MovieBuzz: Forecasting movie's success.

Sanjana Anaokar
Department of Information Technology
K. J. Somaiya College of Engineering
Mumbai, India

Ria Marwaha
Department of Information Technology
K. J. Somaiya College of Engineering
Mumbai, India

Pooja Karkera
Department of Information Technology
K. J. Somaiya College of Engineering
Mumbai, India

Chaitali Tanawade
Department of Information Technology
K. J. Somaiya College of Engineering
Mumbai, India

Abstract* - Bollywood has reached an amazing level in terms of movies produced, its reach in the whole world and providing employment to manpower. The returns obtained are uncertain in nature. Due to this it becomes a matter of interest to develop a model which can forecast the success of movies. In this paper, a model is proposed to forecast performance of Bollywood movies. The proposed work involves collecting data from various websites. Data mining algorithms like multi-linear regression and min-max normalization algorithm are used. The results been generated will help the movie industry as well as common people to take decisions regarding movies i.e. it will act as a decision support system.

***Keywords*— Forecast, Movies, Data, Net gross, Success, Regression, Websites, Box Office**

I. Introduction

In India many movies are released per year but not every movie makes a good business, there are many movies which are high budget but turn out to be flop. Movies act as a source of entertainment; it sometimes also creates awareness in people as well as provides knowledge. In addition to this movie industry provides employment to many people. Movie industry field has seen much technological advancement from black and white movies to color movies and now even variation in dimension 2D, 3D, etc. With advancement in today's technology for the production of motion picture, the movie industry produces about 200 movies per year. Hence a movie's success plays a critical role in determining the fate of the stakeholders associated with it; from directors to actors all are affected by it.

There are no forecasting websites available for Bollywood movies so far, all the available websites only provide data of hit movies top gross movies and all sort of historical data [2]. So our approach of providing forecasting results might help film makers and the audience for taking decisions regarding movies.

The purpose of our paper is to mine publicly available data of Bollywood movies and generate trends for forecasting movie's success. Our objective is to analyze the historical dataset containing records of released movies for forecasting movie's success. More specifically, we focus on forecasting the success of new movies based on the classical parameters associated with it namely- director, actor, actress, genre and the month in which the movie is going to be released. The forecast results will include the performance of movie, region where it will do good business, rating of movie, estimated revenue it will earn, and suggestions which may help to increase its overall performance.

II. Background and Related work

A. Literature survey

In the study conducted by Krushikanth et al [1] derived results by using normalization algorithm ,K-Means Clustering tool of Weka and Weka's J48 decision tree classifier to generate a predictive model.The contribution by this study was that sentiments and comments count used in their project were not identified relevant in their project.

Arundeep and Gurpinder Kaur [2] conducted a study in India,Jalandhar determined that, their study intended to develop model to forecast the success of movies.

B. Algorithms

1. Normalization – Normalization is done to scale the attribute data so that it fits in a specific range [4]. There are many types of normalization techniques available like Z-score, Min-Max, etc. A simple Min-Max algorithm is used to normalize the data since using it we could normalize our values in the range of 0 to 1.

$$Y = \frac{(X - Min\ value\ of\ X)}{(Max\ value\ of\ X - Min\ value\ of\ X)} * (B - A) + A$$

Y –the transformed normalized value.
X –the value which you want to normalize
If you want to convert the value in to range [0, 1] then A is 0 and B is 1.

2. Linear regression algorithm - Linear regression involves a response variable y and a single predictor variable x. For linear regression value of y will be approximately calculated as follows:

$$y = A * x + B$$

Here A is the slope and B is the intercept.
Many prediction projects use Linear regression algorithm for predicting results [3].But since our verdict is depended on many variables we used Multiple – linear regression algorithm to forecast.

Multiple-Linear regression algorithm - It is used when there is 1 response variable y and many predictor variables for e.g. x1, x2, x3, x4... Here the function will look like

$$y = A0 + A1xi1 + A2xi2 + \cdots + Anxin + \mathcal{E}$$

for i=1,2,…k.

where A0 is the intercept. A1, A2, …, An are partial regression coefficients and $\mathcal{E}$ is the random error [5].

III.Experimental Work

This section describes the experimental setup used for the project.
Data is collected from various websites and pre-processed. Pre-Processing is done using macro enabled Microsoft Excel workbooks which are coded in VBA. The data is then imported to SQL server 2012 in which the Film name acts as a common attribute to link different tables.
After importing the data in database the data was transformed in the required format for calculations. Multi-Linear regression technique is then used to forecast the success of movies based on following parameters:

- Genre.
- Release Month.
- Director.
- Actor.
- Actress.

The steps followed in our project are as follows:

Step 1:- Data collection and Pre-processing:
Data is collected from various websites like Wikipedia [1], Koimoi [7], Boxofficeindia [8] and Bollywood hungama [9]. The collected data is pre-processed in macro enabled Microsoft Excel using VBA coding. Movies details like Movies name, cast, genre, etc. are taken from Wikipedia [1][6]. Net gross of blockbuster movies is taken from Bollywood hungama[9] and region wise net gross of blockbuster movies is taken from Boxofficeindia[8] website. Data of 100 crore club movies, verdict of movies, etc. are taken from Koimoi[7] website.

Step 2:- Training Data Normalization:
The Net gross of the movies are normalized using the min-max algorithm as explained in the algorithms section [4]. The normalization is dependent on the input parameters entered by the user i.e. year-wise normalization of the net gross is done 5 times for the 5 inputs variables entered – Actor, Actress, Genre, Director and Release month of the movie to find out its contribution.

Step 3:- Application of Multiple-Linear regression:
Multiple-Linear regression algorithm is used as output is dependent on 5 input variables [5]. For our project we used the following formula to predict the verdict

$$\mathit{Projected\ Verdict} = (\mathit{Contribution\ of\ genre}) * \mathit{genre} + (\mathit{Contribution\ of\ release\ month}) * \mathit{release\ month} + (\mathit{Contribution\ of\ director}) * \mathit{director} + (\mathit{Contribution\ of\ actor}) * \mathit{actor} + (\mathit{Contribution\ of\ actress}) * \mathit{actress}$$

Step 4:- Generation of forecasting results:
Our project gives 5 forecasted outputs they are –
1. Movie performance in terms of blockbuster, super-hit, semi-hit and flop.
2. Movie rating prediction in the range of 0-10.
3 Region where movie will do good business
4. Estimated Box-office revenue
5. Suggestion to increase movies overall performance

Calculation of each output:-
Output 1:- Movie performance in terms of blockbuster, super-hit, semi-hit and flop.

1. The user enters the following parameters:
Genre, Release month, Actor, Actress and Director
2. Contribution of each parameter to calculate the verdict is as follows:
For e.g. Consider the parameter Genre entered as 'Drama'

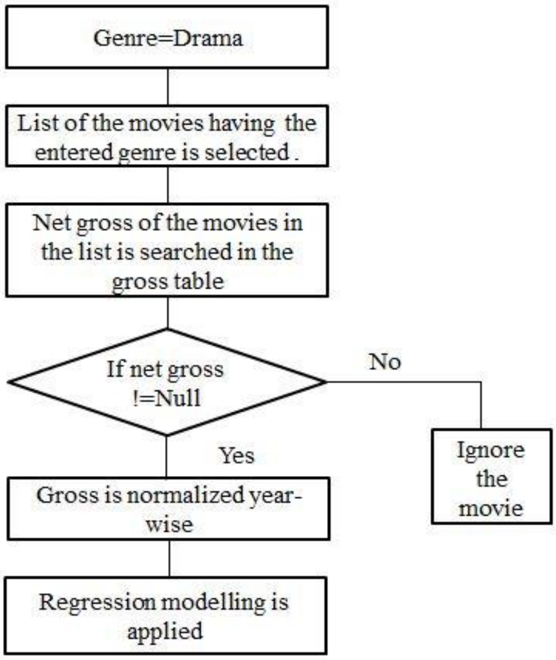

Genre=Drama
List of the movies having the entered genre is selected.
Net gross of the movies in the list is searched in the gross table
If net gross !=Null
No
Yes
Ignore the movie
Gross is normalized year-wise
Regression modelling is applied

Fig1. Flow of calculation

In the same manner the contribution of other parameters is calculated.

3. Projected verdict = Average of contribution of (genre, actor, actress, release month, director)

Calculated projected verdict	Verdict
<35	Flop
>=35 and <43	Semi-hit
>=43 and <65	Super-hit
>=65 and <=100	Block-buster

Table 1. Classification of Projected verdict

Output 2:- Movie rating.

$$\mathit{Movie's\ Rating} = \frac{\mathit{Projected\ verdict}}{10}$$

Output 3:- Regions where movie will do good business

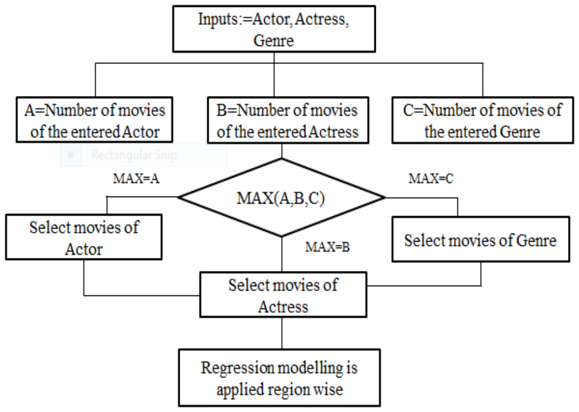

Inputs:=Actor, Actress, Genre
A=Number of movies of the entered Actor
B=Number of movies of the entered Actress
C=Number of movies of the entered Genre
MAX=A
MAX=C
MAX(A,B,C)
MAX=B
Select movies of Actor
Select movies of Genre
Select movies of Actress
Regression modelling is applied region wise

Fig 2. Region-wise calculation process

Output 4:- Estimated revenue prediction
It is calculated as follows

$$\mathit{Estimated\ net\ gross} = \frac{(\mathit{Projected\ Verdict}) * (\mathit{Maximum\ of\ parameter\ contribution})}{100}$$

Output 5:- Suggestions for movie’s release month. The releases month suggestion calculation is as follows:

- Actor, Actress and genre Parameter is considered.
- Net gross of movies consisting of actor or actress or genre is considered.
- Regression Modeling is applied which is grouped month-wise.

IV.RESULTS AND DISCUSSIONS

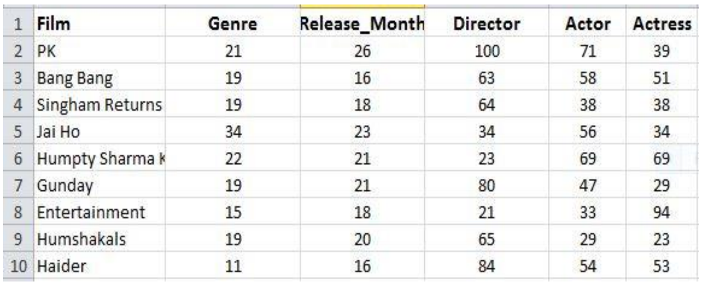
1. Verdicts related forecasting:

The system was tested against training data of movie verdicts. The training data was found to be approximately 86% accurate for the verdicts been projected.

2. Region related forecasting:

As per our study been done most of the movies do good business in the region of Mumbai.
So Mumbai is considered as the best region for earning gross.

3. Star Power: Actors, Actresses and Directors influence the success of movies more as compared to genre and the month in which the movie is going to be released. This gives a hint that the potential of directors and star cast forms the base in deciding the success of movie.



1	Film	Genre	Release_Month	Director	Actor	Actress
2	PK	21	26	100	71	39
3	Bang Bang	19	16	63	58	51
4	Singham Returns	19	18	64	38	38
5	Jai Ho	34	23	34	56	34
6	Humpty Sharma k	22	21	23	69	69
7	Gunday	19	21	80	47	29
8	Entertainment	15	18	21	33	94
9	Humshakals	19	20	65	29	23
10	Haider	11	16	84	54	53

Fig3: Contribution of each parameter

V. CONCLUSION

Our Project forecast the success of movies based on input parameters like Genre, Release month, Director, Actor, and Actress of the movie. Our model forecasts the following results: (1) Movie Verdict. (2) Movie Rating. (3) Suggestion for release month. (4) Box-office revenue estimate of the movie. (5) Region where the movie will go good business. The availability of the numerical data is low and hence mining the data was difficult. Various success parameters of movies are not available. Despite of these problems we performed some meaningful forecasting on the available data and uncovered some useful information related to success of movies. The project will be beneficial to producers and owners of film production studios in order to plan their future releases as well as audience who can decide which movie to watch depending upon its success in Box-office.

ACKNOWLEDGEMENT

We thank Prof. Deepti Patole. Without her guidance, this paper could never have been accomplished.

REFERENCES

[1] 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”, ACM International Conference on Advances in Social Networks Analysis and Mining,2013 IEEE

[2] Arundeep Kaur, 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.

[3]Mahesh Joshi, Dipanjan Das, Kevin Gimpel and Noah A. Smith.” Movie Reviews and Revenues: An Experiment in Text Regression” NAACL-

HLT, 2010

[4]http://intelligencemining.blogspot.in/2009/07/data-preprocessing-normalization.html

[5] http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm

[6] https://en.wikipedia.org/wiki/Lists_of_Bollywood_films

[7] http://www.koimoi.com

[8] http://www.boxofficeindia.com

[9] http://www.bollywoodhungama.com/box-office/top-grossers