Prediction of Movie Success through the Data Mining

Merim Babu[1], B. Muni Archana[2]
[1]*Student, Master of Computer Applications, SKIIMS, Srikalahasti, Andhra Pradesh India*
[2]*Asst.Professsor, Master of Computer Applications, SKIIMS, Srikalahasti, India*

***Abstract*- Predicting a movie's opening success is a difficult problem, since it does not always depend on its quality only. External factors such as competing movies, time of the year and even weather influence the success as these factors impact the Box-office sales for the moving opening. The topic of movies is of considerable interest in the social media user community. Posting online reviews is a new trend set for people to share with other users their opinions and sentiments toward products and services. E-commerce websites provides the venues and facilities for people to publish their reviews. Those online reviews present a wealth of information. In this project, reviews of viewers from movie fertility are collected and processed. When the movie trailer releases various reviews are posted by users on social media sites. We are mining those reviews and predicting performance of the movie. These predictions will be used by shareholders and box office for the movie business. We label the prediction in three classes, Hit, Neutral and Flop.**

***Index Terms*- YouTube, Online reviews, Opinions, Data Mining, Prediction, Regression Model.**

I. INTRODUCTION

In this era, we developed a mathematical model to predict the success and failure of the upcoming movie based on several attributes. Some of the criteria in calculating movie success included budget, actors, director, and producer, set locations, story writer, and movie release day. Competing movie releases at the same time, music, release location and target audience. Predicting the outcome of events and the success of products is a fundamental problem in data mining and predictive analytics. A variety of techniques have been proposed to address real world prediction problems arising in different domains. In this work, I address the problem of predicting movie success based on two indicators:

• **Box-office income**: the gross revenue of the movie combined for all theatres showing the movie on the opening weekend

• **Average Rating**: the rating users provided on the Internet Movie database after the opening weekend.

Many models have been proposed in order to predict viewer ratings and box office incomes, for example with the help of social media [Oghina et al., 2012] or news analysis [Zhang and Skiena, 2009]. The success of a movie can be measured by many different aspects. The main criteria tough are quality, how the audience liked the movie and box office, as in economical success. Movie success prediction has a lot of use for companies to plan their resources. For example,

In this era social media is becoming more popular where netizens can express themselves, gives reviews etc. data generated through social media is nearly 10TB per day. With increase in such large amount of data it is necessary to develop a system which will make use of such large amount of data to perform analysis and predict future with social networking. So we are developing a system which makes use of twitter data for predicting box office collection of movie. Opinion mining or Sentiment analysis refers to a broad area of Natural Language Processing and text mining. It is concern not with the topic a document is about but with opinion it expresses that is the aim is to determine the attitude (feeling, emotion and subjectivities) of a speaker or writer with respect to some topic to determine opinion polarity.

II. RELATED WORK

Another study [1] K-Means clustering, Polynomial and Linear Regression was applied on 2510 movies released1990 onwards to study and build a predictive model to get the expected revenue. They achieved accuracy of 36.9%.

Another study [2] applied Text regression on critics' film reviews to predict the opening weekend revenue for the metadata collected for 2005-2009 movies. The dataset consisted of 1718 movies. The authors used seven meta data features including Movie Running Time (in minutes), Budget, the number of opening weekend screens, genre, MPAA rating, opening time (whether summer or holiday), total number of actors, high grossing actors

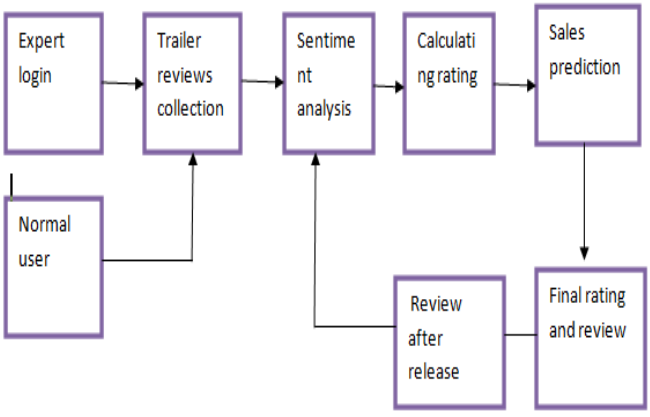
count and whether the movie had any Oscar winning actors and directors.

Another study [3] was conducted over a span of five years (1998- 2002) in which the authors classified nine classes from flop to blockbuster. They applied neural network algorithm on 7independent variables and found that number of screens, high technical effects and high star value contribute a great deal to a movie's success.

Another study [4] researcher proposed the idea to integrate classical and social media factors to improve the prediction accuracy of the movie success. They collected classical attributes (genre, budget etc.) from IMDB and social attributes (Tweets, views) rom social websites like YouTube, Twitter. The study suggests that by increasing the data set, a higher accuracy than the one obtained (70%) through linear regression, can be achieved.

III. PROPOSED SYSTEM

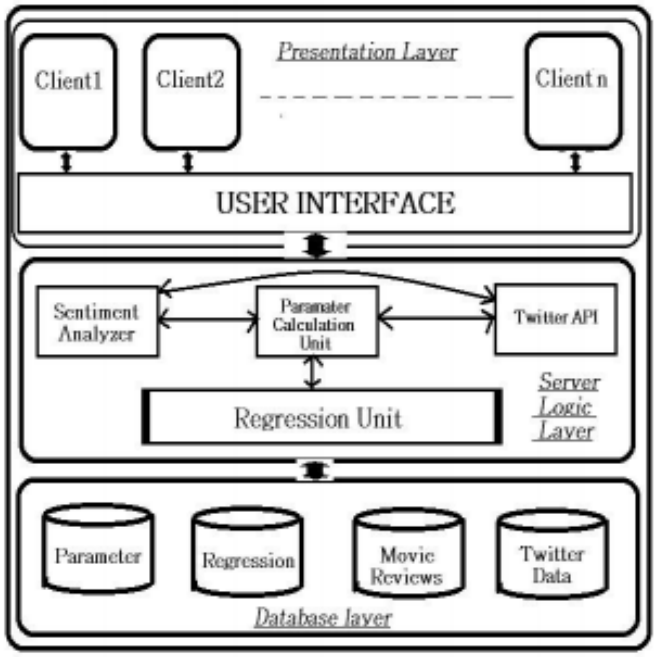
The aim of proposed system is to develop a system of improved facilities. The proposed system can overcome all the limitations of the existing system. The systems provide proper security and reduce the manual work and easily maintain the movie information. The existing system has several disadvantages and many more difficulties to work well. The proposed system tries to eliminate or reduce these difficulties up to some extent. The proposed system will help the user to reduce the workload and mental conflict. The proposed system helps the user to work user friendly and he can easily do his jobs without time lagging.



Data flow of proposed system

System Architecture:-

The given bellow daigram is system architecture for movie success prediction. In this we have 3 layers one is presentation layer and second one is regression layer and database layer. In presentation layer users can see movie trailers and review collections and relesing dates for movies in this layer this is called user interface.second layer is server side logic we can perform sentiment analyzer rating calucations this type of working process perform in server side. In database layer can maitaine the details of the regression, movie reviews. twitter data.



IV. DATA COLLECTION AND PROCESSING

A. Dataset Collection

The initial dataset to be used will be collected. It will consist of movies that were released from 2000 to 2012.for making more accurate predictions; we refined the movie list by removing movies those are not released in the America. Out of those movies we selected movies are in English, in the expectation that it will form a dataset for precise forecast. The most complete historic movie data, that is freely available. It includes over 2.5 million titles and 5 million people and their roles in movies, as well as release dates and ratings. In order to import data from the plain text files, I extended software written by Kosara et al.

B. 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. we mainly used Rotten Tomatoes and Wikipedia. The main problem with datasets was missing filed problem we adopted a method which uses the central tendency for the attribute.

C. Data Analysis

Data analysis is a process of inspecting cleansing, transformation, and modeling data. In movie prediction analysis of data is very important. For visualization I used the software Tableau [Tableau Software, 2014] also via MySQL connection. Visualizing the data in various ways helped me identify important features and correlations, how exactly I will show later. Tableau is a very powerful tool, because it can handle huge data sets (the set I worked with included approximately 35000 movies) and display them in many different ways.

D. Data Integration and Transformation

Data obtained from three different resources Wikipedia where than integrated into one database. In this step the integrated data are transformed or consolidated so that the regression process may be more efficient and easier. Database mixed with both nominal and numeric attributes.

But regression process, we need all attributes to be numerical. We use the measure of central tendency of movie reviews to convert nominal attributes to numerical.

Type	Features
Nominal	Actors, Director, Producer
Numerical	Budget, Rating, Status, Reviews, User Rating.

V. REGRESSION MODEL

Regression model deals with estimate of an output value base on input values. When used for classification, the input values are values from the database and output values represent the classes. Regression can be used to solve the classification problem. Liner regression formula

$$Y=c_0+c_1x_1+\ldots\ldots+c_nx_n$$

Most researchers use simple methods such as linear regression analysis. These methods are known to work well under some conditions. Social media is produced on a complex system and thus more likely than not the predictors and prediction outcomes have non-linear correlation. Furthermore, combination of methods might lead to breakthrough. In such combination, a surface learning agent, such as instantaneously trained neural networks, quickly adapts to new modes and emerging trends on social media. And a deep learning agent focuses on long-term patterns. In a nutshell, we should try some non-linear methods and find out the suitable methods and/or combinations for each prediction realms. The data can then be used to fit a linear regression model using least squares. The parameters of the model include:

- A: rate of attention seeking
- P: polarity of sentiments and reviews
- D: distribution parameter

Let y denote the revenue to be predicted and ǫ the error. The linear regression model can be expressed as:

$$y = \beta a * A + \beta p * P + \beta d * D + \epsilon \quad (4)$$

Where the β values correspond to the regression coefficients. The attention parameter captures the buzz around the product in social media.

5.1 Viewer Rating

Rating (1-10):

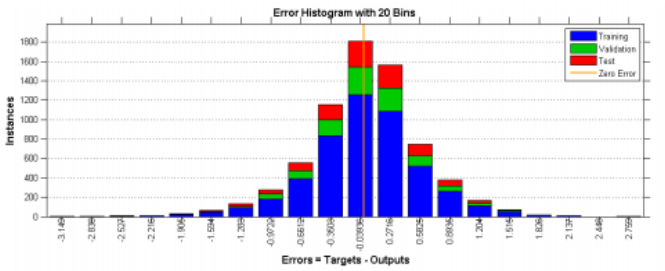
The rating of a movie determined by users**,** this is the target.

Number of ratings:

This shows two things, first, how statistically trustworthy the rating is, second, roughly how many people watched the movie. After visualizing the number of ratings and the ratings see figure 3.1, it became clear, that the number of ratings n correlates with the rating, even more for movies with ($n \geq 50{,}000$). This is also the reason why I took the log10 of the number of ratings, because only the order of magnitude is important. I can only speculate why they correlate, but my first guess would be, that film studios tend to know quite well how to satisfy their target customers.

Average actor rating:

The weighting works the same way as with directors, but I chose the first 5 actors in the billboard listing of a movie, most of the time these also are the main characters of the movie.



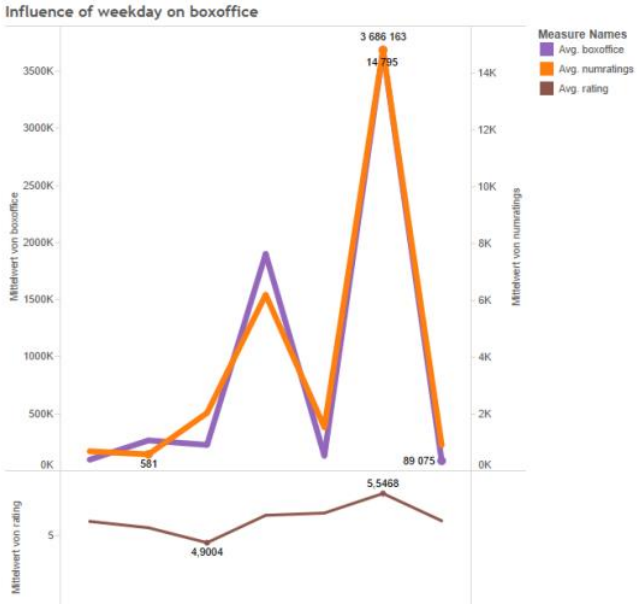
The error histogram for the trained neural network to predict viewer rating

5.2 Box office

Almost always the turnover resulting from ticket sales at the theatres is referred to as box office. The box office has a lot of external dependencies and especially on the opening weekend tends to be more important than the quality of the movie itself. Popular directors and actors tend to draw people to go to the movies they participate in, but there is no easy way to see popularity of i.e. an actor from the provided data. Therefore, I mainly focused on the external dependencies, them being the release date, budget, number of theatres the movie is shown on the opening weekend, social media trends and even the weather. The number of theatres is important due to the fact, that it presents an upper boundary for the possible box office earnings. Due to the restrictions of the challenge, participants were only allowed to use data, unfortunately they do not provide budget information and the number of theatres a movie is shown in. This shifted the focus of box office prediction to the analysis of social media data and the release date, also taking the genres into consideration.

Release date

The release date plays a major role for the box office and the success of a movie. Major studios plan their movies and also the release dates for them years in advance. Differences between release dates can have an impact as big as up to a factor of 3 as the analysis of historical data shows, this can be seen in figure 3.5. This figure is very meaningful for it shows, 3 important things. First, the release date has no influence on the movie's rating.



Ratings of weekdays on box office

VI. CONCLUSION

This article helps to find out the review of the new movie. User can easily decide whether to book ticket in advance or not. Visitor can easily get the information about the movie and industry. Reduce the man power, and also reduce the time. This application helps to find out the review of the new movie. The model I used to predict movie success is quite simple, but still powerful enough to make good predictions. Compared with other proposed methods. In this article, we have shown how social media can be utilized to forecast future outcomes. We also analyzed the sentiments present in tweets and demonstrated their efficacy at improving predictions after a movie has released.

REFERENCES

1. Sharda, R., &Delen, D. (2006): "Predicting box-office success of motion pictures with neural networks". Expert Systems with Applications, 30(2), 243-254.277

2. Nikhil Apte, Mats Forssell, and A. Sidhwa, "Predicting Movie Revenue". 2011.

3. Joshi, M., Das, D., Gimpel, K., & Smith, N. A. (2010). "Movie reviews and revenues: An experiment in text regression". In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

4. Demuth, H. and Beale, M. (1993). Neural network toolbox for use with matlab.

5. Assady et al., 2013] El Assady, M., Hafner, D., Hund, M., Jäger, A., Jenner, W., Rohrdantz, C., Fischer, F., Simon, S., Schreck, T., and Keim, D. A. (2013). Visual analytics for the prediction of movie rating and box office performance. In IEEE Int. Conf. on Visual Analytics Science and Technology (VAST Challenge Paper).