Movie Success Prediction using Historical and Current Data Mining

Partha Chakraborty
Dept. of Computer Science & Engineering
Comilla University
Comilla - 3506, Bangladesh

Md. Zahidur Rahman
Dept. of Computer Science & Engineering
Comilla University
Comilla - 3506, Bangladesh

Saifur Rahman
Dept. of Computer Science & Engg
Comilla University
Comilla - 3506, Bangladesh

ABSTRACT

Movie industry is a multi-billion-dollar business. Lots of movies are being released in every year. All of these movies have different budgets and different cast crew but one thing in common - all want to make profit from movies i.e. make a good box office record. Success of a movie depends on various factors of past and present. Identifying the right factors can predict the profitability of a movie. Some of the factors in predicting movie success are budget, actors, director, producer, IMDb rating, IMDb metascore, IMDb vote count, rotten tomator's tomatometer, actors and director social fan following, wikipedia views, trailer views etc.. The success prediction of a movies plays an indispensable job in film industry since it includes immense investments. Be that as it may, success can not be predicted based on a specific property of a movie. To predict success one have to consider all the properties which can affect movie's success and see how these properties affecting movie's success over time. In this paper, researchers proposed a model where they consider several factors, each factor is assigned by a weight and success/failure of the upcoming movies is predicted based on the factor's value.

General Terms

Data mining, Linear Regression

Keywords

Data mining, Movie success, Prediction, Factors, Weight

1. INTRODUCTION

Predicting society's reaction to a new product in the sense of popularity and adaption rate has become an emerging field of data analysis [1]. There are a huge number of data about movies is available in the web. By studying these data anyone can find connection between some attributes of movies over their success. These factors can be used by producers and production houses to make their creation profitable. The Hollywood movie industry is a massive profitable sector. In 2015, the global box office gross reached an all-time high of $38 billion, with 5 films grossing over 1 billion, the most in the history of Hollywood [2]. Most of the highest grossing films came from just 6 studios: 20th Century Fox, Marvel Studios, Walt Disney Pictures, Columbia Pictures, Paramount and Warner Bros [3]. However movies does not always see the face of success. Though movie industry is profitable place, lots of producers and production houses face a massive loss. Success of a movies is very complex matter because it has larger investment. Larger investment comes with larger risk. The CEO of Motion Picture Association of America (MPAA), J. Valenti mentioned that "No one can tell you how a movie is going to do in the marketplace. Not until the film opens in darkened theatre and sparks fly up between the screen and the audience" [4].

Therefore, the motivation for this research work has come from the recent growing interest about success/failure of movies in the film industry that inspires researchers to work in this context.

2. LITERATURE REVIEW

In 2004, Saraee, White and Eccleston performed analysis of online movie resource of over 390,000 movies and television shows [5]. In 2006, Sharda and Delen worked with predicting financial success of movies even before the movie is released [6]. Classification approach is used where the movies were categorized from flop to blockbuster. Facts and relationships among alternatives can be made by making use of data mining. Some of the factors considered were movie budget and movie popularity relationship, movie cast and movie success relationship. This work helped discover important findings. However, due to copyright rules, there was a challenge involved accessing the data. In 2009, Zhang and Skeina worked on utilizing news analysis to make movie predictions [7]. It was determined that using news data resulted in performance as good as using the IMDb data. Even better performance was achieved using both IMDb data and news data. In 2010, Asur and Huberman worked on predicting outcomes based on social media content [8]. The movie success prediction was based on social media success count, and historical data. The predictions can be made about new movies using this study. However, success prediction cannot be made before the movie is released. In 2015, Lash and Zhao proposed a way to predict decisions about movie investments [9]. This work provided help with investment decision making early in movie production. Historical data was utilized for this work. Some of the features of this work were matching who with what and when with what. The profit was calculated mainly based on box office revenue. However, for many movies, there are other sources of revenue, for example, merchandise.

2.1 Contribution of Researchers in this Work

In this paper, researchers will develop a model which will be used to foresee the success of upcoming movies relying upon specific factors/criteria. The related work discussed can not predict the outcome before it is released. But this model can predict success of movies even before it released. This work utilizes not only historical data but also some present factors at the time of movie released so that it can effectively predict success of a movie waiting to be released.

3. METHODOLOGY

In this work, data mining technique will be used to extract patterns which can be useful in anticipating films success. Data mining approach is significant since it can recognize the connections among different factors[10]. These relationships can in turn help in identifying sequence of events, classification, clustering and predicting future events. Data mining techniques could be used in countless scenarios. Some examples are profit prediction, investment decision, weather forecast, simulations, visualization tools, and medicinal purposes.

4. SYSTEM ARCHITECTURE

The architecture of the system is depicted in fig 1.

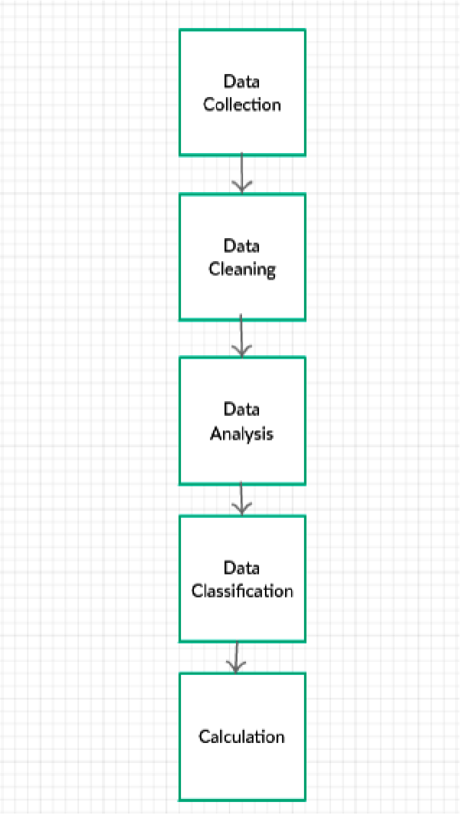


Fig. 1. System Architecture

The system is decomposed into the following modules: -

(1) Data Collection - Collecting data from web using data mining techniques
(2) Data Cleaning - Clean, integrate and store collected data
(3) Data Analysis - Analysis of strong correlation between factors and movie success
(4) Data Classification - Add weight to each factor
(5) Calculation - Calculated mean from all weighting and predict success

The detailed description of these modules will be provided in section 4.1, 4.2, 4.3, 4.4 and 4.5

4.1 Data Collection

The data set was populated with information scraped from IMDb and Rotten Tomato. Trailer views were taken from YouTube. There were information of 1000 movies initially. The features collected include the following:

—Movie Name.
—Release Year.
—Stars (IMDb Credits) - Based on historical data, success rate etc.
—Directors (IMDb Credits) - Based on historical data, success rate etc.
—Production Studio (IMDb Credits) - Based on historical data, success rate etc.
—Budget.
—Total Gross - Box Office collection all over the world.
—IMDb Meta Score (Metacritic) - Critics reviews.
—IMDb Rating - User Ratings.
—IMDb User Vote - User Reviews.
—Tomato meter(Rotten Tomato) - Critics reviews.
—YouTube Trailer Views.
—YouTube Trailer Like-Dislike Ratio - Users reaction.
—Genres - Action, Comedy, Thriller etc.
—Top Actors Instagram Followers.

4.2 Data Cleaning

From the initial data set, movies with incomplete information or junk values are fixed as follows:

(1) Deleting the line with the missing values
(2) Filling up empty fields with specific values
(3) Filling up empty fields with calculations

After completing described procedure, a clean data set with 839 movies information is found.

4.3 Data Analysis

The data set collected for prediction purposes have movies from year 2006 to 2016 [fig 2]. All these movies have all the factors described earlier.

Now lets see how these factors are connected with the success of movies. Correlation of these factors and Box office collection of movies can find by regression analysis between them.

Linear regression endeavors to model the relationship between two variables by fitting a linear equation to observed data. One variable

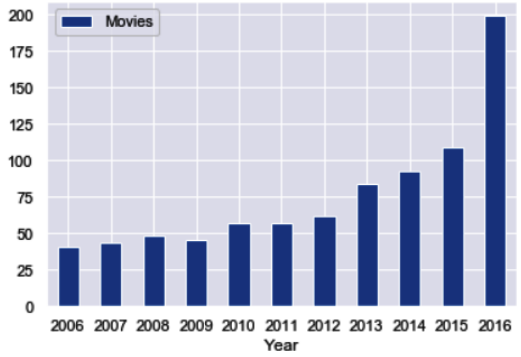


Fig. 2. Movies count from each year

is viewed as an explanatory variable, and the other is viewed as a dependent variable. A linear regression line can be presented as

$$Y = a + bX \tag{1}$$

where X is the explanatory variable and Y is the dependent variable.A multiple regression has multiple ‘X’ or independent variables in one formula [11]. Multiple linear regression endeavors to model the relationship between at least two explanatory variables and a response variable by fitting a linear equation to observed data. Formally, the model for multiple linear regression can be presented as,

$$Yi = b0 + b1Xi1 + b2Xi2 + ... + bpXip + ei \tag{2}$$

where researchers consider n observations of one dependent variable and p independent variables b0, b1, etc. represent parameters to be estimated, ei is the i'th independent identically distributed normal error.

Researchers will use linear regression for showing the correlation between movie success (Box office collection) and factors individually such as IMDb rating, IMDb metascore, IMDb vote counts, Tomatometer etc.

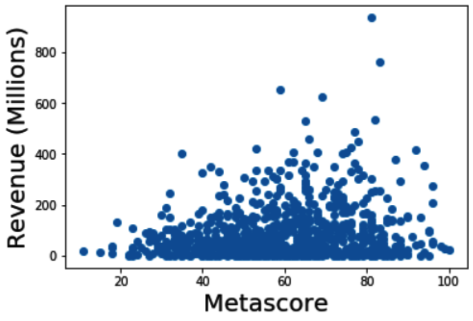
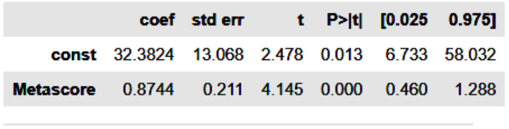


Fig. 3. IMDb metascore influence on revenue

At First, researchers will see the effect of IMDb metascore of movie’s dataset of multiple years. From fig 3, it is seen that IMDb metascore has strong influence in Box office gross of movies throughout the year. This can be made sure from OLS regression table. From observing the OLS regression table (fig 4), the p value is very low, so researchers can reject the null hypothesis and say that Revenue and Metascore are independent and the two attributes are strongly correlated.



	coef	std err	t	P>|t|	[0.025	0.975]
const	32.3824	13.068	2.478	0.013	6.733	58.032
Metascore	0.8744	0.211	4.145	0.000	0.460	1.288

Fig. 4. OLS regression table for revenue (in millions) and metascore

Now lets find the correlation between IMDb ratings and movie revenue (in millions)(fig 5)

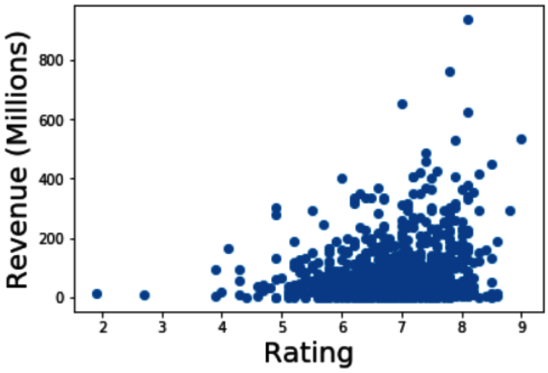
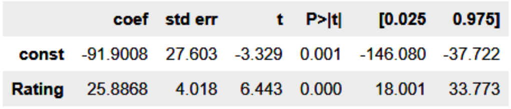


Fig. 5. IMDb ratings influence on revenue

Researchers have found that IMdb ratings also has strong influence in Box office gross of movies. From observing the OLS regression table (fig 6), the p value is also very low. So, it can be said that Revenue and IMDb ratings both attributes are strongly correlated.



	coef	std err	t	P>|t|	[0.025	0.975]
const	-91.9008	27.603	-3.329	0.001	-146.080	-37.722
Rating	25.8868	4.018	6.443	0.000	18.001	33.773

Fig. 6. OLS regression table for revenue (in millions) and IMdb ratings

Researchers have also found the same result for IMDb vote counts. IMDb vote counts has a strong inflence and strongly connected with revenue (fig 7,8).

From these factors analysis researchers have found a strong connection between described factors. These factors plays a great rule in movie success. By using these factors researchers can easily calculate success rate of upcoming movies.

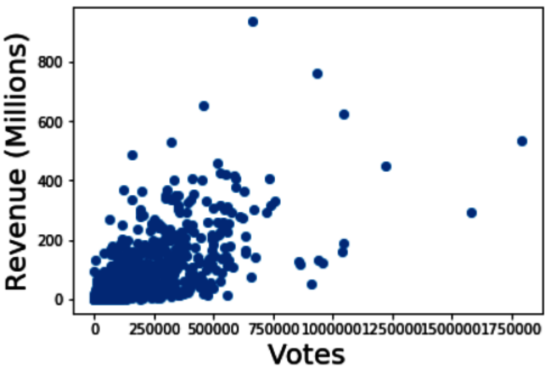
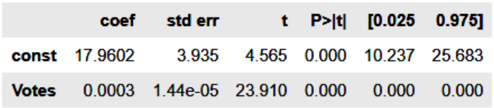


Fig. 7. IMDb vote counts effects on revenue



	coef	std err	t	P>|t|	[0.025	0.975]
const	17.9602	3.935	4.565	0.000	10.237	25.683
Votes	0.0003	1.44e-05	23.910	0.000	0.000	0.000

Fig. 8. OLS regression table for revenue (in millions) and IMDb vote counts

4.4 Data Classification

Every categories of data from the data set are classified in 4 categories[fig 9]. Each categories are assigned with a weight from 0.3 to 1.0.

—Poor - 0.3

—Average - 0.5

—Good - 0.7

—Very Good - 1.0

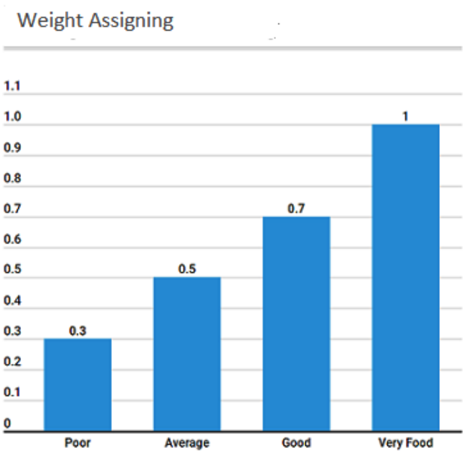


Fig. 9. Assigning weight into factors

The table (fig 10) described how these classifications are determined relative to every actor. Such as Actors Credit from IMDb database indicate the historical analysis of an actor. For current data we have used YouTube trailer views, like-dislike ratio etc. All these factors data are classified into categories described earlier and assigned with a weight value to calculate the success probability.

Factor	Poor	Average	Good	Very Good
Actors Credit (IMDb)	0-20	21-40	41-60	More than 60
Directors Credit (IMDb)	0-20	21-40	41-60	More than 60
YouTube Trailer Views (in millions)	20	20-50	51-100	More than 100
Wikipedia Page Views (in millions	0-2	2-5	6-10	More than 10
YouTube Like-Dislike Ratio	More than 0.5	0.3 -0.5	0.2 -0.3	Less than 0.1
Production Studio Credit (IMDb)	0-20	21-40	41-60	More than 60
IMDb Meta Score	Less than 40	40-60	61-80	More than 80
Rotten Tomato Meta Score	Less than 40	40-60	61-80	More than 60
Stars Instagram Followers (in millions)	Less than 1	1-5	5-10	More than 10

Fig. 10. Data classification

4.5 Calculation

From the classified data, now researchers can predict the success of a movie with weight assigned to every factor.
The prediction model is

$$X = \sum(W)/N \tag{3}$$

Here, 'X' is success probability; $0 < X > 1$ and 'W' is the weight of every factors and 'N' is the number of factors.

Researchers have classified a movie's success into 6 categories by considering the value of 'X' (fig 11).
These are-

—Disaster - If X is bellow 0.3

—Flop - If X is between 0.3 and 0.4

—Average - If X is between 0.4 and 0.5

—Hit - If X is between 0.5 and 0.6

—Super Hit - If X is between 0.6 and 0.8

—Blockbuster - If X is more than 0.8

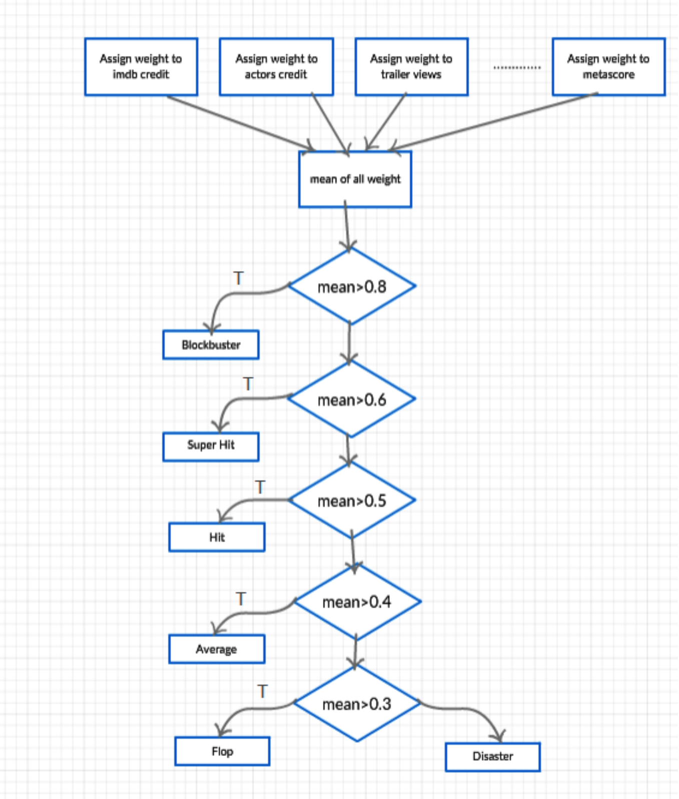


Fig. 11. Prediction of success

5. EXPERIMENTAL RESULT

Researchers have experimented the model with some movies from year of 2018. The experimental results are shown below.

Table 1. Experimental results

Title	Mean	Prediction
The Curse of La Llorona	0.45	Average
Under the Silver Lake	0.575	Hit
Little Woods	0.52	Hit
Avengers: Endgame	0.81	Blockbuster
Penguins	0.34	Flop
Captain Marvel	0.75	Super Hit
Aladin	0.64	Super Hit
High on the Hog	0.45	Average
The White Crow	0.48	Average

It can be seen from the table that this model can successfully predict the outcome of these movies with given attributes.

6. CONCLUSION

In this work, researchers have developed a model to find the success of upcoming movies based on certain factors. But, a movie success does not depend only on those features related to movies. The number of audience plays a vital role for a movie to become successful. Because the whole point is about viewers, the entire industry will make no sense if there is no audience to watch a movie. The number of tickets sold during a specific year can indicate the number of viewers of that year. And the role of movie audience depends on many situations like political conditions and economic stability of a country. So these facts play a vital role in an ultimate success of a movie. So, for future work, researchers suggest considering these features.

7. REFERENCES

[1] Nahid Quader, Md. Osman Gani, Dipankar Chaki, and Md. Haider Ali, “A Machine Learning Approach to Predict Movie Box-Office Success,”in 20th International Conference of Computer and Information Technology (ICCIT), December 2017.

[2] Forbes (2016) ”Experts Predict a Drop in Box Office Revenue In 2016 After a Record Year for Hollywood”, https://www.forbes.com/sites/simonthompson/2016/01/05/experts-predict-a-drop-in-box-office-revenue-in-2016-after-a-record-year-for-hollywood/402059897195

[3] Subramaniyaswamy V., Viginesh Vaibhav M., Vishnu Prasad R. and Logesh R., “Predicting Movie Box Office Success using Multiple Regression and SVM,” in the International Conference on Intelligent Sustainable Systems (ICISS 2017).

[4] J. Valenti (1978). Motion Pictures and Their Impact on Society in the Year 2000, speech given at the Midwest Research Institute, Kansas City, April 25, p. 7.

[5] M. Saraee, S. White, and J. Eccleston. A data mining approach to analysis and prediction of movie ratings, 2004.

[6] Ramesh Sharda and Dursun Delen. Predicting box- office success of motion pictures with neural networks. Expert Systems with Applications, vol 30, pp 243-254, 2006.

[7] W. Zhang and S. Skiena. Improving movie gross pre-diction through news analysis. In Web Intelligence, pages 301-304, 2009.

[8] Sitaram Asur and Bernardo A. Huberman, Predicting the Future with Social Media, http://arxiv.org/abs/1003.5699, March 2010.

[9] Michael T. Lash and Kang Zhao, Early Predictions of Movie Success: the Who, What, and When of Profitabil-ity, June 2015.

[10] Jiawei Han, Jian Pei, and Micheline Kamber. Data Mining Concepts and Techniques, 2012.

[11] Linear Regression,http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm