



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirccce.com

Vol. 5, Issue 8, August 2017

Ensemble Model for Movie Success Prediction

P.Deepthi¹, S.Bhargav²

Student, Department of Computer Science Engineering, SRM University, Chennai, India¹

Student, Department of Computer Science Engineering, SRM University, Chennai, India²

ABSTRACT: Movies are a booming business and a hub for investors like the producers, distributors and shareholders. The motivation to create a movie success predictor is to efficiently estimate the success of a movie that a producer or a distributor would like to invest in. The objective of this paper is to propose a novel method to forecast the outcome of a movie without human intelligence or intuition. Data mining techniques enable us to predict the outcome of a movie at the box office. We introduce a training model to our mining tool and further incorporate the results into an ensemble algorithm that is accurate and feasible. This model allows the investors to try various combinations to evaluate which would be most successful.

KEYWORDS: Logistic regression, correlation attribute evaluation, Weka, weighted average algorithm

I. INTRODUCTION

Movies are a popular mode of entertainment in today's world. Due to this increase in popularity worldwide, the demand for good and successful movies is also on the rise. Unfortunately, not every movie is successful. And as a result, it has become extremely difficult to predict the outcome of a movie. In this project a novel approach is taken to help predict the most likely outcome of a movie even before its release. This can be done in an efficient way using Data Mining. We aim at streamlining our efforts in the benefit of producers and distributors. Most producers and distributors of movies invest a large amount of money even though the outcome of the movie is uncertain and returns are not guaranteed. With this tool, they can estimate the probability of success of a movie, given the various factors that impact the success of the movie. This ensures that they make an informed decision about investing in an upcoming movie.

For this purpose, data of a set of sample movies with known results are collected and fed to the Weka Tool for mining. The dataset contains various attributes which affect the overall success of a movie like popularity of director, actor, actress and production house expectation, genre, budget, etc. We formulate an ensemble model combining the results of a classification algorithm with a feature selection algorithm. [2] The result of the mining algorithm is used to formulate a custom algorithm and determine the probable outcome of a future movie given just the above mentioned attributes. This integration will ensure that the investors compare various options before arriving on the most profitable one.

The movie success predictor can help determine the relationship and effects of various attributes on the movie's success. It can determine the probable outcome of a future movie and producers can wisely choose on which movie to invest on given this tool.

II. DATA DESCRIPTION

A. Data Acquisition

We acquired data from various websites such as [6] Wikipedia, [7] IMDB, [8] Wikishark, [9] Youtube etc. The data was obtained by the method of web scraping using python API. [1] The name of the movie, director, actor, genre, release date were obtained from its Wikipedia page. We also obtained the number of page views of the directors, actors, and production house from the Wikishark website. We took the number of recorded views of the movie's trailer from Youtube. Finally, the movie's budget and length were obtained from the IMDB page of the movie.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

B. Data Cleaning

The acquired data was cleaned using the following steps

- Any movie with missing data was removed from the data set.
- If the movie's director or lead actor did not have a Wikipedia page, then that movie was omitted.
- If the trailer of the movie hadn't released on Youtube, then the movie was removed from the data set.

C. Data Normalization

- Popularity Attributes: The number of page views for Director, Actor, Actress and Production House of a movie are normalized into 3 bins namely "high", "medium" and "low". The popularity was classified as "high" if the maximum page views per day exceeded 60,000, "medium" if it was between 40,000 and 60,000, and low if it was below 40,000.
- Genre: The Genre of the movie is taken from its Wikipedia page. If multiple Genres are present then only the first one is retained as processing a single genre will make the algorithm less complex.
- Expectation: The expectation of a movie is determined by normalizing the number of views its trailer recorded on Youtube. It is classified as "high" if the views exceeded 5 million, "medium" if it has views between 1 million and 5 million, and "low" if the number of views were below 1 million.
- Budget: The budget was converted to USD and grouped into the same three categories as the other attributes. It was classified as "high" if it exceeded \$100 million, "medium" if it was between \$50 million and \$100 million, and "low" if it was under \$50 million.
- Length: The length of the movie's screen time in minutes was sorted into "long", "medium", and "short". The movie was a "long" one if it surpassed 120 minutes, was a "medium" one if its screen time was between 90 and 120 minutes, and was a "short" movie if it had less than 90 minutes of screen time.
- Time of Release: The time of release of a movie was categorized into the four seasons. It was a "spring" release if it released between the months of January and March, a "summer" release if it released between the months of April and June, an "autumn" release if it released between the months of July and September, and a "winter" release if it released between the months of October and December.
- Series: This attribute gives information on whether the movie is a part of a series or not. It takes binary values "Yes" or "No".
- Success: Finally, the class variable or the success of the film take binary values which are "yes" or "no" respectively and the value is determined by the box office collection of the film. A film is classified successful only if it had grossed 1.5 times the amount of its total budget in our criteria.

D. Data Attributes

The final data set required has 10 attributes and a class variable. Which are,

- Popularity of director { High, Low, Medium }
- Popularity of actor { High, Low, Medium }
- Popularity of actress { High, Low, Medium }
- Popularity of production { High, Low, medium }
- Genre { Action, Horror, Comedy, Drama, Sport, romance, Fantasy, Animation, Thriller }
- Expectation { High, Low, Medium }
- Budget { High, Low, Medium }
- Length { Long, Short, Medium }
- Time of Release { Summer, Winter, Spring, Autumn }
- Series { Yes, No }
- Class variable: Successful { Yes, No }.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

III. PROPOSED ALGORITHM

A. Description of the proposed algorithm

The classification method we use for building the model is logistic regression. [4] Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variable. As we have a dichotomous class variable, logistic regression is a suitable and efficient classification technique. It calculates the Log odds for each possible instance of an attribute, which we will further use in our algorithm to calculate the probability of success given a set of possible values for the attributes as shown in Fig. 1.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p) = -\log\left(\frac{1}{p} - 1\right).$$

Figure 1: Logistic Regression Formula

Correlation Attribute evaluation is a special version of feature selection available in the Weka toolkit. [3] It evaluates the worth of an attribute by measuring the Pearson's correlation coefficient between the attribute and the class variable. Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is arrived at via a weighted average. A ranker search algorithm is implemented, which ranks attributes based on their individual evaluations. In general, [5] the ranker method determines which attributes should obtain high or low rank according to the selected attribute in the given datasets. Ranker is providing a rating of the attributes, ordered by their score in the evaluator. In our dataset, on applying correlation attribute evaluation technique to it, we arrive at a result that ranks the attributes according to its influence on the class variable or numerically, the Pearson's coefficient, which we will treat as the measure of importance of that attribute.

Now that we have obtained the probabilities of success of each instance of an attribute using logistic regression and the associated weights of each attribute through feature selection, we now combine these two using a weighted average algorithm, to determine the success of a movie. Each probability is multiplied with its corresponding weight and these products are summed. The cumulative is then divided by the sum of all the weights to produce the weighted average as shown in Fig. 2.

$$P_i = \frac{(P_{a1} \cdot W_{a1}) + (P_{a2} \cdot W_{a2}) + \dots + (P_{a10} \cdot W_{a10})}{W_{a1} + W_{a2} + \dots + W_{a10}}$$

Figure 2: Proposed Algorithm Formula

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

B. System architecture

The architecture of our proposed model is as shown in Fig. 3.

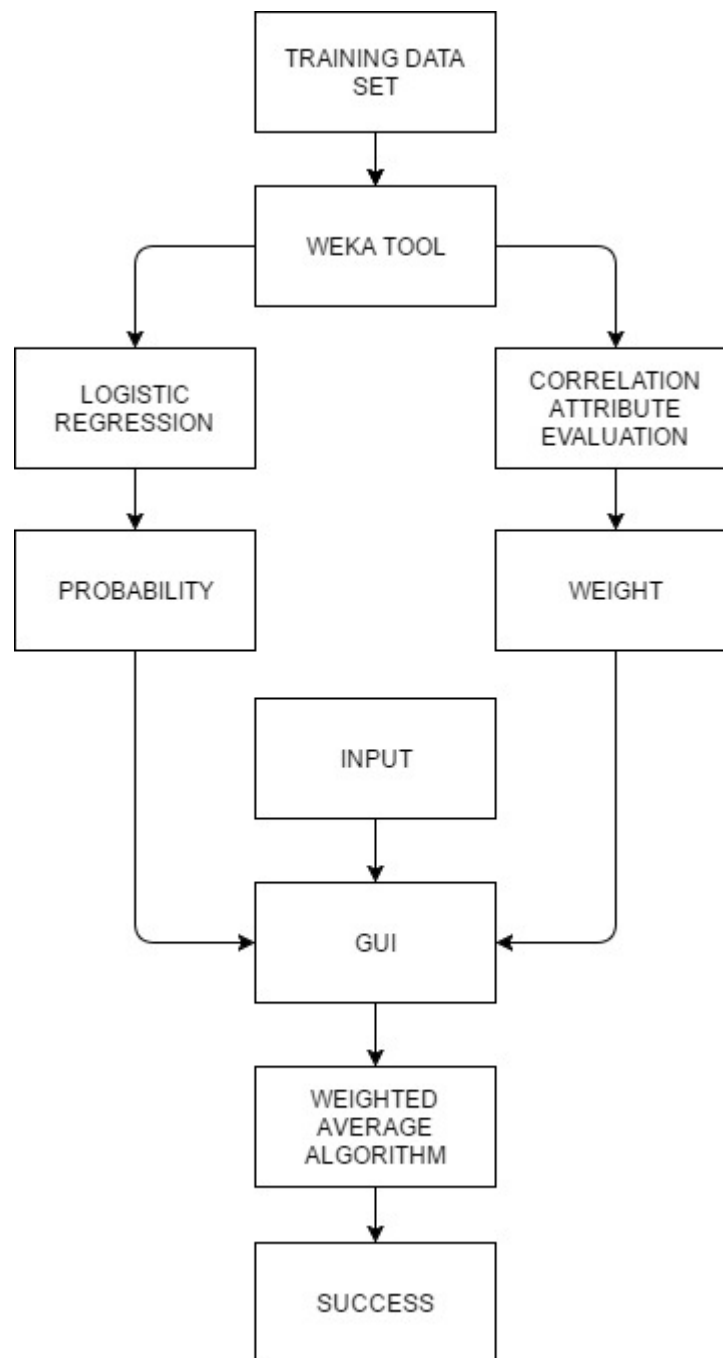


Figure 3: System Architecture



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

IV. RESULTS AND INFERENCES

A. Logistic Regression Outcomes

The output for the logistic regression, giving the Logit odds of each instance of the various attributes from the weka tool is as shown in fig. 4 :

Variable	Odds Ratios...	Class No
Popularity of Director=Low		0
Popularity of Director=High	2.7945956174098585E15	0
Popularity of Director=Medium		2.2108
Popularity of Actor=Low	177665635.6395	0
Popularity of Actor=Medium		0
Popularity of Actor=High		0
Popularity of Actress=Low	349.8527	0
Popularity of Actress=Medium	175844005.4939	0
Popularity of Actress=High		0
Popularity of Production=Low	5.0308066990064415E18	0
Popularity of Production=High		0
Popularity of Production=Medium		0
Genre=Action	3.1698949752151368E19	0
Genre=Horror		0
Genre=Comedy	1895.1952	0
Genre=Drama		0
Genre=Sport		0
Genre=Fantasy	35.2444	0
Genre=Romance		0
Genre=Animation	4.3537542697836855E19	0
Genre=Thriller	9.801392487054845E22	0
Expectation=Low	2.1492140969082043E10	0
Expectation=High		0
Expectation=Medium		0
Budget=Low	0.0002	0
Budget=Medium	4.8635580295416315E11	0
Budget=High	0.0083	0
Length=Medium		0
Length=Short		0
Length=Long	2.0652185152042588E16	0
Time of Release=Winter	0.0001	0
Time of Release=Spring	6.6275	0
Time of Release=Summer	619.6742	0
Time of Release=Autumn	0.9159	0
Series=Yes		0

Figure 4: Logistic Regression Outcomes

B. Feature Selection Outcomes

The outcome of the correlation attribute evaluation using a ranker search algorithm gives us the following weights for the 10 attributes, as shown in fig. 5.

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

```
Ranked attributes:  
0.4257 10 Series  
0.4209 6 Expectation  
0.2758 1 Popularity of Director  
0.275 3 Popularity of Actress  
0.2702 4 Popularity of Production  
0.2215 2 Popularity of Actor  
0.166 7 Budget  
0.1536 9 Time of Release  
0.0934 5 Genre  
0.0444 8 Length
```

Figure 5: Feature Selection Outcomes

C. Proposed Ensemble Model results

We created a rudimentary user interface to accept values of the various attributes on which we apply our algorithm to predict the probability of success for that movie as shown in Fig. 6. Given the nominal values of the various attributes, the interface produces the probability of success in a percentage value, based on the training set as depicted in Fig. 7.

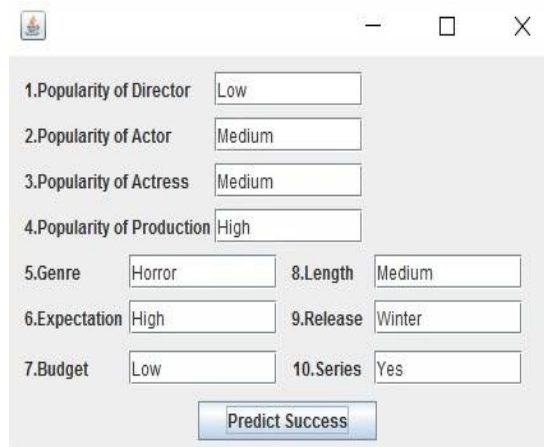


Figure 6: GUI Input

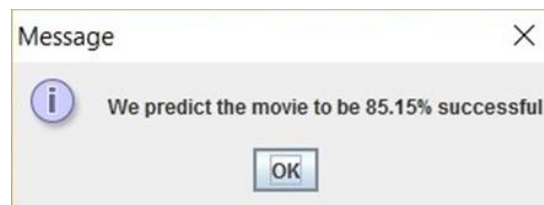


Figure 7: GUI Output

To measure the accuracy of our model, we tested it with our own training data set. The confusion matrix obtained from the result of this is as shown in Fig. 8. The green boxes of the matrix represents the correctly predicted instances whereas the red boxes denote the incorrectly predicted instances. A total of 387 movies out of 500 were correctly classified, hence making out model 77.4% accurate. Although this accuracy is a little lesser than normal classification models, when compared to other ensemble models like bagging, stacking and AdaBoost, our model is more accurate.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

	Predicted NO	Predicted YES
Actual NO	118	39
Actual YES	74	269

Figure 8: Confusion Matrix

D. Inferences

The model is highly accurate and flexible. IT is better in comparison to other models because it allows the investors or the end users to determine which factors influence the success of a movie the most and thereby tweak the factors to their benefit. For example, the investors can change the release time and check if the rate of success increases. Furthermore, the model is simple and highly feasible. It offers adaptability, extensibility and reusability.

V. CONCLUSION AND FUTURE WORK

In this paper, we have successfully devised an algorithm to predict the success percentage of a movie given its various attributes, using data mining techniques. It can be safely concluded that this is a highly efficient GUI based component, achieving the motive to ensure safe business investments. This project is flexible enough to accommodate an increase in the size of the data set for more accurate results. The algorithm used in this project can also be applied in various other domains like short films, TV series, reality shows etc. to predict their success. The interface can be enhanced by allowing the user to input numerical values for attributes like budget and length. Furthermore, the user can give the names of the actors and director and the program will estimate the popularity from a predefined database.

REFERENCES

1. AnandBhave, Himanshu Kulkarni, VinayBiramane, PranaliKosamkar, 'Role of different factors in predicting movie success', 2015 International Conference on Pervasive Computing (ICPC).
2. Travis Ginmu Rhee, FarhanaZulkernine, 'Predicting Movie Box Office Profitability: A Neural Network Approach', 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA).
3. Mark A. Hall, 'Correlation-based Feature Selection for Machine Learning', April 1999.
4. <http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/Logistic.html>
5. <http://weka.sourceforge.net/doc.dev/weka/attributeSelection/CorrelationAttributeEval.html>
6. <https://en.wikipedia.org/>
7. <http://www.imdb.com/>
8. <http://www.wikishark.com/>
9. <https://www.youtube.com/>