Volume: 03, June 2014, Pages: 365-368 International Journal of Data Mining Techniques and Applications
ISSN: 2278-2419

Predicting Movie Success Based on IMDB Data

[1]Nithin VR, [2]Pranav M, [3]Sarath Babu PB, [4]Lijiya A

[1]Student, Department of CSE, National Institute of Technology, Calicut

[2]Student, Department of CSE, National Institute of Technology, Calicut

[3]Student, Department of CSE, National Institute of Technology, Calicut

[4]Assistant Professor, Department of CSE, National Institute of Technology, Calicut

[1]vrnithinkumar@gmail.com, [2]pranavm@gmail.com, [3]srth12@gmail.com, [4]lijiya@nitc.ac.in

Abstract

American film studios collectively produce several hundred movies every year, making the United States the third most prolific producer of films in the world. The budget of these movies is of the order of hundreds of millions of dollars, making their box office success absolutely essential for the survival of the industry. Knowing which movies are likely to succeed and which are likely to fail before the release could benefit the production houses greatly as it will enable them to focus their advertising campaigns which itself cost millions of dollars, accordingly. And it could also help them to know when it is most appropriate to release a movie by looking at the overall market. So the prediction of movie success is of great importance to the industry. Machine learning algorithms are widely used to make predictions such as growth in the stock market, demand for products, nature of tumors etc. This paper presents a detailed study of Logistic Regression, SVM Regression and Linear Regression on IMDB data to predict movie box office.

Key words: Data mining, Logistic Regression, SVM Regression, Linear Regression

I. INTRODUCTION

In the United States of America 1000s of films are released every year. Since the 1920s, the American film industry has grossed more money every year than that of any other country [1]. Cinema in America is a multi-billion dollar industry where even individual films earn over a billion dollars. Large production houses control most of the film industry, with billions of dollars spent on advertisements alone. Advertising campaigns contribute heavily to the total budget of the movies. Sometimes the investment results in heavy losses to the producers.

Warner Brothers, one of the largest production houses had a fall in their revenues last year, despite the inflation and the increased number of movies released. If it was somehow possible to know beforehand the likelihood of success of the movies, the production houses could adjust the release of their movies so as to gain maximum profit. They could use the predictions to know when the market is dull and when it's not.

This shows a dire need for such software to be developed. Many have tried to accomplish this goal of predicting movie revenues. Techniques such as social media sentiment analysis have been used in the past. None of the studies thus far has succeeded in suggesting a model good enough to be used in the industry.In this study, we attempt to use IMDb data to predict the gross revenue of the movies as well as their IMDB rating.

The organization of the paper is as follows. Section II describes the role of dataset collection and preprocessing in data mining. Regression methods have discussed in section III.

Results are shown in section IV and Section V concludes the paper.General Design is shown in the Figure 1.

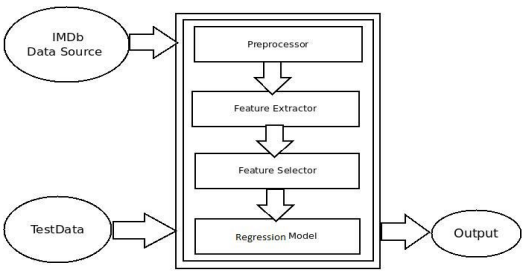

Figure 1 General Design

II. Dataset Collection and Preprocessing

Dataset Collection

The initial dataset to be used will be collected from IMDB. It will consist of movies that were released from 2000 to 2012. Among these movies, we only selected the ones that were released in the United States and are in English, in the anticipation that we would be able to make more accurate predictions on these movies given that their reviews would also be in English. We removed movies which don't have any information about Box office details. We got data regarding 1050 films.

A. Data preprocessing

The data we obtained are highly susceptible to noisy, missing and inconsistent data due to the huge size and their likely origin from multiple, heterogeneous sources [2]. We mainly used IMDb and Rotten Tomatoes and Wikipedia. The main problem with datasets was missing fields. To overcome this missing field problem we adopted a method which uses a measure of central tendency for the attribute. We used both mean and median as central tendency. Then removed duplicate items.

B. Data Integration and Transformation

Data obtained from three different resources IMDB, Wikipedia and Rotten Tomatoes were then integrated into one database.In this step integrated data are transformed or consolidated so that the regression process may be more efficient and easier. Dataset is mixed with both nominal [3] and numeric attributes, but for a regression process, we need all attributes to be numerical. We used a measure of central tendency of Box office revenue to convert corresponding nominal attributes to numerical.

Type	Features
Nominal	Actors, Director, Writer, Production-House, Genre
Numerical	Budget, IMDB Rating, No of Rating, IMDB Votes, Metascore, Tomato Meter, Tomato User Rating, Tomato Reviews, Tomato Fresh, Tomato Rotten.

Table 1 Features

C. Selecting feature Subset

Here we look for correlation between all the different features under consideration and look for correlations with the target variable, i.e. movie revenue in dollars. We also look for correlation between the features themselves in order to avoid redundancy and irrelevant attributes. Also the extent of correlation is noted. Here we are using the greedy backward procedure to get best feature subset. The procedure starts with the full set of attributes. At each step, it removes the worst attribute remaining in the set. At the end of the process, features that are not correlated with the target variable and those which are redundant are eliminated. Major steps involved in the greedy backward procedure have shown in Figure 2

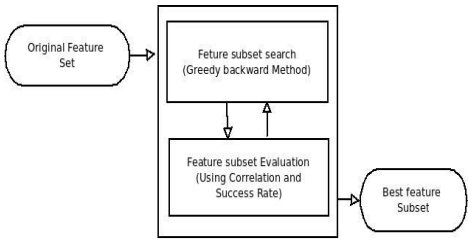

Figure 2: Feature Selection

Supervised learning technique is adopted for this Study. We are using three models to predict the revenue and we will compare the performance of the different methods.

A. Linear Regression Model

In our first model, we use standard least-squares linear regression [4]. To do this, we intend to use stochastic gradient descent. Once we have trained a set of feature weights, we could then generate gross revenue predictions as follows:

$$\text{Gross} = \Theta_0 + \Theta_1 * F_1 + \Theta_2 * F_2 + \ldots + \Theta_n * F_n \quad (1)$$

Where Θ_i are the weights, F_i are the features, and n is the number of features.

B. *Logistic Regression model*

We chose Logistic regression as our second method mainly because it generates a multi-class model with linear weights, most directly comparable to the feature weights given by linear regression [5]. For logistic regression to apply we need to change the problem which is

a regression problem to a classification problem. To achieve this we split the range of the target variable into a finite number of buckets of equal size. To define our classes we draw a histogram of movie revenues to create different buckets for prediction. These buckets are continuous ranges of movie revenues, which covers the entire sample space.

C. *Support Vector Machine Regression Model*

SVMs can also be applied to regression problems [6]. SVM Regression tries to find a function f(x) that has at most ε deviation from the actually obtained targets y for all the training data, and at the same time is as flat as possible. SVM Regression does not care about errors as long as they are less than ε .We used linear kernel function to map the data into a high dimensional feature space where linear regression is performed. Since Training data is not large as compared to the number of features, we used a linear kernel function Hyper parameter C optimization was done using grid search. Grid search is exhaustive search through a manually specified subset of the hyper parameter space. Cross validation on the training was used as the performance matrix

IV. Results

The result we found out using linear regression was about 51% accurate. Where as for logistic regression, we got 42.2% accuracy which is a comparatively low result. SVM approach had a success rate of 39%. The error tolerance for SVM and Linear regression was 20%. For logistic regression the error tolerance was 12.5%. Of all the 20 features in the data-set, budget, director, writer, actor1, actor2, gender, tomato reviews were found to be the most significant features.

Results are shown in the table 2.

Model	Linear regression	Logistic regression	SVM regression
Tolerance	20 %	12.5%	20%
Success Rate	50.7%	42.2%	39.0%
Correlation	0.965		0.965

Table 2 Results

V. Conclusion and future work

After building the three models we found out that the linear model represents the movie features more accurately. The success percentage for all models, while not good enough for industrial use, is in the close proximity of values obtained in previous studies. Some of the results obtained are better than that of some standard libraries and similar studies. Even though results are not good enough for industrial purposes the models built can be used in online applications.

A larger training set is the key to improving the performance of the model. We need to consider additional features to achieve this. News analysis [7], plot analysis [8] and Social Network's data [9] could be done and the information thus obtained could be added to the training set. We can also use Google trends [10] result to improve the result.

REFERENCES

[1] Darin Im, Minh Thao, Dang Nguyen, "Predicting Movie Success in the U.S. market," Dept.Elect.Eng, Stanford Univ., California, December, 2011

[2] Jiawei Han, Micheline Kamber, Jian Pei,Data Mining Concepts and Techniques, 3rd ed.MA:Elsevier, 2011, pp. 83-117

[3] Richard O. Duda, Peter E. Hart, David G. Stork,Pattern Classi_cation, 2nd ed.NewYork: Wiley, 1973

[4] Cohen, J., Cohen P., West, S.G., & Aiken, L.S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences. (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates

[5] Christopher M. Bishop (2006). Pattern Recognition and Machine Learning. Springer. p. 205. "In the terminology of statistics, this model is known as logistic regression, although it should be emphasized that this is a model for classification rather than regression."

[6] Cristianini, Nello; and Shawe-Taylor, John; An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, 2000. ISBN 0-521-78019-5

[7] W. Zhang and S. Skiena," Improving movie gross prediction through news

analysis,", IEEE/WIC/ACM International Conference on Web

Intelligence and Intelligent Agent Technology, Milan, 2009

[8] Sagar V. Mehta, Rose Marie Philip, Aju Talappillil Scaria, "Predicting Movie Rating based on Text Reviews," Dept.Elect.Eng, Stanford Univ., California, December, 2011

[9] Suhaas Prasad, "Using Social Networks to improve Movie Ratings predictions," Dept.Elect.Eng, Stanford Univ., California, 2010

[10] Deniz Demir, Olga Kapralova, Hongze Lai, "Predicting IMDB Movie Ratings Using Google Trends," Dept.Elect.Eng,Stanford Univ., California, December, 2012