Movie Success Prediction Using Data Mining

Alex Wazana Snir Zarchi

Department of Computer Science

Holon Institute of Technology

Abstract:

In the real world, prediction models and mechanisms can be used to predict the success of a movie. Historical data of each component such as budget, actors, director, producer and movie release date influences the movie's popularity. The rating prediction of a movie plays a vital role in the movie industry, because it involves huge investments. In our project, we aim to develop a system based upon data mining techniques using mathematical models to predict the rating of the upcoming movies based on several attributes. However, rating cannot be predicted based on a particular attribute. So, we have to build a model based on relation between several attributes. Each of the criteria will have a weight and then the prediction will base on these. The criteria were not limited just to the ones mentioned above.

Introduction:

A movie's revenue is greatly influenced by various components such as movie's director, actors, budget of the movie, movie's rating, release year and much more. Because of these multiple features, there is no exact formula that helps us analyze and predict how popular an upcoming movie will be. However, by analyzing the revenues generated by previous movies, a model can be built which can help us predict the expected revenue for a particular movie. Such a prediction could be very useful for the movie investors, who consider investing in the movie. The investors and the studios could decide a movie name fitting to the according model, appropriate advertisement for the movie or maybe decide about a different actor. This could be very useful for many movie theatres to estimate the revenues they would generate from screening a particular movie. Nowadays, online review system (such as IMDB) has become a very important part of any cinema business approach. Posting reviews online for products bought or services received has become a trendy approach for people to express opinions and sentiments, which is essential for business intelligence, vendors and other interested parties. We have to remember that in the cinema business, almost every customer checks online for movie reviews before buying a ticket. As we might know, social media contains rich information about people's preferences.

Our project proposes a decision support system for movie investment sector using data mining techniques. In our research, we will be using our own customized mathematics algorithms on dataset which contains different features. According to the calculated values, we will classify the movie into hit, average or flop. Through this project we aim to provide a data mining algorithm which gives the most accurate result for movie success prediction.

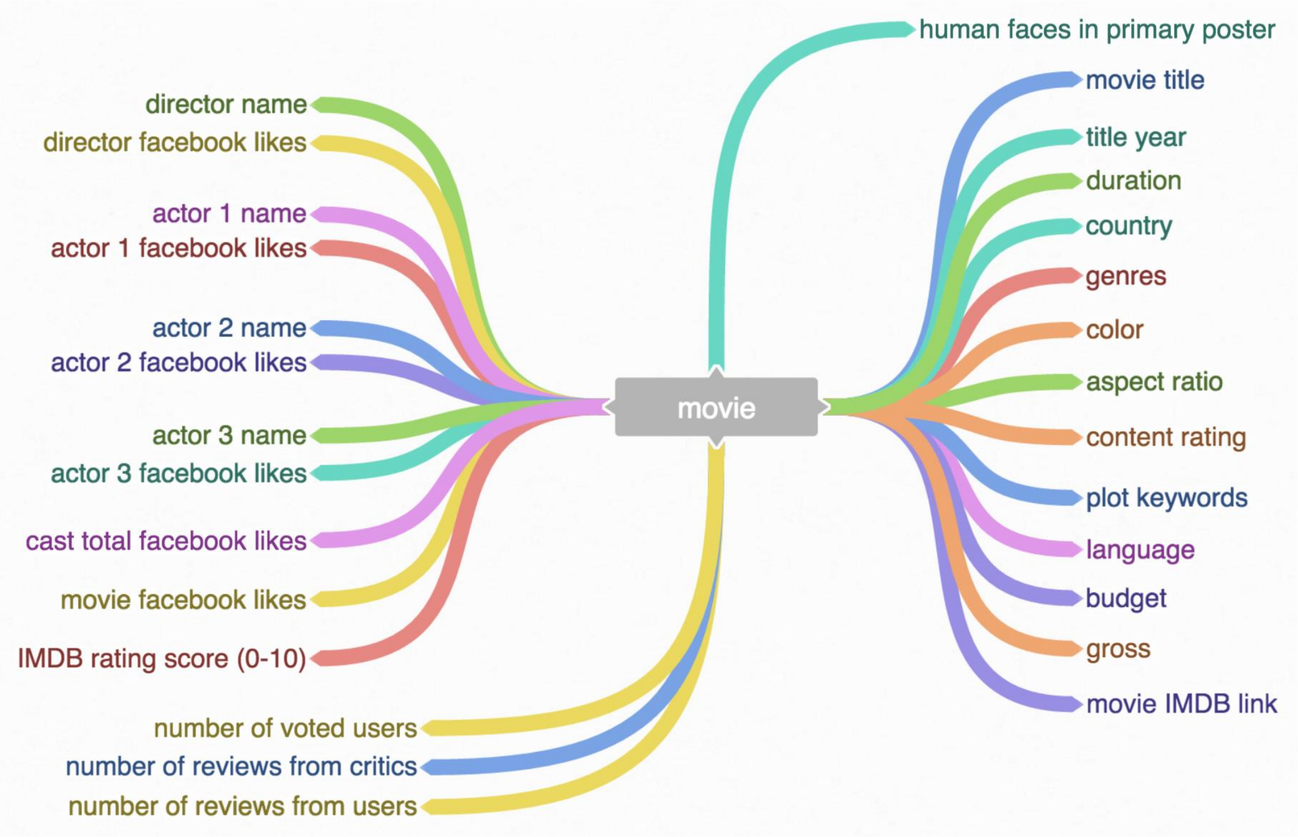
Experiments:

The dataset is from Kaggle website. It contains 28 variables for 5043 movies, spanning across 100 years in 65 different countries. There are thousands unique director names, and thousands of actors.

Based on the massive movie information, it would be difficult but interesting to understand what are the most important features that make a movie more successful than the others. Hence, the first thing we did is to describe the dataset, checked the values and we fixed any NaN/ missing values. Secondly, we used several different plots to implement our dataset, such as director name VS gross, director name VS amount of movies, first actor name VS gross and many more. With plotting our data, it could be very helpful to understand many conclusions.

In addition, we used mathematical algorithms, such as k-NN, k-Means, NB and linear regression, to find the success rating of upcoming movies based on certain features. The goal of this mathematical algorithms is to provide a precise prediction of success. Simulation data was used for this analysis and hundreds of records were cleaned, integrated and transformed. Various variables were trained to provide the movie success prediction. Some of these variables included budget, gross, actors, director, movie release year and many more. Their proposed model consisted of an algorithm that involved finding correlation between these various attributes.

The image below shows all the 28 variables that we used:



Results:

We were especially interested in knowing the answer to the question:

Will the number of human faces in a movie poster correlate with the movie rating?

As we know, movie poster is an important way to make public aware of the movie before the movie releases. It is quite common to see faces in movie posters. It should point out that most of the movies have more than one poster. We assume that a great movie needs to have a ″main″ poster, the one that the director likes most, or long-remembered by viewers. Overall, nearly 95% of all the posters have less than 5 faces. Besides, great movies tend to have fewer faces in posters. As well, if a poster has one or no human faces, we cannot tell if the movie is great simply from poster.

Moreover, out of the 28 features, we were especially interested to know **how does the IMDB rating score correlate with the other features**. From the plot we did in our project, we can infer that United States produced the largest amount of movies.

In addition, we explored **how does the IMDB rating score correlate with a movie release year**. In the last century, it seems that the number of movies produced annually largely increased. This is understandable since the development of filming industry goes hand in hand with the development of science and technology. But we should be aware that along with the big increase of movie industry, there are many movies with low IMDB score. As we know, the social network is a good way to estimate the popularity of certain phenomena. Therefore, it is interesting to know **how does the IMDB score correlate with the movie popularity in the social network**.

From the plot we did in our project, we can infer that, the movies that have very high Facebook likes tend to be the ones that have IMDB scores around 6-8.

Before starting prediction and regression algorithms, we made a **heatmap correlation**. The heatmap reveals that:

- The "cast_total_facebook_likes" has a strong positive correlation with the "actor_1_facebook_likes" and has smaller positive correlation with both "actor_2_facebook_likes" and "actor_3_facebook_likes".
- The "movie_facebook_likes" has strong correlation with "num_critic_for_reviews", meaning that the popularity of a movie in social network can be largely affected by the critics.
- The "movie_facebook_likes" has relatively large correlation with the "num_voted_users".
- The movie "gross" has a strong positive correlation with the "num_voted_users".

Moreover, we developed a **k-NN classification** model for the data. The movie dataset was divided into two parts: 80% of the movies were treated as the training set and the rest 20% belonged to the testing set. First we tested it with k=3 and its accuracy was 0.27. After using GridSearchCV algorithm, we found out that the best k value between 1 to 29 was 29 with 0.349 accuracy.

Regarding the decision tree, we tried to formulate a successful tree to a successful movie rating. The most influencing features were "movie_facebook_likes", "title_year" and "gross".

In the last part of our project, we developed a **k-Means classification** model for the data. After using Silhouette algorithm and a visual way with using a plot, we found out that the best value of k for clustering the data is 2, which means, 2 different clusters.
With describing the clusters data, we found out that the clusters divided by several features. One of the cluster was with high values of: director_facebook_likes, actor_1_facebook_likes, budget, gross, movie_facebook_likes language_num and country_num. The second cluster was with low values of these features.

Conclusions:

Overall, we have found that it is difficult to apply data mining techniques on the data. The data needs extensive 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 which making analyzing it more difficult. Much of the source data could not be integrated at all, without using natural language processing techniques. Despite these problems, we performed some useful machine learning algorithms on the data and uncovered information that cannot be seen by browsing the regular web front-end to the database. Furthermore, we can clearly state that according to our research and dataset, that there isn't a conclusive way to determine wither a movie will be a huge success, but we do manage to see a lean by using mentioned features above for better rating.