2016 International Conference on Internet of Things and Applications (IOTA)
Maharashtra Institute of Technology, Pune, India 22 Jan - 24 Jan, 2016

Relationships between Classical Factors, Social Factors and Box Office Collections

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

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

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

Pranali Kosamkar
Dept. of Computer Engg.
MIT, Pune.

***Abstract*- Every year the number of movie produced and released surpass the previous year's count and so do the total box office collections. So in this quality centric industry, it becomes imperative that the movie succeeds both in terms of box office collections and critical reviews and also renders profit. Due to advent of predictive analytics and big data generated through various social interactions, models to predict accurately the total gross of a movie can be devised, which eventually help the movie studio by giving constructive feedback both in pre-production and post-production phase. So the availability of this data gathered from various social platforms like IMDb, YouTube and Wikipedia can help to gauge the society's reaction and response towards a particular movie. It can also foretell a society's anticipation towards a particular movie. In this paper, we have built predictive models by establishing links between classical features, social media features and the overall success of the movie which includes total box office collection and the critics rating or review. The results show that the prediction model built using integration of classical as well as social factors can achieve higher accuracy rate.**

Keywords : Box office gross; Data Mining; Machine learning; Movie success; Movie, Predictive analytics; Critical review; Rating.

I. INTRODUCTION

Like many innovations, the movie industry has been driven by advances in technology and is mainly dependent on consumer approval and response. Despite a shortening release window and the emergence of digital over-the-top home video alternatives, box office spending worldwide is flourishing, rising a cumulative 32.2 percent between 2008 and 2013. The box office collections are expected to grow, increasing by a cumulative projected 28.1 percent over the next five years due to the digitization of cinemas and growth in the number of screens in some countries [15]. Despite this promising economic scenario, its rapid emergence and growth and its pronounced effect on the everyday life of consumers, the factors for accurately determining the success of a feature film are difficult to establish given the various complexities involved in film-making. For example, the film 'Yuvraj' had great star cast but not a great script, so it was a flop at the box office. But even though 'Shawshank Redemption' had a great star cast and a great script, it was still a moderate hit at the box office at the time of its release. Every year the film industry worldwide produces a large number of movies across various genres. There are many opportunities for movie studios to capture various patterns and to build predictive models, thereby improving its business significantly. Looking at this situation from the movie industry's perspective, it can test the anticipation and approval of the audience through various social media platforms and can make better informed decisions.

In business, predictive analytics models generate interesting patterns from historical and current data to identify various strengths, risks and opportunities to make prediction about future events. The current predictive models available are based on various factors involved directly or indirectly in film-making such as the classical factors like cast, producer, director etc. or the social factors in form of response of the audience on various social media platforms. This strategy lacks to attain the required accuracy level as entire set of features needed for prediction are not considered. In this paper, we have built predictive models by establishing links between classical features, social media features and the overall success of the movie which includes total box office collection and the critics rating or review. For generating a wholesome dataset, we have collected data scattered across internet through various techniques like screen scraping and API calls. Thus data on various social sites such as Box Office Mojo, IMDb, YouTube, Rotten Tomatoes and Wikipedia is taken into account along with the classical factors.

II. LITERATURE SURVEY

The literature on predicting overall success of a new feature film can be classified in two parts based on the set of features involved while predicting, it can be based on role of classical factors or on role of various user responses through social media factors.

A. ROLE OF CLASSICAL FACTORS

The classical factors such as director, producer, cast, runtime and genre play a crucial role in the movie's success. Dan Cocuzzo et al have used naive bayes in which they represented movie as independent combination of associated personas and attributes [2], to predict the movie success. Jason van der Merwe et al have built predictive model using linear and logistic regression. They also have classified movie titles using k-means clustering. [3]. Nikhil Apte et al have implemented Linear Regression, K-means clustering, Weighted linear regression and Polynomial Regression algorithms in which they have considered the effect of inflation rate on movie gross [4]. This was achieved by

978-1-5090-0044-9/16/$31.00 ©2016 IEEE

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. Then the value was multiplied by the current normalized movie ticket price. Steven Yoo et al have categorized the features used for prediction into numeric, text and sentiment factors [5]. The sentiment feature consists of the sentiment score. 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. Sharang et al have considered the features that can be used by producers prior to the beginning work on a movie and thus is mainly suitable for pre-production phase of a movie [6]. The attributes considered are director, actors and genre and rating. Their best performing regression method was boosted decision trees. For increasing accuracy, rather than including classical factors as discrete components, inter-relationships between different classical factors can also be considered [14]. Alec Kennedy has studied the close relationship between the success of the movie and the critical reviews the movie garners. The author in conclusion states that with generally good critical reviews, it is profitable to release a film [1]. Jeffrey Ericson et al also use only the factors that are influential in the pre-release phase [7]. The authors have also tried to understand the impact of the movie title on its overall success. Despite much analysis of various classical factors, predicting the financial success of a movie solely on the basis of classical factors remains a challenging problem. For example, Sharda and Delen have built a neural network and classify movies into one of the nine categories based on their anticipated income [8]. 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

From our understanding of the literature above, to get a substantial accuracy in prediction of the box office collection, all possible features should be considered. For this, response and approval of users on social media should be taken into account along with the classical factors. Our project is based on establishing relationships between classical factors and social features. Gloor et al [9] generated the feedback for the movies in three ways: using posters on movie forums, using web searches and using blog searches. For supplementing their feedback, the authors performed sentimental analysis on IMDb forums to gather the net sentiment towards the movie. The response for a movie or its popularity can be estimated through sentiment analysis of twitter data. Vasu Jain in his paper tries to predict the popularity from sentiment analysis of tweets [10]. The authors generate dataset for each tweet using parameters such as tweet id, user name, tweet text, time of tweet. Then the authors try to classify the movie into three categories: hit, flop, average. Another effective measure is analyzing the Wikipedia usage meters related with a particular movie. It signifies the user anticipation towards a movie. Marton Mestyan et al have considered Wikipedia metrics such as number of views for the movie article, number of users who have contributed to the article, number of edits made to an article, and the collaborative rigor. While building the predictive model, the authors have used multivariate linear regression [11]. An approach to predicting box office success was developed at Google which uses the vast corpus of search data stored to predict the success of a movie [12]. Box Office prediction through YouTube metrics like the views a movie trailer gets, the corresponding likes and dislikes is another important way. Eldar Sadikov et al have built a model for analysis of features extracted from different blogs for prediction of movie sales [13]. While generating the dataset, the authors have used data from spin3r.com for list of features that deal with movie references in blogs.

III. PROCESS WORKFLOW

A. Data Acquisition:

For any accurate movie prediction there is huge requirement of historical data. This data may not be available at single source. It may be spread over many websites. So to amass this data from various sources we used following techniques:

- Web Scraping
- API Calls

The data gathering phase is completely implemented in python language because it has inbuilt packages such as BeautifulSoup, HTML Parser and Wikipedia which made it easy for us to collect data from various websites. The data gathering can be classified into following two parts based on the nature of the attributes:

a. Classical attributes

The classical factors are generally used for quality analysis of movie. So first of all we gathered movie titles with their worldwide grosses from Box Office Mojo website by web scrapping. These movie titles are used to query websites such as IMDb and Rotten Tomatoes to obtain various classical attributes such Cast, Director, Movie studio, Genre, IMDb Rating, Runtime, MPAA rating using API calls.

b. Social Signals

Various social media responses can be gathered through websites such as Wikipedia, YouTube and Twitter. The YouTube hits, likes and dislikes are gathered from their API which is two steps process. First we get video-id of movie trailer through API and then we use this video-id to get the information about movie trailer. Wikipedia movie article views and edit count is obtained through web scrapping of Wikipediaviews.org website and Wikipedia movie article page itself.

B. Data preprocessing:

The data gathered from various websites and APIs may be in different formats and thus it needs to be stored in one uniform format (.csv file). The ambiguity caused by movies with same name but released in two different years is handled by post fixing year to name of respective movie. For example movie Gladiator was released in 1992 and also in 2000. Our database has movies from 2000 to 2015 released under Hollywood or Bollywood. Before preprocessing our database contained more than 8000 movies with their attributes and after pre-processing, the database contains about 5000 movie titles with more than 40 classical as well as social attributes with their groupings.

C. Feature Extraction:

In Feature Extraction phase we discovered various movie attributes and their interrelation which plays crucial role in movie success. For that we plotted many graphs on different attributes using our database. Some of them are shown below

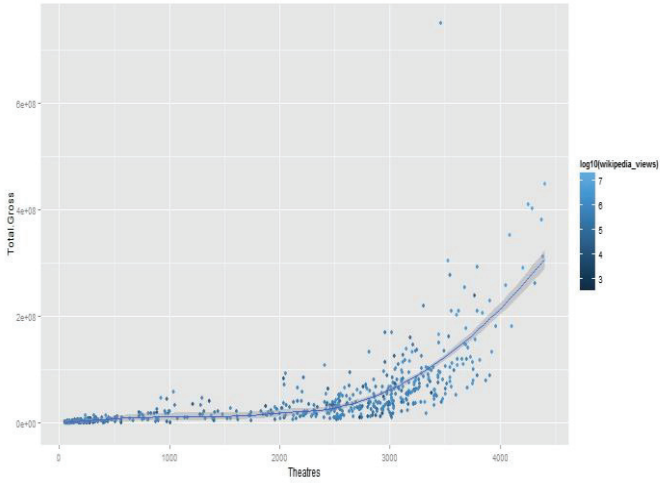

Fig. 1 Theatre vs Total Gross (w.r.t Wikipedia Views)

The Figure 1 represents relationship of Total gross to Theatres with respect to Wikipedia views. As you can see, Total gross is directly proportional to number theatres the movie is released in. Wikipedia views are represented by shade of color i.e. lighter the color larger are the Wikipedia views and vice versa. So as the Total gross increases the Wikipedia views are also increasing. So the conclusion can be drawn that if Wikipedia views of movie article are large then movie can achieve high gross.

The figure 2 represents Actor – Director relationships and its effect on total income or gross of movie. For example movie of Martin Scorsese and Leonardo DiCaprio are more successful than their other movies. Other Interrelationships such as Actor- Genre, Director – Genre and Production house – Genre are also considered.

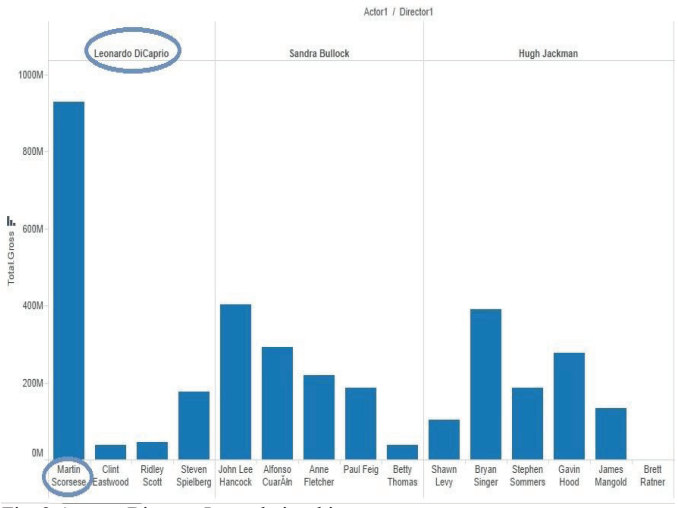

Fig. 2 Actor – Director Interrelationship

D. Analysis of Features:

Data acquired through different sources contain various attributes required for prediction. We used python package called Network-X. This package makes use of Page Rank algorithm to assign weight for each attribute. Using this package we constructed a graph. Nodes of graph correspond to actor or director or genre and movie name, edge between movie title node and any other node denotes that movie has particular actor or director or genre. The total income or gross of movie is assigned as weight of edge. Now running Page Rank algorithm on this graph we get numerical values which represent rank of particular movie attribute such as actor, director, genre etc.

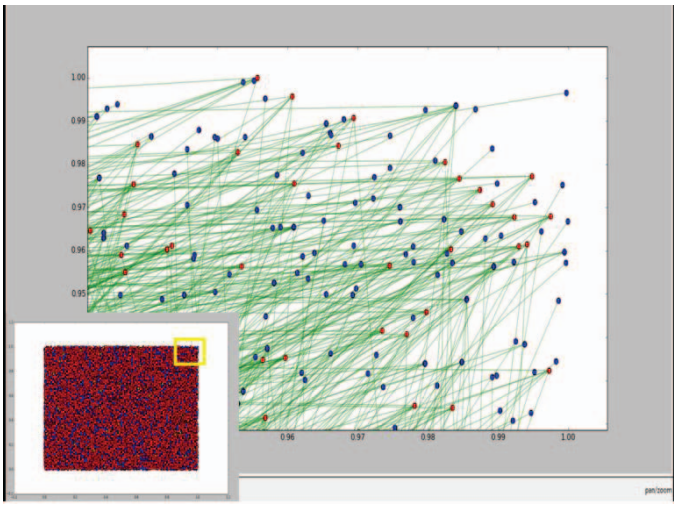

Fig. 3 Network - X Graph

.

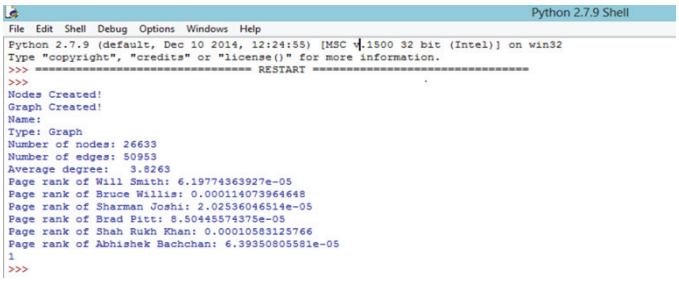

Fig. 4 Summary of graph with some actors' numerical weights or ranks

E. Predictive Model & Accuracy Estimation:

The predictive model is built using the machine learning technique, multivariate linear regression. The general form of multivariate linear regression is

$$y = \sum_{i \in V} \theta_i * X_i + \theta_0$$

Where θ_0 is a constant and θ_i are weights and X_i are the variables. The dataset is given for training the model for maximizing the accuracy. The accuracy of linear regression is measured in terms of R^2 (R-squared). The equation for R^2 is given by: 1 – (SSE / SST), where SSE is the sum of squared error and SST is total sum of squares.

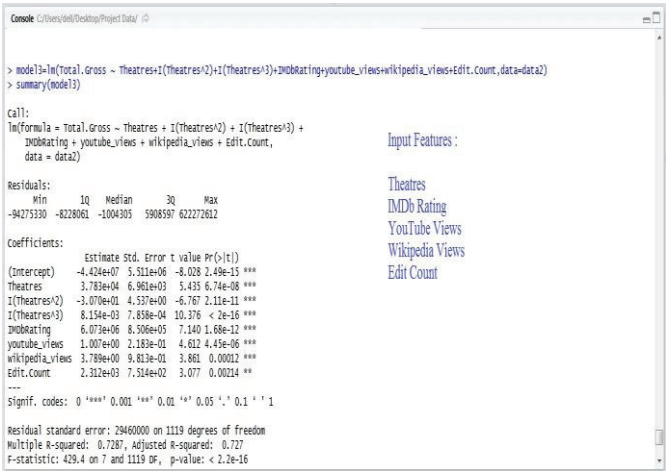

Fig. 5 Predictive model

The figure 5 represents the predictive model built with an R^2 value of 0.7287.

F. Visualization

For visualization we used NetworkD3 package from R language for constructing interactive graph.

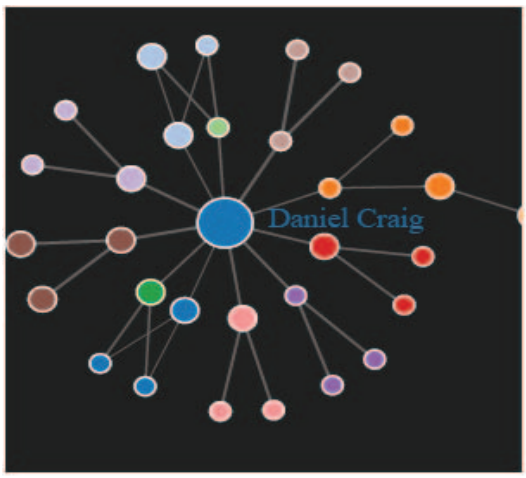

Fig. 6 Visualization

The figure 6 denotes the force networkD3 graph for movie Skyfall. The center node represents the main actor in any particular movie (in our example of Skyfall it is Daniel Craig) and next circular level represents all other movies he has acted in and final subsequent level represents actors in first level movies with links connecting them.

IV. Discussion

A. Analysis in movie production phases

For providing constructive feedback our model can predict the success of movie in two phases such as:

a. Pre-Production Phase

b. Post-Production Phase

In pre-production phase, our model can perform risk analysis and informed decision making by providing the option to try different combinations of classical factors for giving the best possible outcome. This can be helpful for movie stakeholders such as producers for maximizing their return on investment. In post-production phase, stakeholders can look to maximize their movie's outreach by focusing on marketing strategies for social platforms. A general user may test the results of a movie that has already been just released to get the best movie from a list of options.

B. General overview of our product

Our product is a web application for which backend is designed in R language with support of shiny framework.
For hosting the application, services of shiny server, which supports R language, can be utilized.

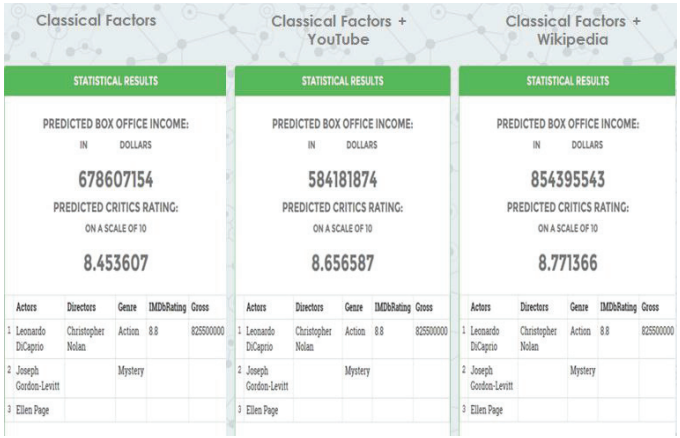

Fig. 7 Comparison of Predictive Models

Figure 7 shows a functionality of our product which provides comparison for predictive models built using different factors for movie Inception.

V. Conclusion

In business, predictive analytics models generate interesting patterns from historical and current data to identify various strengths, risks and opportunities to make prediction about future events. This paper documents the inter-relationships established between various classical factors and social signals used while implementing the predictive model for predicting the total box office collections and critical rating for a particular movie. The results show that the prediction model built using integration of classical as well as social factors can achieve higher accuracy rate. Because the model built can predict the success of movie before its release, it can be used by movie stakeholders for better decision making.

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] Steven Yoo, Robert Kanter, David Cummings; "Predicting Movie Revenue from IMDb Data"; Stanford University, 2011.

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

[7] Jeffrey Ericson, Jesse Grodman; "A Predictor for Movie Success" ; Stanford University, 2013.

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

[9] 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.

[10] Vasu Jain ; "Prediction of movie success using sentiment analysis of tweets"; SCSE 2013.

[11] Marton Mestyan, Taha Yasseri, Janos Kertesz; "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data", 2013.

[12] P.Reggie, C. Andrea ; "Quantifying movie magic with google search"; Google white paper.

[13] Eldar Sadikov, Aditya Parameswaram, Petros Venetis ; "Blogs as predictors of movie success."; Stanford University.

[14] Anand Bhave, Vinay Biramane, Himanshu Kulkarni, Pranali Kosamkar; "Role of different factors in predicitng movie success"; IEEE International Conference on Pervasive Computing.

[15] Global media report , Mckinsey and company.