Analysing Success Rate of Movies Using Data Mining

Rohan Wagle[1] Yashovardhan Pandya[2] Priyanka Jagtap[3] Prof.Priyanka Bhilare[4]

[1,2,3,4]Rajiv Gandhi Institute of Technology, India

Abstract— Given the low success rate of movies, models and mechanisms can be used to predict the success of a movie. It will help the business significantly. Various stakeholders such as actors, producers, directors etc. can use these predictions to make more informed decisions. They can make the decision before the movie release. Historical data of each component such as actor, actress, and director, composer that influences the success or failure of a movie is given due to its weightage. This proposed work aims to develop a model based upon the data mining techniques that may help in predicting the success of a movie in advance thereby reducing certain level of uncertainty. The system is used to predict the past as well as the future of movie for the purpose of business certainty or simply a theoretical condition in which decision making (the success of the movie) is without risk, because the decision maker (movie makers and stake holders) has all the information about the exact outcome of the decision, before he or she makes the decision (release of the movie).

Keywords: Rate of Movies, Data Mining

I. INTRODUCTION

With over two million spectators a day and films exported to over 100 countries, the impact of Bollywood film industry is formidable. In particular, we concentrate on attributes relevant to the success prediction of movies, such as whether any particular actors or actresses are likely to help a movie to succeed. The proposed system reports on the techniques used, giving their implementation and usefulness. The important issue involved in the prediction system is, IMDb is difficult to perform data mining upon, due to the format of the source data. We also found that, the budget of a film is no indication of how well-rated it will be, there is a downward trend in the quality of films over time. Another important factors are the director and actors/actresses involved in a film.

II. LITERATURE REVIEW

In 2012,1. Ajay Shiva Santosh Reddy,Pratik Kasat,Abhiyash Jain published a paper titled "Box-Office Opening Prediction of Movies based on Hype Analysis through Data Mining".This paper emphasises that Pre-release hype is an important factor for estimation of revenue for a movie.However this was quite difficult to accomplish.

In Krushikanth R. Apala, Merin Jose, Supreme Motnam, C.-C. Chan, Kathy J. Liszka, and Federico de Gregorio's paper published in 2013,titled "Prediction of Movies Box Office Performance Using Social Media " the sentiment analysis of social media platforms such as YouTube,Twitter,IMDb using linear regression to determine whether a movie is a hit, flop or neutral. The drawback was that there was no expansion of neutral class and characterization of movie box office performance in terms of net profits and profit ratios.

In 2014,Nithin VR,Pranav M,Sarath Babu,Lijiya A published a paper named "Predicting Movie Success Based on IMDB Data" which used Linear Regression Model, Logistic Regression model, Support Vector Machine Regression Model were used to predict movie success using IMDB data. The limitation was that the success percentage for all models, are not good enough of industrial use. The training set used is small in size hence the results could be improved by using larger training sets.

III. LIMITATIONS OF EXISTING SYSTEMS

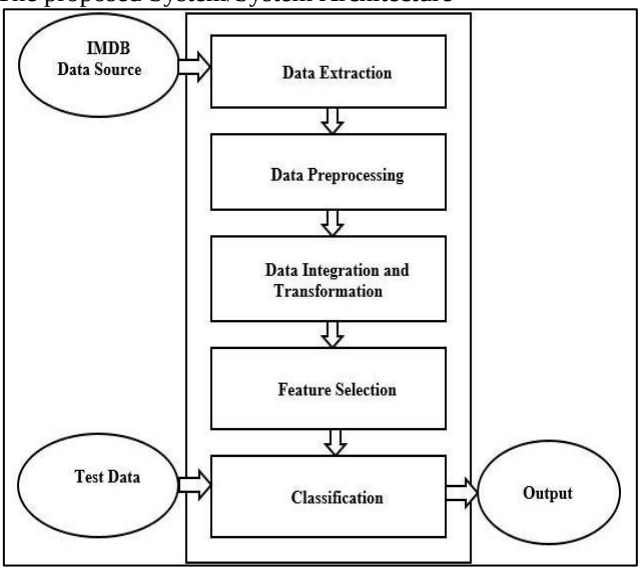
- Hype factor is not a good basis for predicting movie success.
- Discussion of movies in forums cannot be the sole mode of operation in determining the success rate of a movie.
- Success percentages for all models are not good enough for industrial use. Training set is small hence results could be improved.
- Chi square method does not provide us with precise results.

IV. PROPOSED SYSTEM

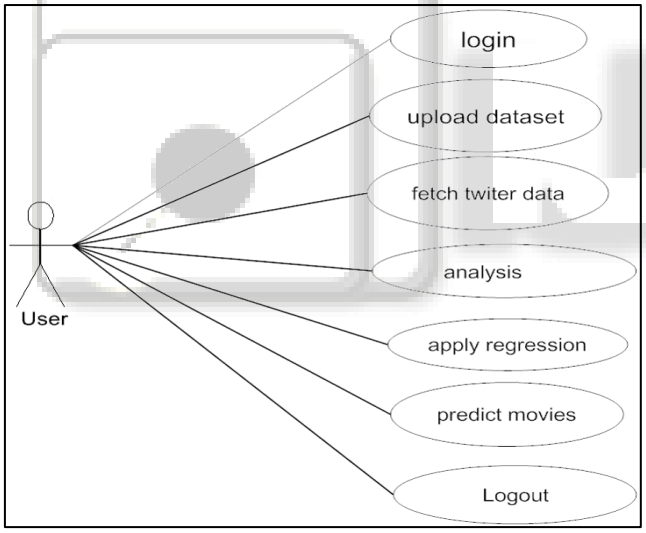
The proposed system aims to develop a system based upon data mining techniques that may help in predicting the success of a movie in advance thereby reducing certain level of uncertainty. Our system makes use of historical data in order to successfully predict the ratings of movies. We will be using multi linear regression based mathematical model to determine the results. Linear regression is a statistical approach that models the relationship between a scalar dependent variable y and one or more explanatory variable X. Multiple Linear Regression is the extension of linear regression which models the relationship between one dependent variable with two or more explanatory variables by trying to fit linear equation on observed data.

V. DESIGN DETAILS

The proposed System/System Architecture

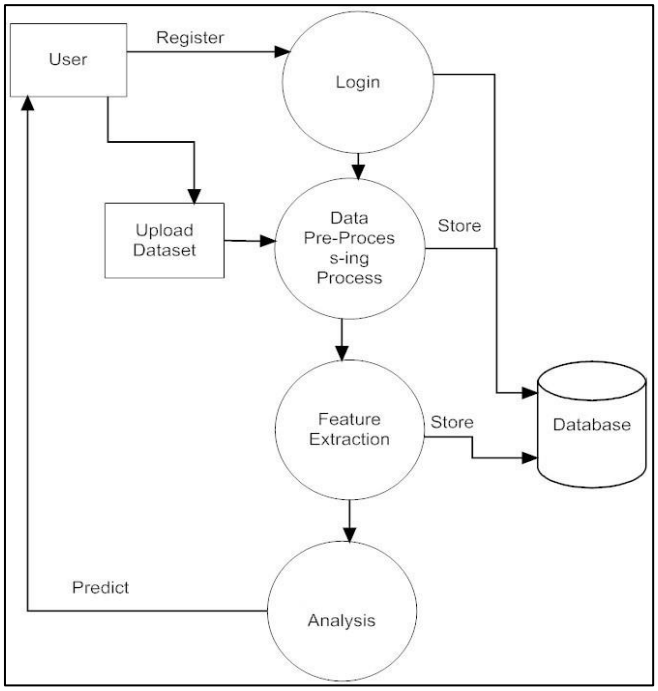


- Our goal is to predict the ratings of new movie based only on the ratings of the training set examples.
- We will be considering 5000 training examples, each of which has features and the corresponding human ratings of some particular attribute. We then form a prediction model that can be used to predict the human rating when the computational features are known.
- The goal is to define a relationship between the prediction value and the features by solving for the linear coefficients, that best map the features to the prediction value. Where, the ratings have been collected in a vector Y. Y is a (m x 1) vector. In our case m=5000.
- The entire system predicts the future ratings based on the training set of 5000 movie details. To fetch the features associated records of the features we will use association rule mining. This movie set will be pruned to select a set of features that have been found to make a major impact on the success or failure of a film. After the identification we find features all the producers, directors, actors and actresses were rated based on their past performance at the Box Office. Similarly months of the year and run times of movie will be assigned a score based on the same criteria. After obtaining a numerical equivalent of the movie database, the features will be normalized to numerical value



We see in the above use case diagram that first the user must login into the system,then upload the IMDB dataset. Once the dataset is uploaded the user can fetch its data and apply regression algorithm in order to make a prediction,once computation is done and result is obtained,the user must logout.

VI. METHODOLOGY



The value that we want to predict is called the dependent variable. Let us consider number of explanatory or independent variables is p. So the variables can be denoted as x1,x2,....,xp. The regression line can be defined as y = 0 + 1x1 + 2x2 + ... + pxp. Example: The selling price of a house can depend on the desirability of the location, the number of bedrooms, the year the house was built and number of other factors. Similarly, Movie success prediction can also depend on month of release, public holidays, star ratings and number of other factors. The aim is collecting number of likes, dislikes and view count of trailer, release date, star ranking. Multiple Linear Regression Algorithm is used which was discussed above for the prediction of earnings of the movie.

VII. CONCLUSION

This project had as a main goal to develop a model able of predicting the box office financial success of a certain set of movies through specific variables and historical data. It was possible to conclude that the percentage of success of the cinematographic revenue prediction is quite different based on the typology of the dependent variable used in the study. The empirical model demonstrated good statistical results when the dependent variable was binary and interval.

ACKNOWLEDGEMENT

We owe our profound gratitude to our subject teacher, Prof. Priyanka Bhilare for giving us the opportunity to carry out the research work. We are thankful for all the support and guidance from the Computer Department of Rajiv Gandhi Institute of Technology who helped us in our project work.

REFERENCES

[1] Javaria Ahmad, Prakash Duraisamy, Amr Yousef†, Bill Buckles, ”Movie Success Prediction Using Data Mining

” 8th ICCCNT 2017 July 3 -5, 2017, IIT Delhi, Delhi, India

[2] Nithin VR, Pranav M, Sarath Babu PB, Lijiya, “A Predicting Movie Success Based on IMDB Data” International Journal of Data Mining Techniques and Applications 2014

[3] Krushikanth R. Apala, Merin Jose, Supreme Motnam, C.-C. Chan, Kathy J. Liszka, and Federico de Gregorio. “Prediction of Movies Box Office Performance Using Social Media.” 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

[4] Ajay Siva Santosh Reddy Pratik Kasat, “Box-Office Opening Prediction of Movies based on Hype Analysis through Data Mining”, International Journal of Computer Applications 56(1):1-5, October 2012. Published by Foundation of Computer Science, New York, USA. DOI: 10.5120/8852-2794

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

[6] Basuroy, S., Chatterjee, S. and Ravid, S. A. (2003). How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets. Journal of Marketing, 67(4), 103–117.

