International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Pre-Release Success Quotient Prediction of Movies

Harsh Taneja[1], Anupam Dewan[2], Vineet Bhardhwaj[3]

[1]Assistant Professor, Dept. of Computer Science, GGSIPU, BVCOE, Paschim Vihar, New Delhi, India

[2,3]UG students, Dept. of Computer Science, GGSIPU, BVCOE, Paschim Vihar, New Delhi, India

Abstract: ***In this research work we have developed a mathematical model for predicting the success quotient of the movies in terms of revenue model and rating model. We have develop a model in which the past data record of each component [e.g. actor, actress, genre, director] that influences the success or failure of a movie is considered. 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 costs a lot.***

Keywords: Movie, Movie Revenue prediction, Big Data, Movie Rate Prediction, Box Office Collection, nlfit, hougen, mrfit, mrval, logistic, MSE, Neural Networks, Regression

1. Introduction

With the spectacular growth of movie industry, we now have access to a vast collection of data which could be used to analyze the past trend and the possible future outcome.These metrics are a great source for making predictions about the success of the products or services they describe and a powerful parameter to be included in existing economic models.

Huge scattered data sets of various fields could be used to develop a model or strategy to predict the following outcome. Here, we have conceptualized how combinations with a robust function would help us to predict the successfulness of a movie. However there cannot be any single strategy to examine whether the movie (upcoming would be a hit or would do badly).Here an attempt has been made to analyze the rate of successfulness of the movie on various parameters.[2]

The research aims at predicting the success of a film at the box office well before it appears in cinemas by looking into the past combinations of an actor, actress, director and revenue earned by those combinations at box office and the how fruitful would it be when a new combination of the above said hit the box office.

The belief is that the results of their research can be used to improve marketing efforts in the film industry and predicting if the film is going to show ducks or miracle in the cinema after release. Moreover a criteria is set up to provide a prerelease rating to a movie which would help the end customers to judge the upcoming movie.

Following are the crucial spacing that we have observed while researching in context of the topic are been highlighted as follows:

- **Lack of Combinational-analysis**
  No, Combinational Analysis has been taken forward (Combination of co-actors, directors, producers etc.) as of now. We hereby attempt to exploit this by using Combinational Analysis in conjunction with other social media buzz.
- **Lack of Comparative Analysis among various proposed algorithm**
  All research paper generally gives an idea related to one single algorithm proposed and its implementation. We hereby try to compare various algorithms on the ground of performance and efficiency**.**
- **Lack of any Data Visualization done so far**
  No, Data Visualization tool is as of now generated to proof the prediction so far. We attempt at providing insightful visualization for the same.

The power of big data and analytics is used to come up with some amazing insights that can be really helpful in determining to the aspect of the fact of pre-release hit and flop of a movie and box office collection.

The objective task of the project is to make an attempt to try at analysing the past performances on the records of the movies of past years. This data would then be used as a training set to come to a point of predicting the gross sales and estimate the possibility of the successfulness of the movie in that particular context.

Later, we also wish to compare on the grounds of various algorithm and device better and more efficient algorithms for the purpose of better possible outcomes.

Moreover, we also wish to use the data visualization schemes to come up with interesting visualizing aspects of data using the modern age tools.

2. Identify, Research and Collect Idea

The basic idea of the research was to use the available movies data on the internet and with the help analytics on the data to come to a meaningful and useful data which would help in predicting the success quotient of a movie prior to its release. It is used as a success of a movies depends on certain important parameters' such as the actors and the director which have been used to predict the movie prerelease outcome.

The combinations of the parameters in the past and how successful were they at that time and, if a new combination among those old parameters is done how would be the outcome is done with the help of the past dataset and analytics tools.

We developed a model that predicts whether themovie flop, hit, or super hit, for the propose we created a historical data set related to different parameters that influence movie success and developed an algorithm that assigned weights and developed a mathematical model to compute and predict movie revenue.

Dataset Collection
The initial dataset to be used was collectedfrom gomolo.com. It consisted all the Bollywood movies sorted according to their names. Among these movies, we selected all the Bollywood movies after the year 2000, in the anticipation that we would be able to make more accurate predictions on these movies given that the data would be richly available and the revenue records could be easily found. The ratings were drawn forth from IMDB and the revenue of the movie was drawn forth from Wikipedia. The dataset includes the year of the release of the movie, the title of the movie, the director, the main lead cast, genre, ratings and the revenue records.

A model which considers a combination of various parameters like director, genre, actor, and actress and checks the association to each other is used to predict the outcome of the new movie.

Implementation Steps
The entire procedure while developing was conducted in a following manner aslisted below

- **Understanding and gathering of data**
  The first step is to understand and gather the data and populate the database with the entries the database may be obtained from many sources Manual database entry filling, from Gomolo.com, IMDB and various other data sources from repositories of Bollywood movies.

- **Cleaning the data and applying the ETL tools**
  The data after gathering is then needed to be cleansed and hence we require ETL strategy to cleans data and get a database ready after performing the elimination of entities that have redundancies in it.

- **Modelling and developing analytics behind the data**
  Next step is to develop and apply the algorithm to churn out the data and come out with the insights and the findings revealed from the data. This step requires application of various effective algorithm to finally land up with some predictive nature of the values so found out.

- **Evaluating and analysing the data**
  The data so obtained now needs to be evaluated and analysed using any modelling technique (Predictive modelling technique here) so as to finally drill down to the results that are nurtured through that data sets and data points so gathered.

- **Deploying the analytics got using a data visualization tool**
  Next step is to deploy the analytics that is been found out in more structured and designed way to visualize it and portray it in a better way by using various data visualization platforms.

- **Monitoring the data so that it gets updated frequently**
  Later the database is monitored so that the information updating can be made regularly so that the tool do not respond in the obsolete fashion.

3. Preparation and Normalization of Data

Dataset Construction and Cleansing
The data collected from various websites as discussed earlier yielded the scattered and unstructured data. The dataset was first constructed keeping in view majorly 6 key attributes as described below. This made the schema for our dataset which was then used as a sole deciding factor for further analysis as shown in Table 1

Table 1: Schema of Dataset used for the prediction

Movie Name	VARCHAR
Genre	VARCHAR
Cast Name 1	VARCHAR
Cast Name 2	VARCHAR
Director Name	VARCHAR
Sales Revenue	DOUBLE
Ratings	FLOATING POINT

But, the dataset so obtained had all the predictors in the character form as a result of which predictions weren't possible and hence there was a need for normalization of the dataset.

Dataset Normalization
For Normalization of the data we considered all the movies of the particular genre, actor 1, actor 2, director individually and averaged the sales revenue to actually come up with a particular value associated with each actor, director, genre and associate a price tag with every individual giving a sense of understanding of the valuation of every individual as per the performance of the movies the person has been associated with as shown in Table 2.[1]

Table 2: Methodology used for Normalization

Genre	∑(sales revenue of all movies of particular genre)/No. of movies
Cast Name 1	∑(sales revenue of all movies of particular casting star 1)/No. of movies
Cast Name 2	∑(sales revenue of all movies of particular casting star 2)/No. of movies
Director Name	∑(sales revenue of all movies of particular director name)/No. of movies
Sales Revenue	No change(already numerical)
Ratings	No Change(already numerical)

This finally helped in Normalization of Dataset to make the dataset fully numerical so that the algorithms could be implemented easily to yield the response from the predicted values.

4. Studies and Findings

Basically, we tried to use the Supervised Learning mechanism to device the basis for our main aim for the project i.e. to come up with an estimation of the pre-release sales revenue of a Bollywood movie in India.

Under Supervised Learning we basically used the approach of Regression to start with the approximate prediction of sales revenue for the movie. Various Algorithms were tested upon the dataset and the values so found out after training the dataset were then used to see the results on the testing dataset. We basically implemented 4 such algorithm approaches to come out with a comparative analysis as to which one performs better than the other.[1]

Linear Regression

Linear Regression algorithm basically is built on the fact that takes into account the linear independencies of various factors (predictors) as per the responses been recorded. The linear regression algorithm considers all the predictor and observes there relation independency to compute the predicted responses asper the significance of each parameter depending upon the training dataset as shown in Fig 1

$$Y=A_1*X_1+A_2*X_2+A_3*X_3+A_4*X_4$$

$A_i = i^{th}$ predictor

$X_i = i^{th}$ numerical coefficient

Y= sales revenue prediction

Figure 1: Linear regression mathematical formula

Non-Linear Regression

Non-Linear regression algorithm basically takes the interdependency between the predicates and generate the responses taking into consideration the fact that predicators may sometimes not always be directly co-related with each other and may show signs of interdependency.Here we divided the data set into training set and testing data set according to which we used the module function *=>hougen()*for computational purposes.The MATLAB function used for Non Linear multivariate regression was *"nlfit"*.From which the MSE was calculated and analyzed for further minimization as shown in Fig 2.

$$y = f(X, \beta) + \varepsilon$$

y :predicted value

$f(X,\beta)$: is the computing model function

(Hougen-Watson model)

$$\text{hougen}(b, x) = \frac{b(1)x(2) - x(3)/b(5)}{1 + b(2)x(1) + b(3)x(2) + b(4)x(3)}$$

ε: Noise

Figure 2: Non Linear Regression Mathematical Formula

5. Logistic Regression

Multivariate Logistic regression algorithm attempts to fits the curve in a more precise and efficient way.It classifies the problem with an upper bound (1) and a lower bound of 0(negative logarithms not possible) .The formulation used for getting the predictions on the testing dataset was as shown in Fig 3:

The Logistic Function

$$\ln\left[\frac{Y}{(1-Y)}\right] = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_iX_n$$

Y: Estimated probability

b_i: (i=0:n) Calculated factors

X_i: (i=0:n) redictors value

Figure 3: Logistic Regression mathematical formulation

The MATLAB function *"mnrfit"* was used to train the dataset and later the testing responses were noted to check for the Root Mean Squared Error. For evaluation of the value of predicted responses we used the function *"mnval"*. The difference between the predicted values to the actual values in the testing dataset defines the error.

6. Neural Networks in MATLAB

MATLAB also contains the statistical Neural Networks tool box that helps in easy estimation and curve fitting tools to construct a Simulink model first and then partition the dataset into 3 categories of 70%, 15% and 15% ratio of training dataset, validation dataset and testing dataset respectively. The dataset are first trained using the training dataset and then pruned using validation dataset and finally the observed error calculation is performed on the pruned using validation dataset and finally the observed error calculation is performed on the testing dataset. The algorithm gets its name from the neural function in our body and how do they function. The reaction released from one axon passes through several neurons simultaneously to finally reach the central controlling nervous system similarly in the data models too the data points are examined and the functions are run on all data points to device a pattern formation of any such kind.The neural network so created is shown in Fig 4 below in the form of a model as generated from Neural Network Toolbox.

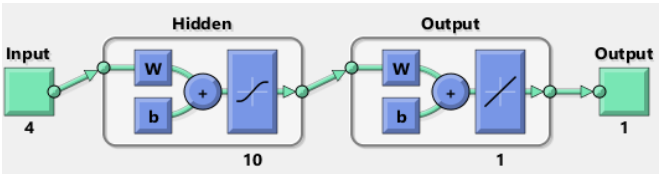

Figure 4: Function Fitting Model in Neural Network

7. Error and Analysis

On running various algorithms simultaneously, following analysis and insights could be withdrawn on the deviation patterns observed from the original values taken up from the internet and the predicted value so obtained.

Below, defined are the approaches taken up and the some data visualization is done from the results so obtained from the MATLAB codes. The visualization has been carried out by MS Excel Tool.

1) Approach 1: Linear Regression

The results obtained by Linear Regression algorithm implementation were highly accurate as per the relative algorithms which were been implemented. It was also

observed that the algorithm worked very well in providing an estimated value for medium and low sales revenue earning movies. The error value deviation in low and mediocre sales revenue movies were quite less and in sync with the actual responses so recorded.

Below shown is a line graph demonstrating the difference between the sales revenue of actual data and predicted data as shown in Fig.5

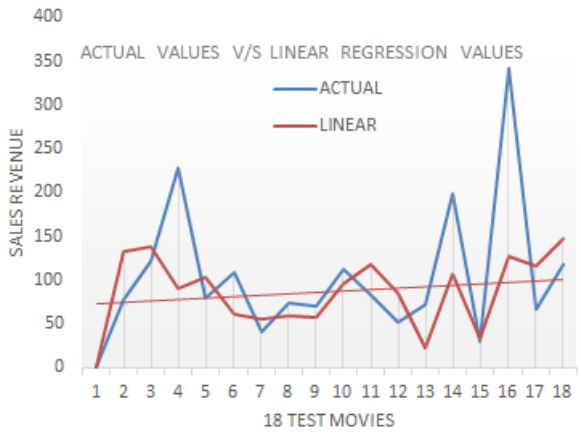

Figure 5: Line graph between Actual v/s Linear regression values

2) Approach 2: Non Linear Regression:

Non Linear Regression algorithm prediction was observed to be highly accurate with some values and deviating to large extend on some other values .Such high deviation introduced good amount of error while on other values the responses almost exactly matched to the predicted values .Such close approximations between the predicted value and actual value was observed on low and mediocre values. But major errors were been recorded on high end movies. [6]

Below shown is a line graph demonstrating the deviation between the sales revenue of actual data and predicted data. Please note the similarities marked by circle on low values as shown Fig. 6.

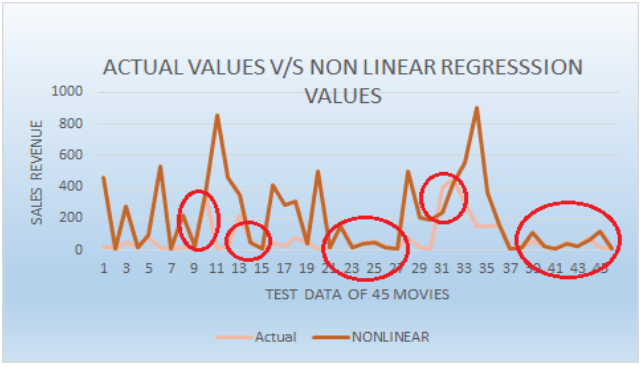

Figure 6: Line graph between Actual v/s Non Linear Regression values

3) Approach 3: Logistic Regression:

For implementation the problem was needed to be rethought altogether for analyses of the problem as a problem for Classification dividing the data into various buckets (>100,100-300,300-600 600-900 900<) where 1 indicated if it belonged to that bucket and 0 indicated otherwise. Finally the data was then associated with weights responding to each level of bucket and then the algorithm was implemented[6]

On implementation of Multivariate Logistic Regressionalgorithm it was observed that the algorithm did good predictions only for a limited range of values staying to be low. As a result it did not performed very well on high end values hence causing deviation and error in prediction

.

Below shown is a bar graph demonstrating the difference between the sales revenue of actual data and predicted data as shown in Fig. 7.

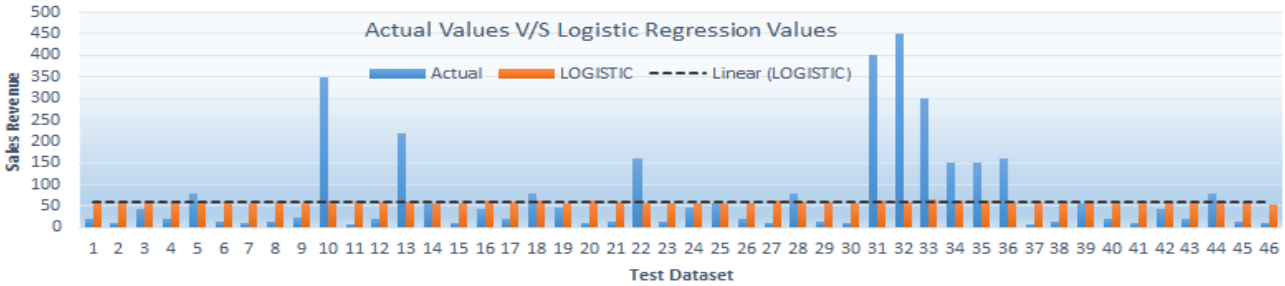

Figure 7: Line graph between Actual v/s Logistic regression values

4) Approach 4: Neural Networks Toolbox

To analyze the performance, we also used the Neural Network Toolbox readily available in MATLAB. The tool helps in importing the data in input and responses and then segregating the inputs into Training, Validation and Testing dataset and then pruning the dataset using the neural network algorithm. The plots so obtained are shown below.

Following Fig 8 shows the zero error and showing various error obtained at various instants.

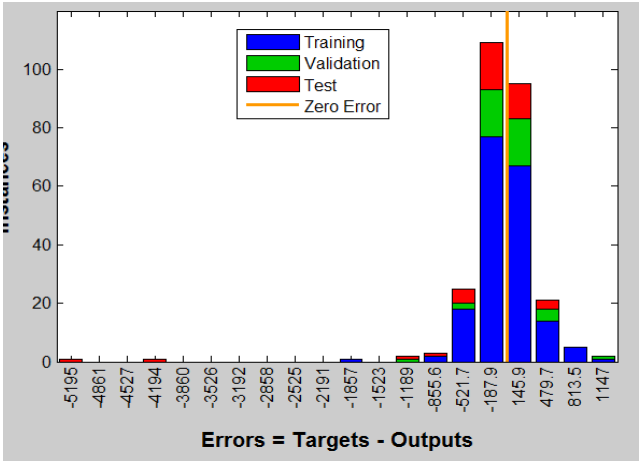

Figure 8: Bar Graph showing the error obtained in the run

The root mean squared value(MSE) was plotted against 12 epochs and the generated following shown graph was yielded as shown in Fig. 9 .As it is evident that there is a valid correlation between the 3 curves namely(Validation, Training and the testing) for some time. There is also a marked point when the validation and the training reaches a convergence point as shown in the graph below.

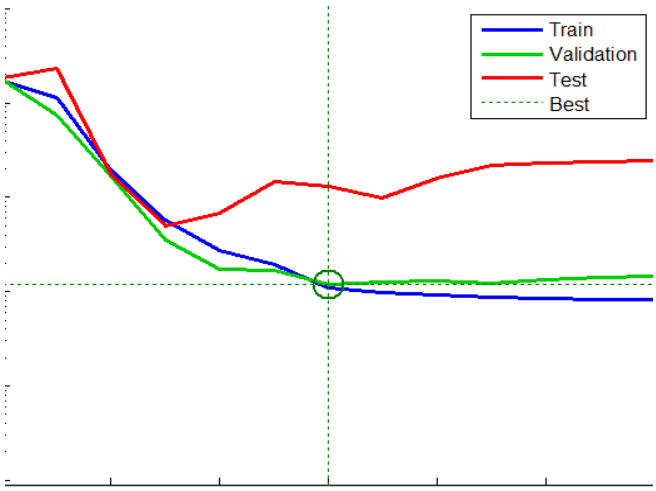

Figure 9: The performance analysis from Neural Networks in MATLAB

8. Conclusion

In this article, we have shown how past data can be utilized to forecast future outcomes.The analyses used here illustrate that the ultimate box office performance of movies can be forecasted with some accuracy given available information. Understanding the trend, it is decided whether if a new combination of the affixed attributes (e.g. actors, directors) are made to work with each other, what would be the possible collection of the movie at box office and possible rating to be given to that particular movie.

Other potentially important sources of revenue outside the theatre or any hall are not considered here. These also include home video, television broadcasting, and network television, all of which could be important source of revenue to movie producers.

9. Acknowledgment

This research paper is made possible through the help and support from everyone, including: teachers, family, friends, and in essence, all sentient beings. Especially, please allow me to dedicate my acknowledgment of gratitude toward the following significant advisors and contributors:

First and foremost, I would like to thank Harsh Taneja, Assistant Professor and Mentor, for his utmost support, valuable guidance and advice. He kindly read my paper and offered invaluable detailed advices on grammar, organization, and the theme of the paper. Second, I would like to thank my peer group to proof read my thesis and to provide valuable advices.The product of this .research paper would not have been possible without all of them.

References

[1] Arundeep Kaur and AP Gurpinder Kaur, Predicting Movie Success: Review of Existing Literature ,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013.
[2] Darin Im, Minh Thao, Dang Nguyen, "Predicting Movie Success in the U.S. market," Dept.Elect.Eng, Stanford Univ., California, December, 2011.
[3] Itedal Sabri Hashmi Bahia “A Data Mining Model by using ANN for Predicting Real Estate :Comaritive Study
[4] Ma’rton Mestyan,Tha Yassari,Jan’os Kertes’z:Early prediction of Box Office Success Based On Wikipedia Activity Big Data
[5] Philip Omentisch :Predicting Movie Sucesss with Machine Learning and Visual Analytics
[6] Nithin VR,Pranav M,Sarath Babu PB,Lijiya A: Predictng Movie Success Based on IMDB Data

Author Profile

Mr. Harsh Taneja, is Assistant Professor, CSE, Bharati Vidyapeeth College of Engineering

Anupam Dewan, B.Tech (CSE), Bharati Vidyapeeth College of Engineering

Vineet Bhardwaj, B.Tech (CSE), Bharati Vidyapeeth College of Engineering