Predicting the Success of a movie before its release

Pravallika Vanukuri
Computer Science,
Utah State University
Pravalli1703@gmail.com

Sahiti Katragadda
Computer Science,
Utah State University
sahitik1994@gmail.com

ABSTRACT

The film industry has grown immensely over the past few decades generating billions of dollars of revenue for the stakeholders. 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. A prediction system to assess the success of new movies can help the movie studio by giving constructive feedback in pre-production phase. . It can also foretell a society's anticipation towards a movie. Predicting the success of the movies is of interest to economists and investors (media and production houses) as well as predictive analysts. The current predictive models available are based on various factors for assessment of the movie success such as the classical factors like 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. The primary goal of our project is to build a classification model by integrating both classical and the social factors and the study of interrelation among the classical factors will lead to more accuracy. The results show that the prediction model built using integration of classical as well as social factors achieved highest accuracy rate of 78.3%.

INDEX TERMS

Movie Success, Predictive Analytics, IMDB, Classification model, Social media

ACM Reference format:

G. Gubbiotti, P. Malagò, S. Fin, S. Tacchi, L. Giovannini, D. Bisero, M. Madami, and G. Carlotti. 1997. SIG Proceedings Paper in word Format. In *Proceedings of ACM Woodstock conference, El Paso, Texas USA, July 1997 (WOODSTOCK'97)*, 4 pages.
DOI: 10.1145/123 4

1 INTRODUCTION

The movie industry worldwide produces large number of movies every year. 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. Prediction of success in business has been of great interest to the economists and financial experts. 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. .Various stakeholders such as actors, financiers, directors etc. can use these predictions to make more informed decisions. With advent of data analytics, the prediction process has been made intelligent by considering the historical data and employing various data analytical techniques to infer the future events. Such studies have been performed in prediction of movies success as well where success and popularity is measured in terms of the IMDB Ratings.

If IMDB rating > 7 then the movie is "Success"
If IMDB rating between 5 to 7 then the movie is "Average"
If IMDB rating < 5 then the movie is "Fail".

The current predictive models available are based on various factors for assessment of the movie such as the classical factors or the social factors. Accuracy for both the methods is very low. Hence a better method is required. Our project implements integration of both the conventional and the social factors to generate the success of a movie before its release and the study of interrelation among these 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 IMDB is considered, along with the conventional factors resulting in effective integration.

New social media tools are constantly appearing which are enabling people to gather information on films and post comments about movies. These comments can influence the initial prediction about the box office success of a movie. YouTube Trailer reviews often come out a few days before the film is released, therefore, help in prediction the movie success and at the same time influence the box office revenue.

In our experiments, we considered several conventional features collected from IMDB such cast, previous success history, competition factor and other features taken from social media such as YouTube and Twitter. These latter types of attributes include Aggregate Followers (equal to sum of followers of top 3 cast, production house, and director from Twitter), Number of views, Number of likes and Number of dislikes, Number of comments and Sentiment Score (all taken from official trailer of movies on YouTube) . We found that most discriminating feature that played vital role in classifying the Ratings and Income was Sentiment score and previous success history. One of our proposed feature, Aggregate Followers, performed better than traditionally used Top Actor followers. Other features proposed including Likes/Dislikes, views, comments on YouTube, also played vital role in prediction. We used these attributes to measure rating of movie. A number of experiments were performed using Classification models. The best performance was calculated using J48, where sentiment score and previous success history came up as the best discriminating attribute. 78.3% accuracy was achieved while predicting the Rating.

2. Related Work

Though there are many factors that constitute a movie's success, and it is not always clear how they interact, our work attempts to determine these factors through the different attributes, social media, conventional features, and predictive analytics. For many years, researchers have been investigating and generating predictive models for the movies performance. They have used conventional features as well as social media features separately to predict the popularity of movies.

A research has been conducted which presented the comparison of Conventional Features with Social Media features in determining the popularity of movies. Their experiments showed that social media features such as Sentiment Score of tweets related to movies, Number of Views and Comments of movies' trailers on YouTube and fan following on twitter can usefully be utilized to predict the popularity of movie[1].Another study discussed 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 using exploratory analysis of features which resulted in understanding the inter-relationships between them [2].

Vasu Jain in his paper tries to predict the popularity from sentiment analysis of tweets [3]. 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. 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 [4]. 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 [5].

Features for movie success: [6]: To detect the features for movie success in our model, we gathered many related sources and integrated them and implemented in our model. There was an interesting study that focused on detecting the features for movie success. This study emphasized that the accuracy of a predictive model depends a lot on the extraction and engineering of features. Three types of features have been explored to determine the success of a movie: audience-based, release-based, and movie-based features.

Audience-based features are about potential audiences' reception of a movie. The more optimistic, positive, or excited the audiences are about a movie, the more likely it is to have a higher revenue, and vice versa. Such receptions can be retrieved from different types of media, such as Twitter, trailer comments, blogs, news articles, and movie reviews.

Release-based features focus on the availability of a movie and the time of its release. The more theaters that will show a movie, the more likely the movie will have a higher revenue. Many movies are targeted for releases at a certain time. For example, holiday release, as well as seasons and dates of releases (spring, summer, etc.), are commonly utilized in the prediction problem [7]. Some studies also attempted to capture the competition at the time of release, which could negatively affect revenues.

Movie-based features are those that are directly related to a movie itself, including who are on the cast and what the movie is about. The most popular feature for cast members is a movie's star power– whether the movie casts star actors. Star powers of actors have been actor rankings, and the number of actors' Twitter followers [8]. It was agreed that higher star powers are helpful for a movie's success. Moreover, the role of directors in a movie's financial success is often overlooked or downplayed. Another interesting feature is derived from a paper [9] which addressed on how Cast members' previous experience also positively influences revenues .In terms of what a movie is about, features such as genre, MPAA rating, whether or not a movie is a sequel, and run time have often been incorporated into success predictions, as well as in other domains [10].

From our understanding of the literature above, to get a substantial accuracy in prediction of the success of a movie, all possible features should be considered. For this, response, and approval of users on social media should be considered, along with the classical factors. Our project is based on establishing relationships between classical factors and social features.

3. PROPOSED METHODOLOGY

Our system is comprised of two major modules, namely Data Collector and Predictive Engine. The overall methodology is shown in Figure 1. Data Collection is the more significant task of the two; it involves data collection from IMDB, Twitter and YouTube and then data pre-processing. Data Collection and Prediction Model are explained below.

3.1 **Data Collection:**

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. The data that is required for predicting the movie success is scattered across the

internet. Data Collector is the major module as it retrieves information about movies from diverse sources including movies web sites i.e. IMDB, and social media including YouTube and Twitter.

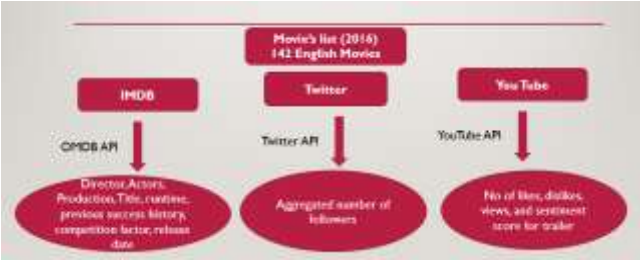

Figure. 1: Data Collection.

3.1.1 Data Description

The data includes the classical factors that are considered for the analysis of the movie, which are obtained from IMDb 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 YouTube view hits, likes, dislikes and sentiment analysis of comments on the pre movie videos. As the data, that we are interested in such as followers count on twitter, ratings on IMDB etc. continuously changes, therefore, we collected the latest data from these web resources by using APIs instead of using already available movies datasets. Attributes are classified into both classification and social media factors.

Classification Factors:

We extract the following movie data from IMDB using OMDB java API [11].

a) Director, Actors, Production, Title:
Cast information like actor, actress and movie director and production team information is collected from IMDB based on English movies titles released in the year of 2016.

b) Runtime:
The total run time of the movie in minutes.

c) Previous success history:
Previous success history represents whether the last movie of the actors, director and production house is a success, failure or average.

d) Competition factor:
Competition factor is calculated based on the number of movies released before and after two weeks of a movie release data.

Competition factor (CF) =1/number of movies released.

e) Release date:
The date, month and year on which the movie got released to theatres.

f) IMDB rating:
The IMDB rating represents the average of the IMDB user ratings of a movie that varies from 0 to 10.IMDB rating information is extracted for the movies released in 2016 to build the prediction model. Verdict of the movie is derived based on IMDB rating.

If IMDB rating > 7 then verdict is "Success"
If IMDB rating between 5 to 7 then the verdict is "Average"
If IMDB rating < 5 then verdict is "Fail".

Social Media Factors:

From social media, we collected following features for each movie based on OMDB data.

a) Aggregated Follower Count:
To understand the popularity of cast, we aggregated the follower count of cast, director and production house. Follower count of cast, director and production house are collected from Twitter using Search Users method of Twitter Search API [7].

To understand how well the movie trailer is received by audience we collected following attributes from YouTube using YouTube Search API [8].

b) Number of Views and Comments:
The number of views and comments of trailer of movies on YouTube are calculated.

c) Number of likes and dislikes:
Similar to the number of views and counts, number of Likes and Dislikes of trailers on YouTube are considered.

d) Sentiment Score:
Sentiment score is calculated for YouTube movie trailer reviews. "R studio" is used to get the sentiment score for each comment on YouTube trailer. This feature is represented by signed integer value. 0 represents neutral sentiment, "+"sign shows the positive sentiment whereas number shows the magnitude. Similarly "–"sign shows negative sentiment. Sentiment score is calculated by retrieving all tweets related to each movie, assigning the sentiment score to each of them and then aggregating the score. It is calculated as below:
Sentiment (Mi) = S1 + S2 … Sn Where, Sentiment (Mi) represents the sentiment score of a movie trailer i. Sn represents the sentiment score of a trailer comment n, and n is the total number of comments linked with movie Mi.

Data set features after Extraction:

Classification Features	Social Media Features
Movie Title Actors Director Release date Producer Movie run time Previous Success History Competition Factor IMDB rating	Sentiment Score Number of Views Number of likes Number of dislikes Aggregated Follower Count

3.1.2 Data Preprocessing

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 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. Data preprocessing involves the conversion of the raw data acquired previously to the usable data. 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 and after 2016. Further, to identify actors or director's twitter account we look for description of the twitter accounts. Additionally, we excluded the twitter accounts with number of followers less than 2000 to identify celebrity accounts. We have considered fan pages with highest followers for the actors who does not have Twitter account.

3.2 Classification Model

As discussed before, we have assumed that popularity is depicted by movie IMDB Rating. For this initial version we classify the movies

into three categories, success, average and failure. In this section we explain our classification model and how it classifies movies into 'success', 'average' or 'failure'.

If IMDB rating > 7 then the movie verdict is Success.

If IMDB rating is between 5 and 7 then the movie verdict is average.

If IMDB rating < 5 then the movie verdict is failure.

We define J48 decision tree classification model that is trained using our pre-processed data set. The model is then validated and tested to predict success or failure of a movie given a set of input data items.

4. RESULTS

We performed experiments with different evaluation strategies such as percentage split, 10 Fold Cross Validation Method; however, best results were obtained with 75% Split of Training and testing data.

$$Accuracy = \frac{Number\ of\ Movies\ with\ exact\ prediction\ of\ rating}{Total\ Number\ Of\ Predictions}$$

Accuracy of our classification model is around 78.3%.

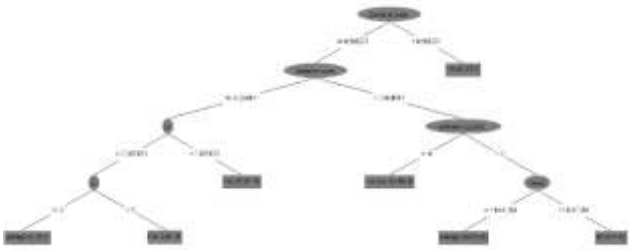

Figure. 2: Decision Tree.

5. ADDITIONAL INSIGHTS

Experiments were conducted to identify the dominant features which played a significant role in predicting movie success. When runtime of the movie is considered, traditionally Indian movies are highly dependent on movie run time. Movies that tend to have more run time are not very successful than the ones with less run time. Therefore, for English movies, we have plotted (Fig 3) and seen if there is any impact of runtime on movie success using the IMDB rating for movies. We noticed that, all the movies have less run time within the range of 94 to 151 minutes. Hence, runtime has no significant role in predicting the rating of a movie

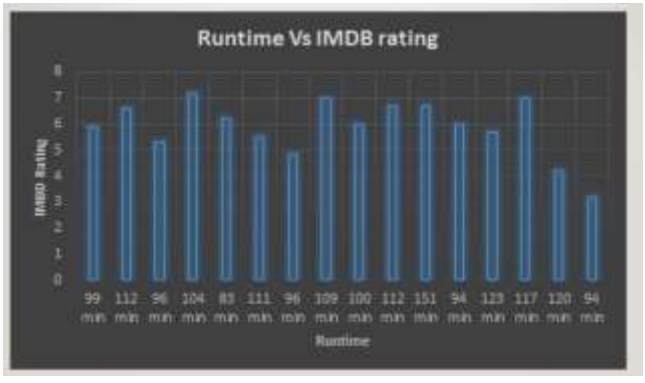

Figure 3: Runtime Vs IMDB rating.

Figure 4 depicts the timeline graph showing number of movies released in a particular month. From the graph we can see that April and September months have highest number of releases compared to other months.

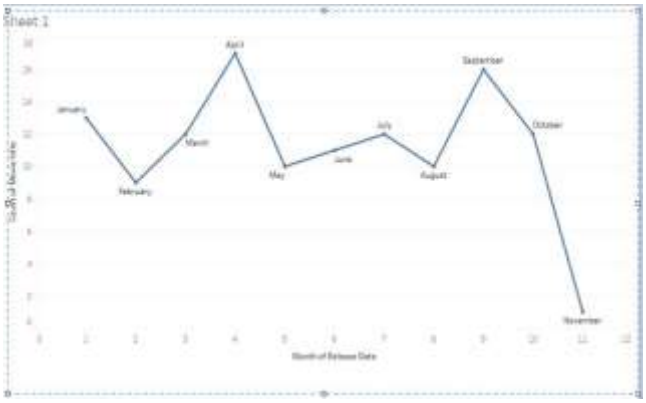

Figure 4: Visualization depicting the release time of a movie

Figure 5 denotes the impact of Average number of followers to predict the rating. This is done by plotting the actual IMDB rating and movie titles. For the same movie titles, average number of followers for cast, director and production house is considered and plotted to see if it follows the same plot.

Figure 5: Visualization showing the impact of Average number of followers on the rating of a movie.

Figure 6 denotes the data set we have taken for the experiment. Total of 125 movies that were released 32 movies were hit, 81 were average and 13 movies were unsuccessful.

Figure 7 depicts the average number of YouTube trailer likes for hit, average and successful movies. From the figure, we can observe that average and hit movies have higher number of likes than flops

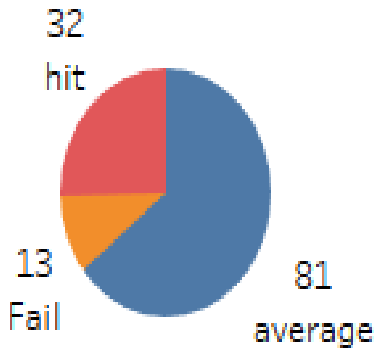

Figure 6: Number of hit, flop, and average movies

.

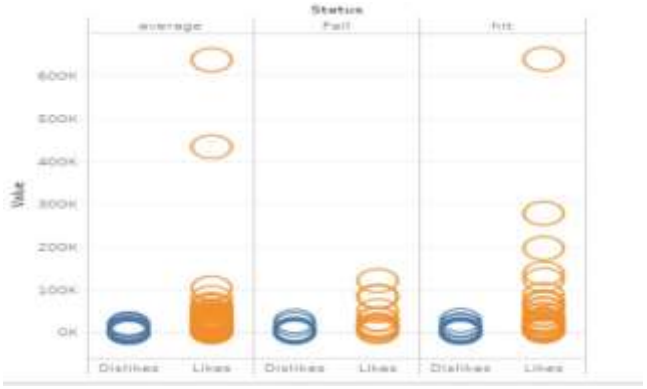

Figure 7: Impact of likes for hit, and average movies is more for Official movie trailer

Figure 8 depicts number of views for average, fail and hit movies. As we can see average movies have higher number of views than flop and hit movies.

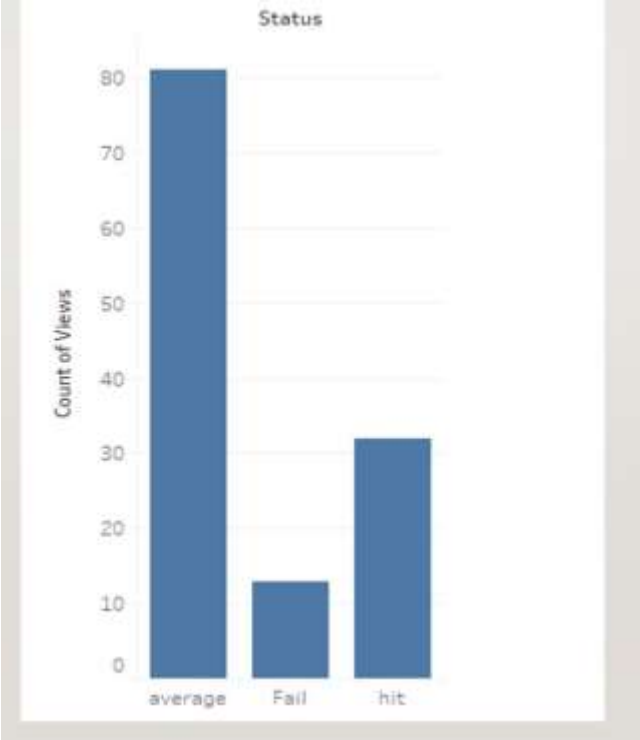

Figure 8: Visualization depicting the views for Official movie trailers from YouTube

6. 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 interrelationships established between various classical factors and social media factors used while implementing the predictive model for predicting critical rating for a movie. Because the model built can predict the success of movie before its release, it can be used by movie stakeholders for better decision making.

Our study suggests that if more data is considered and properly integrated, then greater accuracy can be achieved than considering the classical or social factors individually.

REFERENCES

[1] Mehreen Ahmed, Dr. Awais Majeed, Using Crowd-source based features from social media and Conventional features to predict the movies popularity.

[2] Anand Bhave, Himanshu Kulkarni,Vinay Biramane, PranaliKosamkar; Role of Different Factors in Predicting Movie Success.

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

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

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

[6] Michael T. Lash and Kang Zhao: Early Predictions of Movie Success: the Who, What, and When of Profitability.

[7] Taylor, P.; Simonoff, J. S.; and Sparrow, R. Predicting Movie Grosses : Winners and Losers , Blockbusters and Sleepers. CHANCE, 13, 2 (February 2014), pp 15.

[8] Apala, K. R.; Jose, M.; Motnam, S.; Chan, C. C.; Liszka, K. J.; and de Gregorio, F. 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, Niagra Falls: IEEE Computer Society, 2013 pp 1209–1214.

[9] Meiseberg, B.; Ehrmann, T.; and Dormann, J. We Don't Need Another Hero Implications from Network Structure and Resource Commitment for Movie Performance. Schmalenbach Business Review, 60, 1 (January 2008), pp 74–99.

[10] Abbasi A.; Zahedi F. M.; Zeng D.; Chen Y.; Chen H.; and Nunamaker Jr J. F. Enhancing predictive analytics for anti-phishing by exploiting website genre information. Journal of Management Information Systems, 31, 4 (January 2015), pp 109-157.

[11] http://www.omdbapi.com/.

[12] https://dev.twitter.com/rest/public/search

[13] https://developers.google.com/youtube/v3/docs/