# Movie Success Prediction Using Data Mining

Javaria Ahmad*, Prakash Duraisamy*, Amr Yousef†, Bill Buckles‡
*University of Central Missouri, Lees Summit, Missouri, USA
†University of Business and Technology, Dahban, KSA
‡University of North Texas, Denton, Texas, USA

***Abstract*—In this project, we developed a mathematical model to predict the success and failure of the upcoming movies based on several attributes. Some of the criteria in calculating movie success included budget, actors, director, producer, set locations, story writer, movie release day, competing movie releases at the same time, music, release location and target audience. 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, we have 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 were be additional factors discussed in this work. We have conducted our work with simulation data.**

**Keywords: movie success, data mining, movies, attributes.**

## I. INTRODUCTION

For this work, 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 [1]. 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 uncertainties 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.

We have provided a useful model in this study which can lower chance of failure and can provide the stakeholders with confidence and a visible prediction of success. There are various variables which were studied to provide a movie success prediction. Some of these variables included budget, actors, director, producer, set locations, story writer, movie release day, competing movie releases at the same time, music, release location and target audience. The goal of this mathematical model is to provide a precise prediction of success, hence providing confidence to stakeholders in their investments.

## II. RELATED WORKS

In 2004, Saraee, White and Eccleston performed analysis of online movie resource of over 390,000 movies and television shows [2]. In 2006, Sharda and Delen worked with predicting financial success of movies even before the movie is released [3]. Classification approach is used where the movies were categorized from flop to blockbuster. Facts and relationships among alternatives can be made by making use of data mining. Some of the factors considered were movie budget and movie popularity relationship, movie cast and movie success relationship. This work helped discover important findings. However, due to copy right, there was a challenge involved accessing the data. In 2009, Zhang and Skeina worked on utilizing news analysis to make movie predictions [4]. It was determined that using news data resulted in performance as good as using the IMDB data. Even better performance was achieved using both IMBD data and news data. In 2010, Asur and Huberman worked on predicting outcomes based on social media content [5]. The movie success prediction was based on social media success count, and historical data. The predictions can be made about new movies using this study. However, success prediction cannot be made before the movie is released. In 2015, Lash and Zhao proposed a way to predict decisions about movie investments [6]. This work provided help with investment decision making early in movie production. Historical data was utilized for this work. Some of the features of this work were matching ”who” with ”what” and ”when” with ”what”. The profit was calculated mainly based on box office revenue. However, for many movies, there are other sources of revenue, for example, merchandise.

### A. *Contribution of Our Work*

In our work, we have developed a mathematical model which is used to predict the success and failure of upcoming movies depending on certain criteria. Our work provides advantage in that strong correlations were found between different criteria and movie success rating. Unlike the related work discussed, our work can be used to predict movie success even before it is released. Our work makes use of historical data in order to successfully predict the ratings of movies to be released.

## III. Proposed Model

### A. *Algorithm for Movie Success Prediction*

1) Clean, integrate and transform simulation data.
2) Find $X^2$ analysis between genres and ratings of the movies.
3) Find $X^2$ analysis between movie actors and movie ratings.
4) Find $X^2$ analysis between movie actors and movie genres.
5) Find the correlations from the respective $X^2$ analyses above.
6) Predict success rating from the correlations between various movie criteria.

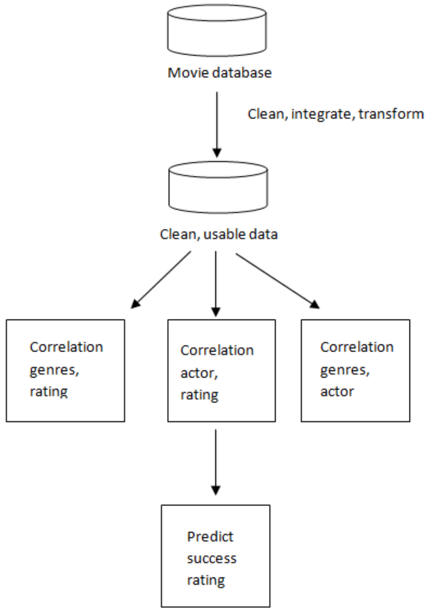
Movie database
Clean, integrate, transform
Clean, usable data
Correlation genres, rating
Correlation actor, rating
Correlation genres, actor
Predict success rating

Fig 1: Data Model

## IV. Evaluation Metrics

Simulation data was used for this analysis and hundreds of records were cleaned, integrated and transformed. A random subset of this data was utilized for each set of analysis.

## V. Data Preparation

Simulation data was used for the movie database. The data was cleaned, integrated, and transformed before the data mining techniques were applied. The data was analyzed based on the following attributes.

- Movie Name
- Year of release
- Genres Drama, Action, Romance, Comedy, Other
- Directors
- Music directors
- Producers
- Languages Hindi

## VI. Mathematical Model

In this study, the mathematical model developed to predict the success and failure of the upcoming movies involved finding correlation between various attributes using $X^2$ analysis.

Correlation is a measure of dependence between two variables. The correlation can be negative or positive. A positive correlation indicated that the two variables increase or decrease in parallel, whereas, a negative correlation indicated that the two variables change in opposite directions.

### A. *$X^2$ analysis: Genres vs. Ratings:*

Correlation between genres and ratings was analyzed first. The $X^2$ results are shown in the table below.

TABLE I
$X^2$ Analysis: Genres vs. Ratings:

| Ratings/Genres | Romance | Comedy | Other | Total |
|---|---|---|---|---|
| 1 | 2(9.8) | 0(4.8) | 16(3.4) | 18 |
| 2 | 0(0) | 0(0) | 0(0) | 0 |
| 3 | 3(10.8) | 16(5.33) | 1(3.7) | 20 |
| 4 | 19(12.5) | 4(6.13) | 0(4.34) | 23 |
| 5 | 25(15.8) | 4(7.7) | 0(5.47) | 29 |
| Total | 49 | 24 | 17 | 90 |

Expected frequencies are calculated as:
$Count(Genres)xCount(Ratings)/n$
$X^2 = 6.2 + 5.6 + 3.38 + 5.35 + 4.8 + 21.3 + 0.74 + 1.81 + 3.4 + 2.1 + 4.34 + 5.47 = 64.39.$
Degrees of freedom $= (4)(2) = 8$.

From observing the chi-square table, the p valve is very low, so we can reject the null hypothesis that ratings and genres are independent and conclude that the two attributes are strongly correlated. This means that movie genres are predicted to have specific ratings.

As shown by the bar graph below, most of the romance movies have the rating of 5, most of the action movies have rating 4, and most of the social movies have rating 3.

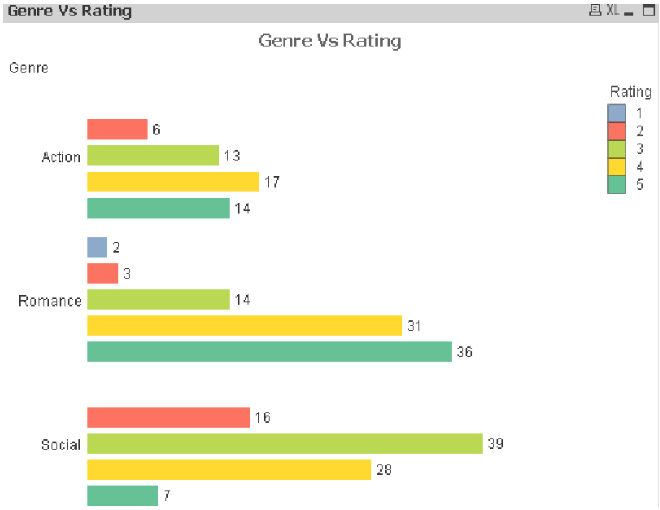

Fig 2: Genre vs. Rating

### B. $X^2$ analysis: Actors vs. Ratings:

Correlation between actors and ratings was analyzed next. The $X^2$ results are shown in the table below.

TABLE II
$X^2$ ANALYSIS: ACTORS VS. RATINGS:

| Ratings/Actors | Shahrukh Khan | Bobby Deol | Total |
|---|---|---|---|
| $\leq 3$ | 1(4.2) | 5(1.8) | 6 |
| $>3$ | 13(9.8) | 1(4.2) | 14 |
| Total | 14 | 6 | 20 |

Expected frequencies are calculated as:

$$Count(Genres)xCount(Ratings)/n$$

$$X^2 = 2.4 + 1.04 + 5.7 + 2.43 = 11.57.$$

Degrees of freedom $= (2-1)(2-1) = 1$.

The computed value is 10.828 ($X^2$ value needed to reject the hypothesis), so the hypothesis is rejected that ratings and actors are independent. Hence the two attributes are strongly correlated.

This means that actors can be predicted to increase the rating values of their movies. As shown by the bar graph below, most of the Salman Khans movies have a high rating of 4.

As shown by the bar graph below, most of the romance movies have the rating of 5, most of the action movies have rating 4, and most of the social movies have rating 3.

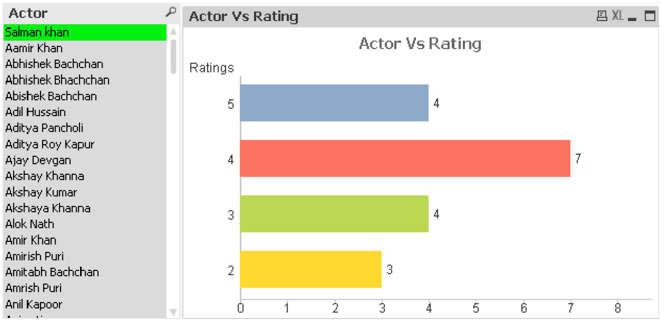

Fig 3: Actor vs. Rating

### C. $X^2$ analysis: Actors vs. Genres:

Correlation between actors and genres as it is shown in the table below.

TABLE III
$X^2$ ANALYSIS: ACTORS VS. GENRES:

| Genres/Actors | Akshay Kumar | Hrithik Roshan | Total |
|---|---|---|---|
| Romance | 0(2) | 6(1.77) | 6 |
| Others | 9(3.6) | 2(3.25) | 11 |
| Total | 9 | 8 | 17 |

$$X^2 = 2 + 10.10 + 8.1 + 0.48 = 20.6.$$

Degrees of freedom $= (2-1)(2-1) = 1$.

The computed value is above 10.828 ($X^2$ value needed to reject the hypothesis), so the hypothesis is rejected that genres and actors are independent. Hence the two attributes are strongly correlated.

This means that actors can be predicted to work in certain genres most of the time. As shown by the bar graph below, most of Akshay Kumars movies are action, and most of the Salman Khan's movies are romance.

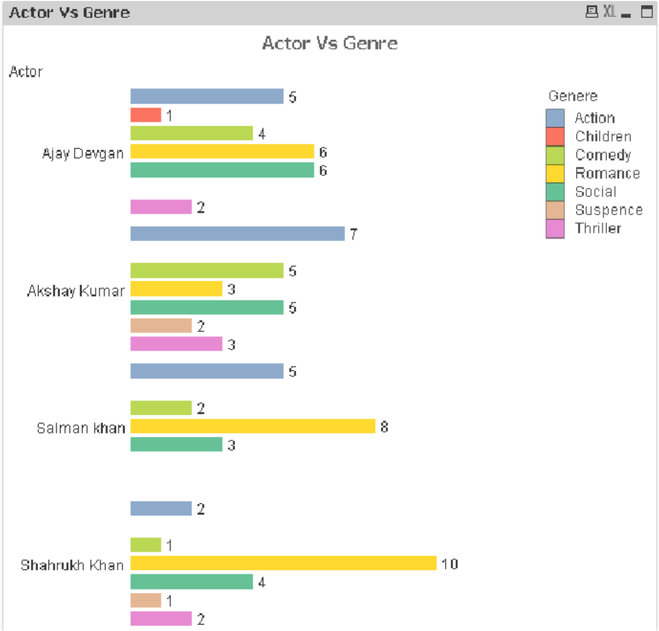

Fig 4: Actor vs. Genre

The following figures explain the statistics of our simulation data.

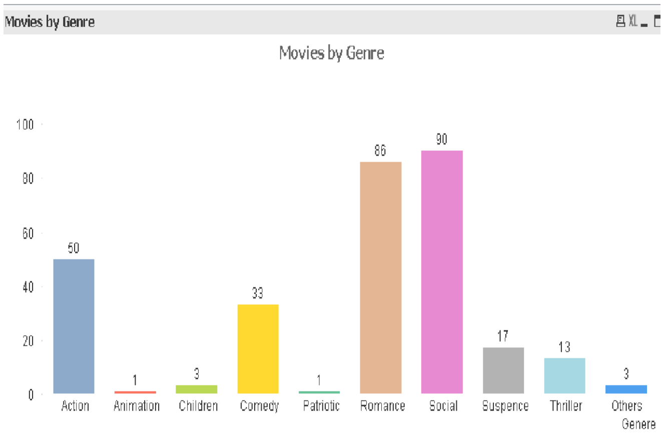

Fig 5: Movies by Genres

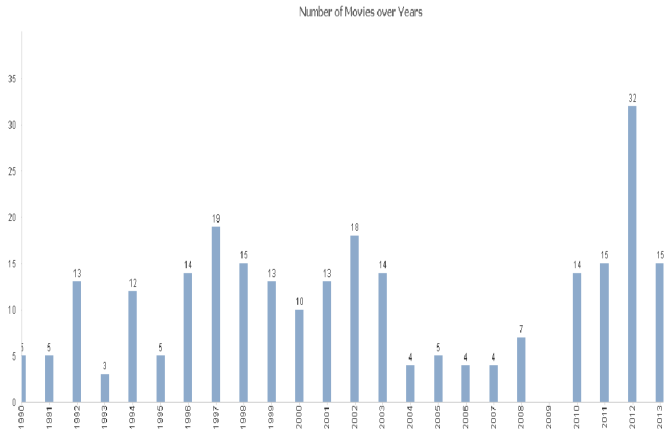

Fig 6: Number of Movies per Year

## VII. Conclusion and Future Scope

In this work, we have come up with a mathematical model to find the success rating of upcoming movies based on certain factors. As per our mathematical model, it was concluded that one factor was movie genres which determined the success rating of movies. It was also determined that the movie success depends on the cast of the movies. There was a strong correlation found between actors and the genres indicating that certain actors tend to work in certain genres. The actors and genres in turn define the success rating of the movie. Our work and results can be used to predict success or failure of upcoming movies by the movie makers as well as by the audience. A limitation of our work is that it focuses on only Bollywood movies currently. In the future, we will expand our model to include Hollywood movies.

## Acknowledgements

The contributions of Aveek Maity and Vamshi Penubakula towards this work are greatly appreciated.

## References

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 box-office 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.