features for analysis, a rough dataset with Tollywood movies was prepared by considering many of the most related attributes for a movie like genre, cast, budget, social media popularity etc. A nominal analysis was performed on this data to find the gross associated features that were considered necessary in predicting the Overall Gross for a film. Those observations quite helped in selecting the features that are mentioned above. It was observed that movie genre has a minimum or no impact on the Overall Gross in Tollywood films, as most of the Tollywood films were mixed genre (includes family drama, comedy, love and romance) and only a very few movies diverge from this trend. Also, the primary actors (Hero or Heroine) in the film and the director of the film are showing some significant impact both in Success and Overall Gross of the film. Therefore, the attribute 'star-value' was introduced to indicate the effect of actors and directors of the film. Subsequently, assuming the impact of festival season on the Overall Gross for a film, the attribute 'festival-time' was added to the dataset. Finally, a dataset was prepared with the above mentioned seven features for predicting Overall Gross and rest of the targets for a movie. The star-value of a movie is calculated based on the mathematical equation given below in Eq. 1. This attribute represents the contributions of hero and the director of the movie towards the success of the film. A high profile star can influence the development of the movie by fetching support from the investors, distributors, and exhibitors. But at times the return on investment in the movies acted by high profile stars is highly debated. From the scrapped data it is observed that on an average only 60% of the high grossing films featured a highly paid star cast. Star value has weak correlation with the box office collection and it is analyzed detailed in Section 4.

*Star-value – sum ((0.7 * actor's past 3 films average gross) + (0.3 * director's past 2 films average gross))

IV. METHODOLOGY

The problem of movie business revenue estimation is considered as supervised learning problem wherein the output variable 'y' (Gross / USA premier collections/ Day 1/ Week 1 shares) has to be predicted from the given input variables 'x_1, x_2, ..., x_n'. From the training set $\{(x_1, y_1),\ldots,(x_n,y_n)\}$ the aim is to find and approximation of function $F(x)$ such that the error estimated using a loss function $L(y, F(x))$ is minimized.

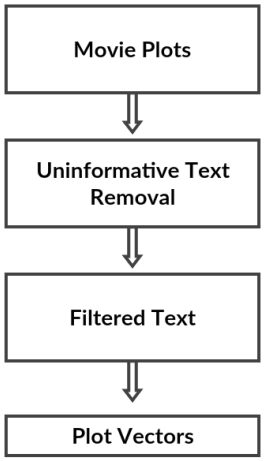


Fig. 4 Text Vectorization and transformation

Before analysis initially the collected data must be converted in to right format suitable for analysis which involves data cleaning, converting text data in to numerical vectors, encoding categorical data numerical data and normalize all the attribute values. The collected raw data was inefficient to perform analysis. Therefore, two pre-processing steps Plot2Vec conversion and Feature Normalization were performed on the raw data before supplying it to the model to yield better results and flow is presented in Fig. 4.

Plot2Vec conversion:

After collecting the necessary amount of data, records containing more null values were deleted. From this dataset, movie plots (collected from Wikipedia) were converted into vector form and supplied as an input to the model. Usually, these plots are composed of many uninformative words such as 'a', 'is', 'the' etc. To eliminate such kind, they were pre-processed before supplying them to the model. After pre-processing the text, informative words representing the main idea in the plot were extracted (also represented as keywords in the plot) and converted into the vector form using Word2Vec model. These vector inputs were supplied along with the six features to the model. Text in every movie plot is a composition of both informative and uninformative chunks of data. Uninformative chunks of data include stop-words ('the', 'is', 'are', etc.), digits, punctuations, named-entities etc.

They are used to add order, meaning and to serve grammatical purposes while reading and understanding the text. But they don't necessarily present the main idea in the text. Informative chunks were more descriptive and present the main idea in the text. To extract informative words, these uninformative chunks need to be eliminated. Punctuations were replaced with empty spaces and stop-words were removed using stop-word removal in NLP. The acquired text was divided into tokens and processed individually. Using pos_tagging in NLP, named-entities were identified from the tokens and removed. On the rest of the tokens, NLP techniques like lemmatization and stemming were applied to convert each token (word) into its respective base form. At the end, the processed tokens were joined to form a filtered text filled with informative words. This method was applied on every plot to filter the informative text before converting to the vector form.

Feature Normalization:

When the data collection process was completed, there existed features with values varying at different range. To reduce the knock-on effect on the model's learning ability; features like budget, pre-release business, trailer view count and audio views were standardized to fixed units. Before normalization the values are standardised to fixed unit values like the movie budget mentioned in lakhs, millions and crores were converted to crores. Similarly, the values of pre-release business were converted to crores (fixed unit) followed by the conversion of trailer and audio views in thousands, lakhs and millions to millions (fixed unit). These standardised features are rescaled in to 0 and 1 range and given as input to the predictive models. The predicted output values were inverted by de-normalizing. Eventually, by the end of feature normalization, values in features were aligned in a specified range to reinforce the algorithm to perform better.

Model construction and evaluation:

Before constructing the regression model, using data visualization techniques few observations were inferred from the extracted dataset. It was observed that data is exhibiting a non-linear relationship among its features, from which it was clear that a linear regression model cannot fit the data. So, three non-linear models were chosen to analyse the dataset. The following regression models namely Support Vector Machine (SVM), k-Nearest Neighbour (kNN), and Random Forest (RF) regression were trained on the normalized data. The Fig. 5 illustrates the flow of inputs to predict the targets from the selected features. The non-linear models generally use an iterative algorithm rather than using a linear methodology to predict the dependent variable.

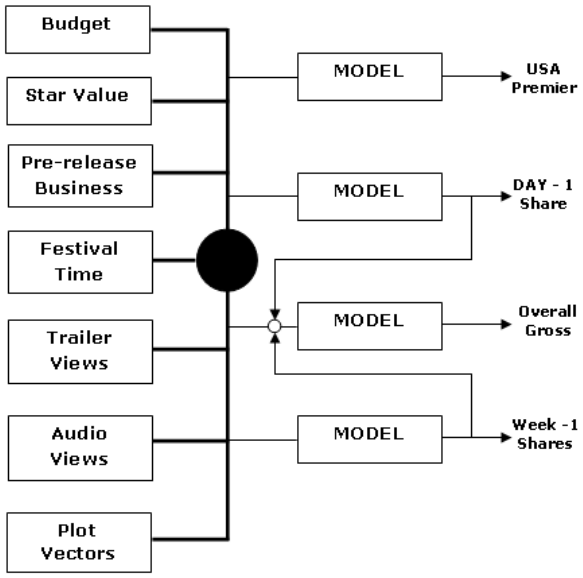


Fig. 5 Information Flow in the predictive modelling process

When compared to other model the training time required for SVM model is higher, due to the presence of higher dimensional plot vectors in the input. k- Nearest Neighbors, as experimented 'k-NN' has adjustable 'k' value which produce different clusters for varied 'k' values, which is not appropriate for a scalable dataset. Compared to the three models, performance of Random Forest Regressor was satisfactory for providing the expected results. The RF algorithm performs better even for high dimensional data, and have ability to handle imbalance and missing values in the data. The performance of the predictive models is evaluated using the Mean Absolute Error and it is calculated as mentioned in Eq. 2.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|\left(y_j - \hat{y}_j\right)\right|$$

Eq.2

V. EXPERIMENTS AND RESULTS

The focus of this work is to predict the box office revenue and USA premier collection of a movie from the historical data available for the recent past released movies. The simple and most widely applied parametric technique for modelling a quantitative continuous dependent variable for a given set of multivariate independent variables is polynomial regression. In this method the relationship between the independent and dependent variable is modelled as n^{th} degree polynomial in x. An alternative method is to use Support Vector Regression (SVR) with a polynomial kernel. The kernel function maps the data in to higher dimensional space where linear separation is possible using a hyper plane. Decision trees with ensemble learning are used in Random Forest (RF) regression where each tree grows using a bootstrap sample of training data. Randomly selected subsets of independent variables are used as candidates for splitting the tree nodes.

The RF regression prediction is estimated by averaging the output of individual trees. Another ensemble learning method based regression technique is Gradient Boosting Regression (GBR) which is capable of selecting optimal set of variables, handle outliers, and missing values. As similar to RF regression the Gradient Boosting constructs the variable importance ranking. K- Nearest Neighbor prediction is based on the mean or the median of the K-most similar instances. The afore-mentioned regression models are constructed using the historical data and the models are tested on unseen data. The dataset with 820 observations has been randomly divided into three buckets: training (60%), validation (20%) and test (20%). The training set has been used to build the models, whereas the validation set data, which would not be seen by the model during training, has been used to compare the out-of-sample errors among the developed models. Then the best performing model is used to predict on the test dataset.

Predictions of SVM Regressor were deviating from the actual targets, while the predictions of KNN Regressor were relatively closer to the actual targets. By comparing the predictions of the two models, results of KNN Regressor seems good but not satisfactory. RF regressor was applied on the dataset expecting some better results. As expected, the predictions of Random Forest Regressor seem to be reliable in certain cases as they were much closer to the expected output. Before constructing the prediction models few visualization techniques have been employed to study the nature of the training data. On the training dataset some of the techniques like correlation analysis, partial dependency finding was performed. The obtained results are visualized using data visualization techniques. Strength of relationship between the attributes, demonstrates the effect of features on the targets. By performing correlation analysis, the degree of association between the variables is analysed. The value of correlation coefficient indicates the strength of the relationship between the attributes. On the normalized data correlation analysis was performed using Pearson correlation coefficient and the observations obtained can be visualized in the Fig.6.

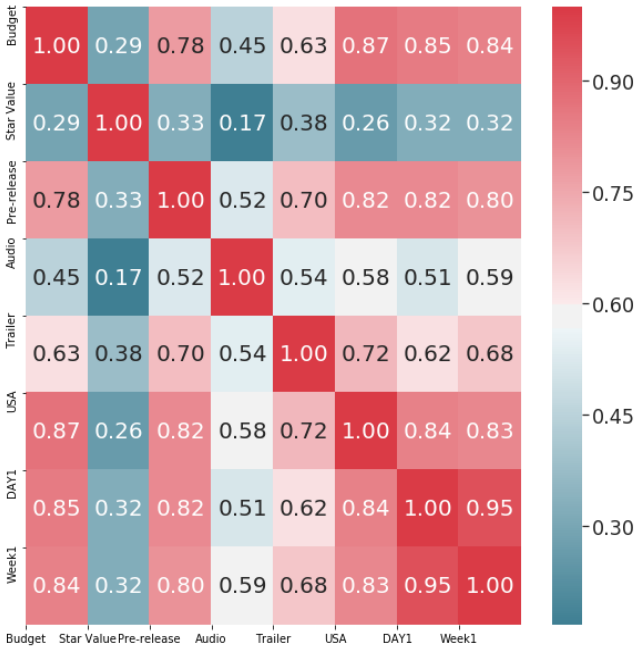


Fig. 6 Heat map of Correlation Analysis

From the above heatmap, it was observed that USA-Premiere collections have strong correlation with budget followed by Day-1, Week-1 and Pre-release business. Similarly, Day-1 collections have strong correlation with Week-1 followed by budget, USA-Premiere and pre-release business. Week-1 collections have strong correlation with Day-1 followed by budget, USA-Premiere collections and pre-release business. Random Forest Regressor can perform well on the normalized dataset. Using the partial dependency plots, each variable or predictor affecting the model's predictions is visualized. The following visualization plots show in Fig. 7 exhibits the effect of strong correlated features with the Overall gross. Observing the figure, it was clear that, with increase in any of the features namely budget, pre-release business, trailer, Day-1 and Week-1; the Overall Gross of a film increase which was shown through the visual results from the experiments. The error in prediction with any regression technique is due to either bias or variance during the learning process. The bias error is caused due to assumptions in the learning algorithm and the variance error is influenced by the fluctuations in the training data set. High variance will make the model to over fit the random noise present in the training data set. The GBR is based on weak learners which have high bias and low variance. From the results plotted in the Fig. 8 – 11, it can be observed that the boosting algorithm reduces the overall error in prediction by reducing the bias and the variance up to a possible extent by aggregating the output estimated by various decision trees.

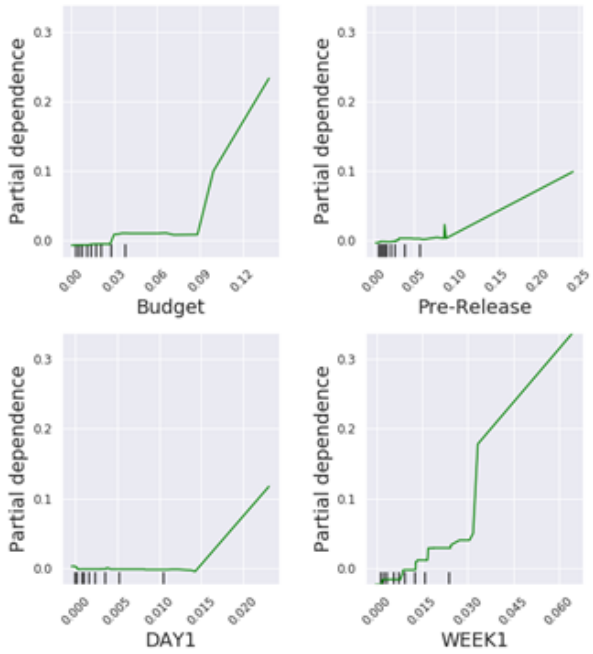


Fig. 7 Partial Dependency Plot

PCA based regression transforms the data such that a large number of features are represented with a limited number of components. The principal components are not intuitive and empirical results indicate that their prediction errors are high when compared to other regression techniques.

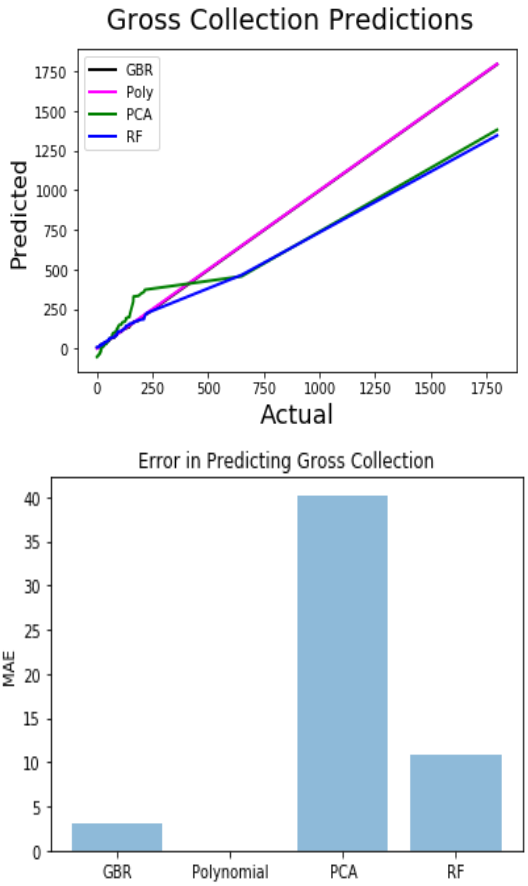


Fig. 8 Plots of gross collection predictions and error

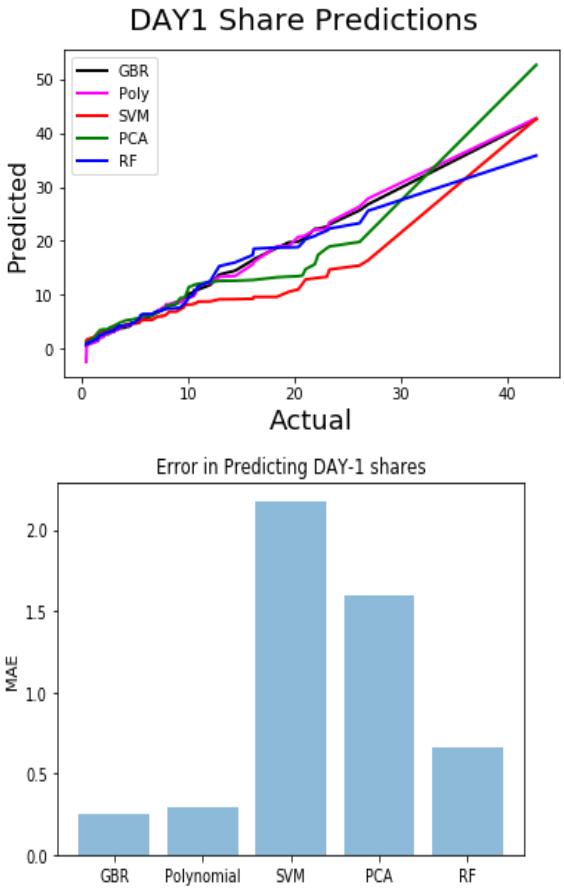


Fig. 9 Plots of Day 1 shares predictions and error

In contrast the RF uses fully grown decision trees which are characterized by low bias and high variance. It reduces the error by reducing variance but it cannot reduce the bias. During the experiments the size of forest is maintained with a large number of unpruned trees to reduce the bias in the initial stages.

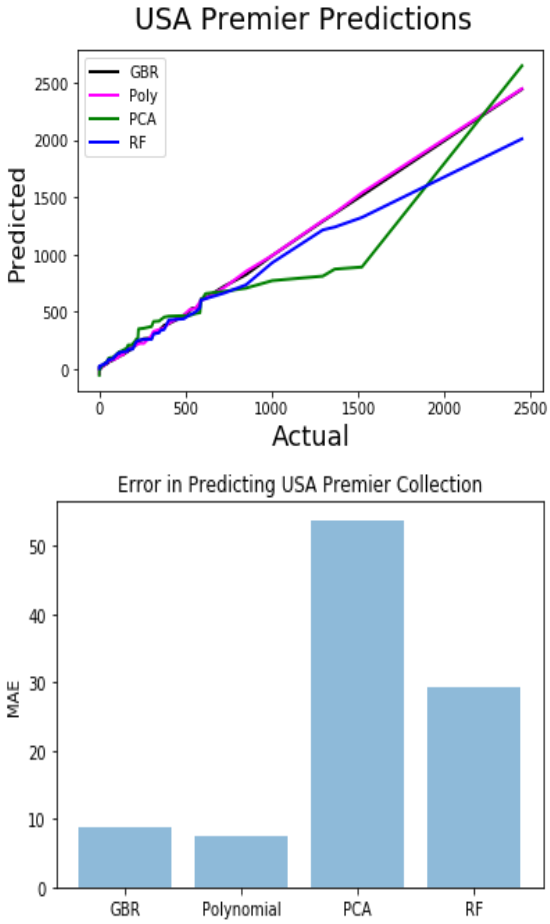
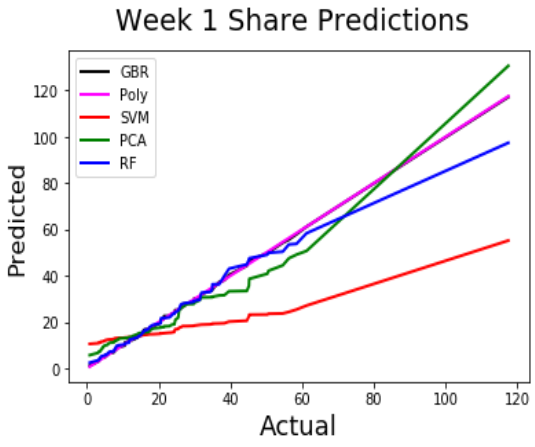


Fig. 10 Plots of USA Premier collections predictions and error



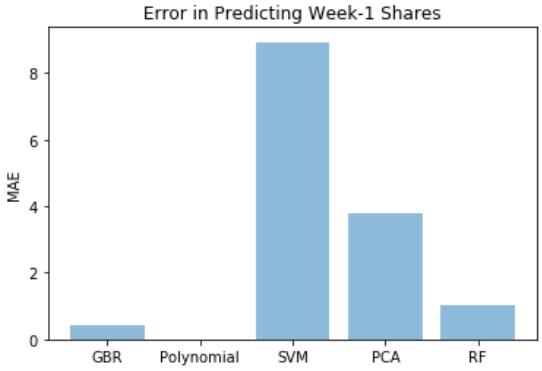


Fig. 11 Plots of Week 1 shares predictions and error

The performance of SVR for predictions of Week 1 share and USA premier collections are poor due to inappropriate values of model hyper parameters. Using grid search technique optimal parameter values have to be estimated after analysing various models with different values of parameters.

VI. CONCLUSION

This paper presented the experiments performed on the regional movie dataset that was prepared as a part of the curiosity towards finding the features which shows an impact on overall gross in Tollywood films. By exploring and performing the experiments on the regional movies data, the features showing the significant impact on the overall gross were acknowledged. Based on the work and the visual representation of the results, it was clear that the three of the features supplied from the six essential input features were showing very significant impact on the overall gross in parallel to the day-1 and week-1 collections. Through a general study and after performing analysis it was observed that story and screenplay contributes much both on the movie success and the overall collections of that movie. This work is encouraging to perform further analysis on the regional data where the availability of regional movies data in the internet is challenging to perform any analysis. This work analysed suitability of various regression techniques in predicting the parameters related to box-office collection of a movie and from the results it was observed that Gradient Boosting and Polynomial Regression techniques were performing better when compared to other regression techniques.

REFERENCES

1. https://economictimes.indiatimes.com/industry/media/entertainment/media/film-industry-in-india-to-hit-3-7-billion-by-2020-says-report/articleshow/60998458.cms
2. George, Gerard, and Yimin Lin. "Analytics, innovation, and organizational adaptation." *Innovation* 19.1,16-22, 2017 .
3. Reddy, Ajay Siva Santosh, Pratik Kasat, and Abhiyash Jain. "Box-office opening prediction of movies based on hype analysis through data mining." *International Journal of Computer Applications*, 56 (1), 1-5, 2012.
4. M. S. Rahim, A. Z. M. E. Chowdhury, M. A. Islam, and M. R. Islam, "Mining trailers data from youtube for predicting gross income of movies," in 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2017.
5. M. Zhang, G. Johnson, and J. Wang, "Predicting Takeover Success Using Machine Learning Techniques", *JBER*, vol. 10, no. 10, pp. 547-552, Sep. 2012.
6. V. Uma Ramya and K. Thirupathi Rao, "Sentiment Analysis of Movie Review using Machine Learning Techniques," International Journal of Engineering & Technology, vol. 7, no. 2.7, p. 676, Mar. 2018.
7. K. R. Apala, M. Jose, S. Motnam, C.-C. Chan, K. J. Liszka, and F. de Gregorio, "Prediction of movies box office performance using social media," in Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13, 2013.
8. N. Quader, M. O. Gani, and D. Chaki, "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), 2017.
9. K. Lee, J. Park, I. Kim, and Y. Choi, "Predicting movie success with machine learning techniques: ways to improve accuracy," Information Systems Frontiers, vol. 20, no. 3, pp. 577–588, Aug. 2016.
10. El Assady, Mennatallah, et al. "Visual analytics for the prediction of movie rating and box office performance." IEEE VAST Challenge USB Proceedings (2013): 3-4.
11. A. Oghina, M. Breuss, M. Tsagkias, and M. de Rijke. Predicting imdb movie ratings using social media. In ECIR, pages 503–507, 2012.
12. Lu, Yafeng, Feng Wang, and Ross Maciejewski. "Business intelligence from social media: A study from the vast box office challenge." IEEE computer graphics and applications 34.5 (2014): 58-69.