Role of Different Factors in Predicting Movie Success

Anand Bhave
Dept. of Computer Engg.
MIT, Pune.

Himanshu Kulkarni
Dept. of Computer Engg,
MIT, Pune.

Vinay Biramane
Dept. of Computer Engg.
MIT, Pune.

PranaliKosamkar
Dept. of Computer Engg.
MIT, Pune.

***Abstract*--Due to rapid digitization and emergence of social media the movie industry is growing by leaps and bounds. The average number of movies produced per year is greater than 1000. So to make the movie profitable, it becomes a matter of concern that the movie succeeds. Given the low success rate, models and mechanisms to predict reliably the ranking and or box office collections of a movie can help de-risk the business significantly and increase average returns. The current predictive models available are based on various factors for assessment of the movie. These include the classical factors such as cast, producer, director etc. or the social factors in form of response of the society on various online platforms. This methodology lacks to harvest the required accuracy level. Our paper suggests that the integration of both the classical and the social factors (anticipation and user feedback) and the study of interrelation among the classical factors will lead to more accuracy.**

Keywords : Box office; Forecasting; Gross income; Machine learning; Movie success;Movie; Predictive analytics.

I. INTRODUCTION

The movie industry worldwide produces a large number of movies per year. However, very few movies are a success and are ranked high. Given the low success rate, models to predict reliably the box office collections of a movie can help by improving the business significantly and increase average returns. Looking at this situation from the movie industry's perspective, if there is a link between critical reviews and getting people out to see a movie, this could help with distribution decision making. If a movie does well in test screenings or if they anticipate good reviews from the critics then they can decide to release it on opening weekend in more theaters in hopes of bringing in more revenue[1].Various stakeholders such as actors, financiers, directors etc. can use these predictions to make more informed decisions. Predictive analytics encompasses a variety of statistical techniques from modeling, machine learning and data mining that analyze current and historical facts to make predictions about future or otherwise unknown events.

In business, predictive models exploit patterns founding historical and transactional data to identify risks and opportunities. Models, actual relationships among many factors, allow assessment of risk or potential associated with particular set of conditions guiding decision making for candidate transactions. The current predictive models available are based on various factors for assessment of the movie such as the classical factors such as cast, producer, director etc. or the social factors in form of response of the society on various online platforms. This methodology lacks to harvest the required accuracy level. Hence a better method is required. Our paper suggests that the integration of both the classical and the social factors to generate the result and the study of interrelation among the classical factors will lead to more accuracy. To achieve this, collecting the data scattered across internet is necessary and thus data on various platforms such as YouTube, Twitter, and Wikipedia etc. is taken into account along with the classical factors resulting in effective integration.

II. LITERATURE SURVEY

Though there are many factors that constitute a movie's success, and it is not always clear how they interact, this paper attempts to determine these factors through the different attributes, social media etc.and predictive analytics.

A. ROLE OF CLASSICAL MOVIE ATTRIBUTES

The classical movie attributes such as cast, director, producer, and genre play a crucial role in the movie's success. Dan Cocuzzo et al have used Naive Bayes and Support vector machine to predict the movie success. In Naive Bayes algorithm, they represented movie as independent combination of associated personas and attributes [2],which was given by,
P(rating | movie) proportional to P(movie | rating) * P(rating) ,where P(movie | rating) is product of individual conditional probabilities for each persona.

Jason van der Merweetal have built Linear Regression and Logistic Regression models [3]. In linear regression least mean square method, specifically stochastic gradient descent, was used to learn the weight vectors. In order to include the movie title in the feature vector, the movie title was given a score. The movie title was included since the title of a movie does have an effect on the movie's success. To accomplish this, K-Means clustering was used. The accuracy was increased to 52% by implementing K-Means clustering on the

978-1-4799-6272-3/15/$31.00(c)2015 IEEE

titles. Nikhil Apte et al have implemented Linear Regression, K-means clustering, Weighted linear regression and Polynomial Regression algorithms [4]. The authors have also considered effect of inflation rate on movie gross. This was done by dividing the global box office collection and the movie budget, by the values of the normalized price of a movie ticket for the year of its release and then multiplying it by the current normalized movie ticket price. Besides traditional movie attributes Jeffrey Simon off et al used additional variables for measuring star power. Linear regression for predicting movie grosses was used [5]. Steven Yoo et al categorized the features into numeric, text and sentiment [6]. The numeric features consist of budget, average rating, duration, user vote count and critics review count. The text based features consist of MPAA rating, director and genre. The sentiment feature consists of the sentiment score. Sharang et al have considered the features that can be used by producers prior to the beginning work on a movie [7]. The variables considered are director, actors and genre. Also the audience rating was used as an extra criterion variable. Four different regression techniques, support vector regression, Ada boosted decision tree regression, gradient boosting regression and random forest regression were used in this project. The best performing regression method was boosted decision trees. Alec Kennedy has studied the interrelationship between the success of the movie and their critical reviews. The author concludes that along with effective marketing strategies and favorably good critical reviews, it is profitable to release a film [1].Jeffrey Ericson et al use only the attributes that are influential in the pre-release phase [8]. They also tried to analyze the impact of the movie title on its success. Despite much effort with various approaches, predicting the financial success of a movie remains a challenging problem. For example , Sharda and Delen have trained a neural network to process pre-release data, such as quality and popularity variables, and classify movies into nine categories according to their anticipated income, from ‘‘flop’’ to ‘‘blockbuster’’ [9]. Seven different types of independent variables were used. A neural network treats these pseudo variables as different mutually exclusive information channels. For test samples, the neural network classifies only 36.9% of the movies correctly, while 75.2% of the movies are at most one category away from correct.

B. Role of user anticipation and response

To get the higher accuracy in the estimation of the box office collection, all possible criterion should be considered. For this, response of users on social media should be taken into account along with the classical factors. Part of the hypothesis of the project is that the anticipation and social media feedback helps to predict movie success as discussed by Gloor et al [10]. They generated the feedback for the movies in three ways: using web searches, using blog searches and using posters on movie forums. In addition to determining the feedback the author performed sentimental analysis on IMDb forums to gather the general mood towards the movie. An important step is to measure the movie title’s relative importance on web and other such forums. The user feedback or movie popularity can be estimated through sentiment analysis of twitter data. Twitter, a micro blogging website plays an important role by conveying information about user feedback and preferences. It can be done by measuring the extent of positive or negative words in tweets. Vasu Jain in his work tries to predict the movie popularity from sentiment analysis of tweets [11]. The data fields for each tweet such as tweet id, user name, tweet text, time of tweet are stored. The author tries to classify the movie into three categories: hit, flop, average. Lyric Doshi, in his project explored the effectiveness of collective intelligence, social network analysis and sentiment analysis in predicting trends by mining publicly available online data sources [12]. To determine general sentiment about movies, the author considered posts from IMDb forums, Oscar Buzz, Film General. The author used single variable and multi variable linear regression models which proved to be effective. Another effective measure is determining the Wikipedia metrics associated with a particular movie. It signifies the user interest or anticipation towards a movie. Marton Mestyan et al in their paper have considered Wikipedia metrics consisting of the following parameters such as V: Number of views of the article page, U: Number of users, being the number of human editors who have contributed to the article, E: Number of edits made by human editors on the article, and R: Collaborative rigor of the editing train of the article. The authors used multivariate linear regression [13].Alexander Jagar et al developed visual analytics tool based on tweets and IMDb data [14]. The authors in this project displayed the tweets’ content as a graph structure to get feeling for actors, associations and sentiments. The MooVis tool was then used to get an overview about the movie itself. A popular approach to predicting box office success was developed by Google this approach utilizes Google’s vast corpus of search data to predict box office performance using query volume [15].Movie success prediction through YouTube metrics is another important way. The metrics such as view count or likes, a particular movie trailer gets can be influential for predicting the box office performance. Eldar’s Sadikov et al implemented a model for analysis of comprehensive set of features extracted from blogs for prediction of movie sales [16]. The authors used the blog data set from spin3r.com for

comprehensive list of features that deal with movie references in blogs.

III. PROPOSED METHODOLOGY

A. Data Acquisition:

The data that is required for predicting the movie success is scattered across the internet. Various sources can be useful to gather the relevant data for the estimation of movie success. The data includes the classical factors that are considered for the analysis of the movie, which are be obtained from IMDb and similar websites and at the same time the data that is generated due to the social media which can lead to the conclusion regarding the popularity of the movie. These social factors include various responses on the social media such as sentiment analysis of the tweets made on Twitter, YouTube view hits on the pre movie videos such as trailers and the increment rate of the views as the movie release date approaches and the edits made by the users on the Wikipedia page of the movie or movie related articles. Since our method involves the integration of the classical factors as well as the social factors, the first step involves the acquisition of this data through various sources available. These contain the various files consisting of data about classical factors as well as social media analysis. Hence here we gather the data through various APIs and file formats available.

B. Data preprocessing and formation of usable dataset:

The data acquired needs to be stored systematically in the database so that it can be used as the training or the testing dataset. This data base acquired initially can be considered as the raw data which is not directly applicable as it may have many redundancies, incomplete data and other inconsistencies. Some of the data might be in the form of lists while other can be present as the API network calls. Hence converting all this data to one single usable format is necessary. Data preprocessing involves the conversion of the raw data acquired previously to the usable data. Initially this involves the conversion of all the data in one single uniform format such as SQL database. In the later step among all the data acquired only the relevant data is to be stored in the database. This involves removal of all the redundant data such as removal of the entry of the movie tuple which has some of its classical factors missing or removal of the data that is out of the scope such as movies released before 1990 or those movies which are not released under Hollywood or Bollywood. Once only the relevant data is stored in the database we convert it to the directly applicable dataset by normalizing the database. Here we introduce various linking factors among entries and improve the accessibility of the database.

C. Feature Extraction:

The output for a particular movie success can be determined on two platforms as follows:

A. Pre-Production Phase (Classical factors and Interrelations)

B. Post-Production Phase (Classical factors, their interrelations and Social media)

A user may want to test the results of a movie for a movie that has already been declared; is fully or partially produced and whose data is available on the social media. In this situation the user may provide only the title of the movie for result estimation. On the other hand a user may be a movie stakeholder who would want to make most efficient choice for his crew so as to maximize the grossing of his movie and get maximum returns. In this case he can try various combinations of the input parameters and predict the result. Thus extraction of proper features from the user which would give the desired result is necessary.

D. Classification Model:

The analysis of features involves setting up the relationships among the various features or the parameters which is the important advancement in our result generation mechanism. In the movie industry there are certain relationships established among stakeholders over time. The previously established relationships will affect the weights for various parameters that would minimize the error rate. The training dataset is applied to the machine learning algorithm and weights of the various parameters are modified such that the error rate of hypothesis function is minimized and the maximum accuracy is achieved.

E. Result Generation and Accuracy Estimation:

The test dataset is input to the classification model (hypothesis function). Based on the accuracy achieved the parameter weights can be adjusted or modified. The movie success is measured in two parameters:

- Gross box office collection.
- Critics rating.

Since these two parameters judge the success of movie on different independent and unrelated levels the estimate for the success can be fairly made. Once the results are generated, then the movie tuple for the movie under analysis, along with the results can be stored in the training dataset for the purpose of machine learning and can be further used for accuracy estimation through the testing dataset.

IV. Discussion and Analysis

The current predictive models are usually based on classical factors such as Cast, Producers, Directors, Genre, movie revenue, movie production budget.

Scope for improvement as a basis of hypothesis:

A. The interrelation among the classical factors

Every model we have seen previously considers all these parameters or subset of these parameters. The interrelation among the classical factors is generally not considered. To increase accuracy we are taking into account the interrelation between these classical factors. If particular actor or actress works with particular production house, their films perform well on box office. For example movies of Ajay Devgan with RohitShetty as director are generally blockbuster hits.

B. Integration of classical factors and social media interactions for improving overall accuracy rate

Along with the classical factors or the main movie attributes, considering the user anticipation or feedback improves the prediction success rate.

a. YouTube

Prior to movie release, movie teaser or trailer is available on YouTube. So YouTube hits or views of a movie trailer provide a opportunity to predict the popularity of movie and hence the success. YouTube provides API for accessing data related to particular video.

b. Twitter

Sentiment analysis of tweets on twitter can contribute to improve accuracy of model. Through requisite API the sentiment analysis of tweets gives insight about user feedback or response for a particular movie.

c. Wikipedia

From Wikipedia, the view and edit counts of a particular movie can be obtained which provides information about popularity of a movie.

V. Conclusion:

Thus, the overall success of an unreleased film can be accurately predicted by considering the classical features as well as the user anticipation or feedback through social media channels. The exploratory analysis of features resulted in understanding the inter-relationships between them. For example, considering features like IMDb rating, YouTube view count and number of theatres a movie is released in and using multi-variate linear regression, the multiple R-squared value obtained was 0.7057. Our study suggests that if more data is taken into account and properly integrated, then greater accuracy can be achieved than considering the classical or social factors individually. Hence integration of the classical and social factors and using the basis of interrelation among the classical factors to assign the weights will provide us with the higher accuracy.

REFERENCES

[1] Alec Kennedy; “Predicting box office success: Do critical reviews really matter?”; UC,Berkeley

[2] Dan Cocuzzo, Stephen Wu ; “Hit or Flop: Box Office Prediction for Feature Films”; Stanford University , 2013.

[3] Jason van der Merwe, Bridge Eimon; “Predicting Movie Box Office Gross”, Stanford University, 2013.

[4] Nikhil Apte, Mats Forssell, Anahita Sidhwa; “Predicting Movie Revenue” ; Stanford University.

[5] Jeffrey S. Simonoff, Ilana R. Sparrow; “Predicting movie grosses: Winners and losers, blockbusters and sleepers”.

[6] Steven Yoo, Robert Kanter, David Cummings; “Predicting Movie Revenue from IMDb Data”; Stanford University, 2011.

[7] Sharang Phadke, Shivam Mevawala; “BoxOffice: Machine Learning Methods for predicting Audience Film Ratings”; The Cooper Union for Advancement of Science and Art.

[8] Jeffrey Ericson, Jesse Grodman; “A Predictor for Movie Success” ;Stanford University, 2013.

[9] Ramesh Sharda, Dursun Delen; “Predicting box office success of motion pictures with neural networks”; Elsevier (2005).

[10] P. Gloor, J. Krauss, S. Nann, K. Fischbach and D. Schoder; “Web science 2.0: Identifying trends through semantic social network analysis.”;In IEEE Conference on Social Computing, Vancouver, August 2009.

[11] Vasu Jain ; “Prediction of movie success using sentiment analysis of tweets”; SCSE 2013.

[12] Lyric Doshi; “Using Sentiment and Social Network Analysis to Predict Opening-Movie Box-Office Success.”; Massachusetts Institute of Technology, 2010.

[13] Marton Mestyan, Taha Yasseri, Janos Kertesz; “Early Prediction of Movie Box Office Success Based onWikipedia Activity Big Data”,2013.

[14] Alexander Jagar, Daniel Hafnar ; “MooVis – A visual analytics tool for prediction of movie viewer ratings and box office.”; University of Konstanz, Germany.

[15] P.Reggie, C. Andrea ; “Quantifying movie magic with google search”; Google white paper.

[16] Eldar Sadikov, Aditya Parameswaram, Petros Venetis ; “Blogs as predictors of movie success.”; Stanford University.