IJESMR
International Journal OF Engineering Sciences & Management Research

A DATA MINING APPROACH TO LANGUAGE SUCCESS PREDICTION OF A FEATURE FILM

Nikhil Chaudhari *[1], Karthik Vardhrajan[2] , Shashank Shekhar[3], Prabhjit Thind[4] and Swarnalatha P[5]
[*1,2,3,4,5]SCOPE, VIT University, India [*1]Department, Collage, Country

Keywords: *Success Prediction, Decision Tree, RapidMiner, Celoxis*

ABSTRACT

The paper aims to develop a tool, which can predict the success of movie being a hit or flop. As this factor is important for everyone involved in the movie, for example : If a movie is flop, it exacerbates the image of actor or director. The tool will use searching algorithms and then use of bespoke system to predict the percentage of success of movie which is yet to be released. This paper details our analysis of the data collected from various resources like IMDb, Kaggle. We gather a series of interesting facts and relationships using a variety of data mining techniques such as Bayes Classification Algorithm, Decision Tree etc. Subsequently, a classifier is learned and used to classify new movies with respect to their predicted box-office collection. Experimental results show that the proposed approach improves the classification accuracy as compared to a fully independent setting.In particular, we discover the rate of success with respect to various parameters such as language, country, budget, Facebook likes of the actors and actresses etc and focus on relevant details such as the relationship between the budget of the movie and rating of the movie, language and rating, facebook likes and rating etc. The data mining techniques used will enable us to uncover information which will both confirm or disprove common assumptions about movies, and also allow us to predict the success of a future film given select information about the film before its release

INTRODUCTION

The production of movie is both an industry and art. Producers around the world are concerned about the success of feature films. Many researchers have used user ratings from social media websites or specific movie review websites. In this paper, other attributes like actor's facebook like, duration of the movie, language, IMDb score, aspect ratio, a movie's facebook likes, country, content type etc. The number of attributes used in this method allows high accuracy.

The high accuracy of movie success prediction has a lot of use for companies to plan their resources. For example, a Hollywood studio, that expects its newest movie to be highly successful will rent more theatre rooms in advance, increasing revenue if the prediction turns out to be true. If it rent less theatre rooms, not all viewers might have been able to watch the movie in its opening weekend.

In this paper, the approach used to complete this project and effort is discussed in detail. The outcome of this research is that it classifies the data based on a certain label attribute and provides with tools and techniques to transform the dataset into a suitable format. The data initially obtained had a lot of anomalies which has to be rectified. After the data has been suitably cleaned and integrated, it will then need Selection and Transformation, to translate the textual information (where necessary) into numerical information which can be better analysed by data mining processes. This stage will also discard irrelevant data, and may select a subset of the data to be mined, since the original set may still have hundreds of thousands of records.

A classifier is learned and used to classify new movies with respect to certain parameters. Experimental results show that the proposed approach improves the classification accuracy as compared to a fully independent setting.

Celoxis which is an integrated and collaborative web-based platform to manage your projects, finances, resources and business processes online, was used for management of the project related tasks.

The paper consists of the analysis which explains the life cycle approach, calculates the cost and estimation of the project in detail, the methodology adopted and the risk analysis.

IJESMR
International Journal OF Engineering Sciences & Management Research

RELATED WORKS

Movie success prediction is very important for actors, artists, directors and mainly producers. Earlier Latif [2] used machine learning approaches to predict movie popularity. Cook[3] et al. experimented with a large number of attributes and achieved 65% accuracy using random forest classifier.

In 2002, various specialists assessed regardless of whether a film's case office execution could be anticipated by evaluating the likelihood of a film's income passing a specific threshold[6]. Through examination of movement on Wikipedia pages, another calculation was made to anticipate whether a film flops or turns into a blockbuster[7]. There are additionally numerous assets assessing lifetime gross of a film in light of their prosperity amid opening weekend[8]. The main difference between these related works and ours, is that none of them predict movie success before the production begins

M. Saraee, S. White & J. Eccleston[9] analysed Internet Movie Database(IMDb) which is a free, online source for 3,90,000 movies, television series and video games, containing movie descriptions and user ratings. They found out that it is difficult to mine data from IMDb because of the format of the source data. They firstly concluded that the budget of the film gives no indication on the movie rating. Secondly it was observed that there was a downward trend in the quality of films over time. Thirdly the director and the actors/actresses involved in the film play the most important role in the success of the film.

ANALYSIS

- **Life Cycle Approach**
  Our project is adopting Waterfall Model i.e. it is one-shot and once-through. It does not involve any increments or evolution as such. Our project will basically implement the classification of data to predict the movie success rate and help in finding which movie will succeed in the market and which might flop.
- **Estimate of Effort and Time** :
  Size is a fundamental measure of work. Our project lies in the category of medium size i.e. it lies between larger projects and smaller projects when estimation is done.

Based on the estimated size, two main parameters are estimated:

1. **Effort**:

One person-month is the effort an individual can typically put in a month. In our project, one person works for about 10 hours in a month. On this basis, one person month for our project and project team will be equal to –

One Person-Month = 10/30 = 0.33333

Thus the amount of effort required for the given project is quite low.

2. **Duration**:

Table 1 : Table for activity and time period for each activity

Activity	Time Duration Required
Planning - A	2 weeks

Analysis (Data Collection and Analysis) - B	2 weeks
Design - C	3 weeks
Implementation - D	3 weeks
Testing - E	2 weeks
User Feedback - F	1 week
Total Time (Required)	**13 weeks**

Our project follows the bottom up approach as our project does not have any past project data. Therefore our project is quite time consuming in nature.

Project Network

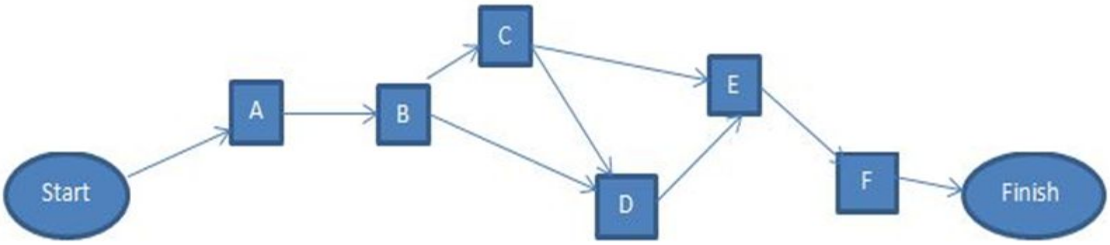

Figure 1 Network diagram for the project

The following estimates were calculated while performing the project.

Effort Estimates :

On applying **COCOMO81** model,

Effort = c * (size)^k

Our project is of **semi-detached model** :

So, c=3 and k=1.12

So,**Estimated Effort** = 3 * (0.525)^1.12 = 1.457 person-months

Actual Effort = 1.5 person-months

- **Effort variance** = (Actual Effort – Estimated Effort)/ Estimated Effort
  - = (1.5-1.457)/1.457 x 100
  - = 2.951

- **Size variance** = (Actual size – Estimated size)/ Estimated size x 100
  - Actual Size = 0.525
  - Estimated Size = 0.450
  - Size Variance = (0.525 – 0.450)/0.450 x 100 = 16.67%

- **Project Productivity** = Actual Project Size / Actual effort
  - = 0.525/1.5
  - = 0.35

- **Productivity (defect detection)** = Actual number of defects (review + testing) / actual effort spent on (review + testing).
  - =3/0.5
  - =6

- **Productivity (defect fixation)** = actual no of defects fixed/ actual effort spent on defect fixation.
  - =3/0.5
  - =6

- **Defect removal efficiency**: It basically quantifies the efficiency with which defects were detected and prevented from reaching the customer.

    - **Defect removal efficiency** = (1 – (total defects caught by customer/ total no of defects)) x 100

    - =1-(1/3)x100 = 66.67%

- **Review efficiency**: It is basically defined as the efficiency in harnessing/ detecting review defects in the verification stage.

Review efficiency=(number of defects caught in review)/ total number of defects caught) x 100

=(2/3)x100

=66.67%

The tool used for the generation of decision tree is RapidMiner. The following figures explain the architecture used for decision tree.

RISK ANALYSIS

The following risks were identified during the review of project.

Risk 1:
Problem: Incomplete data or tuples with Null values reduces the accuracy of the decision tree.

Solution: When the tuples with Null values, in a large dataset are less in number, then these values can be ignored.

Risk 2:
Problem: Time is one limiting factor to an organization using a decision tree tool. They take time to construct and can tend to be so annoying.
Solution: This time constraint can be reduced by carefully choosing the predictor parameter.

Risk 3:
Decision trees will only work effectively if the accurate guiding information is provided. Introduction of inaccurate information at the top affects all the decisions to undertake. You will therefore be forced to correct the entire decision tree. This means delay, followed by unsatisfied customer because misleading information was passed.

Risk 4:
They are not favorable for quick decision making. Offering efficient and effective customer care in such an organization needs complete investment in time.

METHODOLOGY

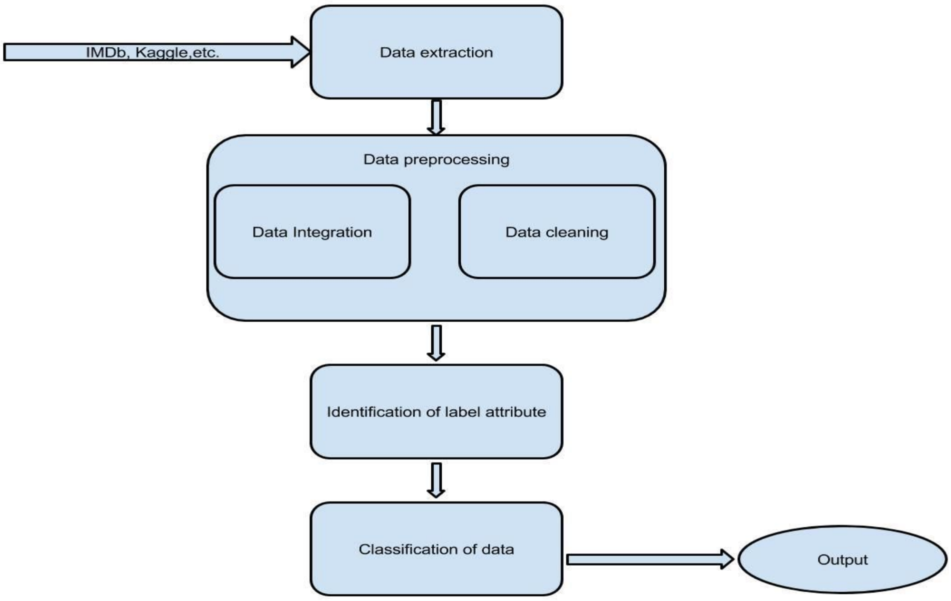

Figure 2 Process of data mining

The Role of a quality mirrors the part played by that property in an ExampleSet. Changing the part of a quality may change the part played by that property in a procedure. One quality can have precisely one part. This operator is utilized to change the part of at least one qualities of the information ExampleSet. This is an extremely basic operator, you should simply to choose a characteristic and select another part for it. Distinctive learning operators require traits with various parts.

The Split Validation operator is a settled operator. It has two subprocesses: a preparation subprocess and a testing subprocess. The preparation subprocess is utilized for learning or building a model. The prepared model is then connected in the testing subprocess. The execution of the model is additionally measured amid the testing stage.

A decision tree is a tree-like diagram or model. It is more similar to an upset tree since it has its root at the top and it becomes downwards. This representation of the information has the preferred standpoint contrasted and different methodologies of being significant and simple to decipher. The objective is to make an arrangement model that predicts the estimation of an objective characteristic (regularly called class or mark) in view of a few info qualities of the ExampleSet.

DECISION TREE ALGORITHM

Decision Trees are produced by the recursive partitioning. Recursive partitioning means repeated splitting on the values of the attributes. In every recursion, the algorithm follows the following steps:

1. An attribute A is selected to split on. Making a good choice of attributes to split on each stage is crucial to generation of a useful tree. The attribute is selected depending upon a selection criterion which can be selected by the criterion parameter.
2. Examples in the ExampleSet are sorted into subsets, one for each value of the attribute A in case of a nominal attribute. In case of numerical attributes, subsets are formed for disjoint ranges of attribute values.
3. A tree is returned with one edge or branch for each subset. Each branch has a descendant subtree or a label value produced by applying the same algorithm recursively.
4. In general, the recursion stops when all the examples or instances have the same *label* value, i.e. the subset is pure.

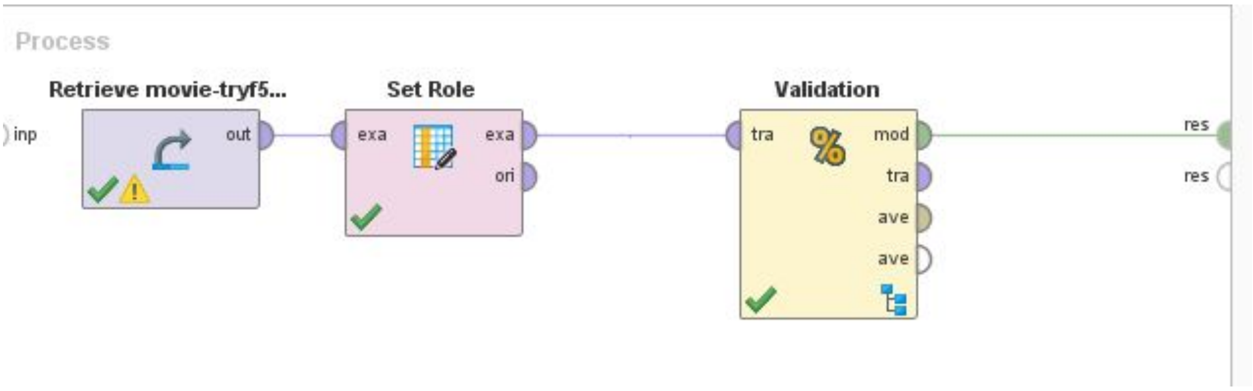

Figure 3 Process blocks of main process

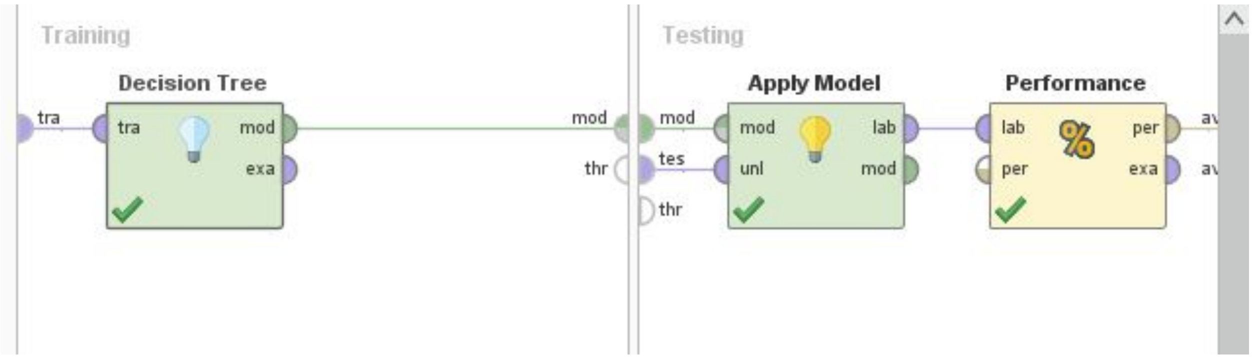

Figure 4 Block for training and testing model

RESULTS

The following tree was obtained after the process was completed.

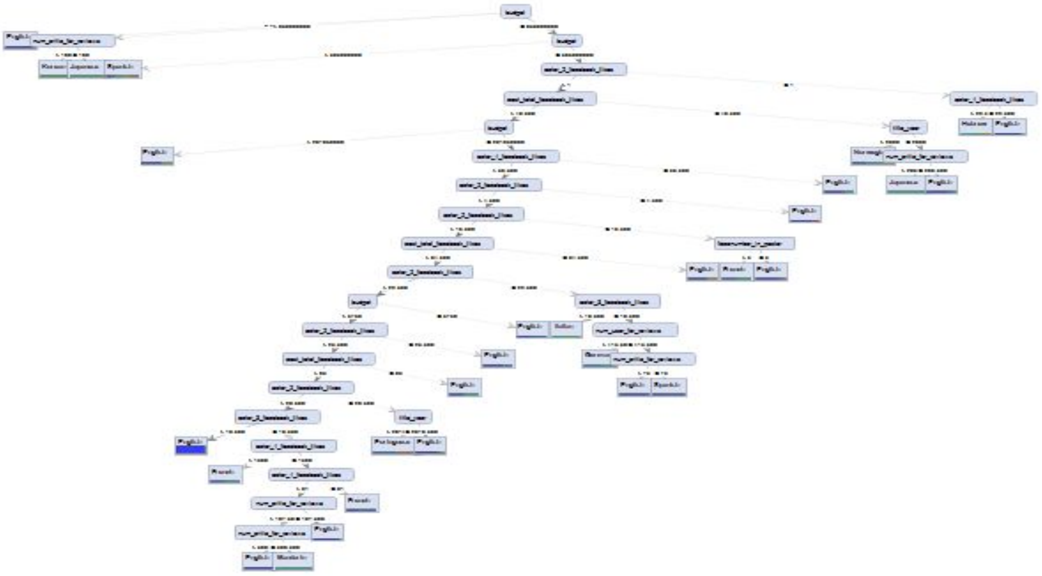

Figure 5 Output Tree

The accuracy found for country label is 78% and language is 94%.

Confusion plot

A confusion matrix is a table that is regularly used to portray the execution of a grouping model (or "classifier") on an arrangement of test information for which the genuine qualities are known. The moderately easy to see, however the related phrasing can be fuddle.

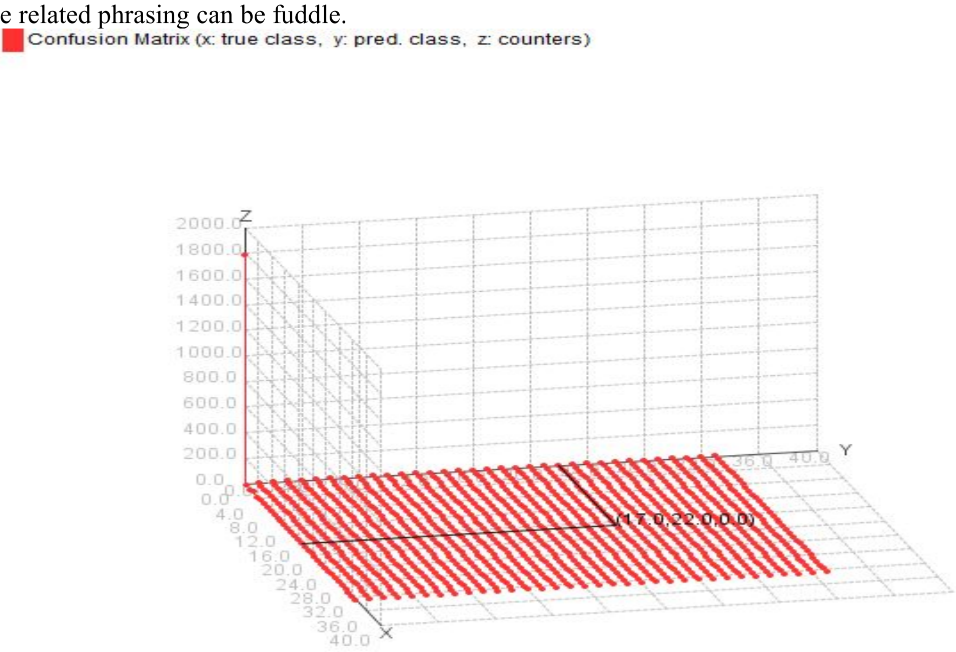

Figure 6 Confusion plot

CONCLUSION

Overall, we have found the accuracy for each classification method as well as analyzed the dataset obtained from the resource. The data obtained has been analyzed only after 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, making mining more difficult. Much of the source data could not be integrated at all, without using natural language processing techniques. More importantly, we believe that our research shows promise for further development in this area. Given additional time to incorporate more of the source data available, and some use of natural language processing techniques, other interesting patterns in the data may become apparent. A more accurate classifier is also well within the realm of possibility, and could even lead to an intelligent system capable of making suggestions for a movie in pre-production, such as a change to a particular director or actor, which would be likely to increase the rating of the resulting film.

REFERENCES

1. Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early prediction of movie box office success based on Wikipedia activity big data. *PloS one*, *8*(8), e71226.
2. Latif, M. H., & Afzal, H. (2016). Prediction of Movies popularity Using Machine Learning Techniques. *International Journal of Computer Science and Network Security (IJCSNS)*, *16*(8), 127.
3. Cook, C., Cunningham, B., Reading, E., Sedgewick, M., & Tilcock, K. Predicting Blockbuster Success.
4. Krauss, J., Nann, S., Simon, D., Gloor, P. A., & Fischbach, K. (2008, June). Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis. In *ECIS* (pp. 2026-2037).
5. Omenitsch, P. (2014). Predicting Movie Success with Machine Learning and Visual Analytics. *Technische Universutat Wien*.
6. A.Colins ,et al. ,"What makes a blockbuster? Economic analysis of film success in the United Kingdom", Managerial and Decision Economics, vol.23, John Wiley Sons Ltd, 2002, pp. 343-354.
7. M.Mestyan, et al. (2013). "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data" . [Online]. Available FTP : http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0071226
8. P.Perner ."Machine Learning and Data Mining in Pattern Recognition ," New York, NY, FinalRep., July 2013.
9. M. Saraee, S. White & J. Eccleston (2004). "A data mining approach to analysis and prediction of movie ratings ", 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9.