International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 04 | April - 2021 ISSN: 2582-3930

MOVIE SUCCESS PREDICTION USING DATA MINING

KABINA P
Asst. Prof. Mr. R. SATHISHKUMAR
Krishnasamy College of engineering and technology, Cuddalore, Tamil Nadu, India.

ABSTRACT - The success prediction of a movie plays a vital role in movie industry because it involves huge investments. However, success cannot be predicted based on a particular attribute. So, proposed system has built a model based on interesting relation between attributes. The movie industry can use this model to modify the movie criteria for obtaining likelihood of blockbusters. Also, this model can be used by movie watchers in determining a blockbuster before purchasing a ticket. Each of the criteria involved was given a weight and then the prediction was made based on these. For example, if a movies budget was below 5 million, the budget was given a lower weight. Depending on the number of actors, directors and producers past successful movies, each of these categories was given a weight. If the movie was to be released on a weekend, it was given higher weight because the chances of success were greater. If with the release of a movie, there was another high success movie released, a lower weight was given to the release time indicating that the chances of movie success were low due to the competition. The criteria were not limited just to the ones mentioned. There was being additional factors discussed in this work. The work was conducted with simulation data.

Key Words: blockbusters, Data mining, success prediction, rating, decision tree algorithm

1. INTRODUCTION

Data mining process was used to extract patterns and trends which can be beneficial in predicting movies success. The data mining techniques were applied to a movie database, but before the mining techniques could be used, the data went through the cleaning and integration process. Data mining deals with discovering trends and patterns in a given data. Data mining approach is important since it can help to identify the hidden patterns and relationships among various variables. These relationships can in turn help in identifying sequence of events, classification, clustering, and predicting future events. Data mining techniques could be used in countless scenarios. Some examples are profit prediction, investment decision, weather forecast, simulations, visualization tools, and medicinal purposes. Due to the powerful data mining techniques and predictions, this approach was used for movie success prediction. Movie success prediction is important because it involved significant time and investment. For this reason, it is important for the shareholders to have less uncertainty involved. They can achieve this very well using data mining techniques. Movie success predictions, trends and variable dependence can very well be determined using data mining. Movie success prediction is also significant for the movie watchers who need to know in advance the quality and success rating of a movie before monetary resources can be utilized for a movie. If data mining modeling is not used to predict an outcome, uncertainty increases and success confidence is lowered. This is particularly risky for stakeholders who have invested their significant resources. It is important that there is an outcome prediction and confidence before an important investment is made which is achieved by using the data mining techniques.

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. 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. That also found, 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. Other important factors are the director and actors/actresses involved in a film.

Extensive description and data is available for Hollywood movies (IMDB, Rotten Tomatoes etc.). The IMDB site predominantly contains data for Hollywood movies. A structured database or a central repository of Indian movie data is difficult to find. Four billion tickets for Bollywood films are sold annually. It affects producers, distributors, actors, rentals agencies and Bollywood fans. Indian movie industry produces the maximum number of movies per year. However, very few movies taste success. The stakeholders concerned depend on film critics for movie reviews. However, the movies are reviewed after their release.

IMDB rates movies according to true Bayesian estimate.

$$WR = (v \div (v+m)) \times R + (m \div (v+m)) \times C \quad (1)$$

where:

R = Mean Rating

v = votes for the movie

m = minimum votes required to be listed in the Top 250 (currently 25000)

C = the mean vote across the whole report (currently 7.0)

Indian Hindi Cinema industry popularly known as Bollywood has reached staggering proportions in terms of volume of business (184.3 billion), manpower employment (over 6 million workers), movies produced (more than 100 in a year) and its reach (exported to more than 100 countries worldwide). With so much at stake and highly uncertain nature of returns, it is of commercial interest to develop a model which can predict success of a movie. This however, is not an easy work, since movies have been described as experience goods with very less shelf life; it is difficult to forecast demand for a movie. There are number of parameters that may influence success of a movie like – time of its release, marketing gimmicks, lead actor, lead actress, director, producer, genre, music director – being some of the factors.

2. LITRATURE VIEW

1. "Circle-based recommendation in online social networks", written by X. Yang, H. Steck, and Y. Liu. Online social network information promises to increase recommendation accuracy beyond the capabilities of purely rating/feedback-driven recommender systems (RS). As to better serve users' activities across different domains, many online social networks now support a new feature of "Friends Circles", which refines the domain-oblivious "Friends" concept.

2. "A matrix factorization technique with trust propagation for recommendation in social networks", whose author is M. Jamali and M. Ester. Recommender systems are becoming tools of choice to select the online information relevant to a given user. Collaborative filtering is the most popular approach to building recommender systems and has been successfully employed in many applications.

3. "Achieving Efficient Cloud Search Services: Multi-Keyword Ranked Search over Encrypted Cloud Data Supporting Parallel Computing", written by Z. Fu, X. Sun, Q. Liu. A large number of data are outsourced to the cloud by data owners motivated to access the large-scale computing resources and economic savings. To protect data privacy, the sensitive data should be encrypted by the data owner before outsourcing, which makes the traditional and efficient plaintext keyword search technique useless.

3. PROPOSED SYSTEM

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).

The methodology has 4 major components, these are *1. Data Collection2. Data Cleaning 3. Data Transfer 4. Data Analysis and Prediction*. Cleaning the dataset and discarding the irrelevant data from the IMDB dataset as well as through detailed study of the dataset, we get the attributes that can affect the prediction of success of a movie. The textual data is transferred to numeric data and converts it into CSV format. Different decision tree algorithms are then studied so that the best suited algorithm according to our problem could be determined. The best suited algorithm should be the one which gives the most accuracy and least error. The test data when tested according to the algorithm should give as accurate results as possible.

Advantages

- Optimum utilization of system is possible. All basic functionalities are provided.
- The wastage of time is reduced.
- More user friendly environment
- More flexible, it means we can continue to use the same system even the business extends up to maximum level.

4. MODULES

Data Collection: The raw IMDb dataset is structured in such a way that most of its attributes and information is organized and stored separately in compressed plain text files. For instance, all of the roughly 600,000 movie ratings from the database are stored in the compressed text file ratings. List (e.g. ratings.list.gz), which includes textual information about the data as well as a table of film rank, the number of votes and film titles. Thus, some sort of cleaning, integration and preprocessing is likely to be required in order to make good use of the data for the purpose of data mining through supervised machine learning techniques. The data was collected using IMDB (java movie database) which contains the IMDB movie dataset of more than 30,00,000 movies in the dataset. The dataset was transferred to MySQL, in form of tables.

Data Cleaning: Several SOL queries were run on the relational database to clean the data in order to reduce the

data and select only the relevant attributes which would help in data analysis and prediction of movie success.

Data Transfer: The relational database table data was exported to excel files and stored in the CSV (comma separated values) format for further analysis.

Data Analysis and Prediction: The dataset is divided into training dataset and test dataset which contains the classes like Hit, Flop and Average and predicting variables like actor, actress, composer, genre, director producer and music director k-means clustering is used to analyze the training dataset to develop models which can be used for test dataset for analysis decision tree algorithm is used for predicting which factors.

Admin Module: The Administrator is maintain the user Details, Movie details, Theatre details.

User Module: The user can first Registration in enter the Personal details, and User login and If you want to update personal Details and you and update. The user collect all information like Movies details, Theatre Details.

Ticket Booking Module: The Users will search for the movie and then go for theatre then booking the tickets online.

5. CONCLUSION

To apply data mining technique to the data in the IMDb dataset. It requires proper cleaning and integration, and this consumed a large proportion of the time available for this analysis. In addition, much of the data is in textual rather than numerical format, making mining more difficult. The source data could not be integrated easily. By using natural language processing techniques the data can be integrated properly. For overcoming these problems, we performed some useful data mining technique on the IMDb data, and uncovered information that cannot be seen by browsing the regular web front-end to the database.

6. REFERENCE

[1] Jiawei Han, Jian Pei, and Micheline Kamber. ”Data Mining Concepts and Techniques”, 2012.

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

[3] Ramesh Sharda and Dursun Delen. ”Predicting boxoffice success of motion pictures with neural networks”. Expert Systems with Applications, vol 30, pp 243-254, 2006.

[4] W. Zhang and S. Skiena. ”Improving movie gross prediction through news analysis”. In Web Intelligence, pages 301-304, 2009.

[5] Sitaram Asur and Bernardo A. Huberman, ”Predicting the Future with Social Media,” http://arxiv.org/abs/1003.5699, March 2010.

[6] Michael T. Lash and Kang Zhao, ”Early Predictions of Movie Success: the Who, What, and When of Profitability”, June 2015.

[7] Cohen, J., Cohen P., West, S.G., and Aiken,L.S. ”Applied multiple regression/correlation analysis for the behavioral sciences”, 2003.

[8] Christopher M. Bishop. ”Pattern Recognition and Machine Learning. Springer”, 2006.

[9] Cristianini, Nello and Shawe-Taylor,John. ”An Introduction to Support Vector Machines and other kernel-based learningmethods”, 2000.