Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 02, February-2020

Predicting the Conceptual Appeal of Movies using Data Analytics

M. Vaishnavi, Sourabh S Kulkarni
Kusuma H
(Associate Professor)
Dayanand Sagar College of Engineering

***Abstract-* With increasing volumes and data types and piquing interest to use data to produce invaluable insights, it has become the most major areas of study in the present era. Huge datasets are available for predictive analysis of several aspects of movies and many domains are available for making predictions. It is beneficial to all varieties of people associated with the art of movie-making and watching. Stakeholders like producers can know the risks and advantages of investing in particular movies. Movie watchers can determine if the movie is up to the mark and worth their money. This paper aims to explore the different techniques used for predictive analysis. We also seek to explore what factors are necessary to predict the quality of a movie in terms of its concept and how to establish a relationship between different categories.**

I. INTRODUCTION

A. Background and motivation

Data analytics is the analysis of raw data to make conclusions about a particular information. Most of the techniques and processes of data analytics are automated into algorithms which work over unprocessed data for human use. Predictive analysis is one such field of advanced analytics, used to make predictions concerning the unknown occurrences in future. It uses several techniques like statistical learning, modelling, and machine learning to analyze present data and make future predictions. Predictive models for analytics can capture relations between various factors to study risks with particular sets of conditions to assign a weightage or score. Examples of the use cases are a quality assurance of a product, sentiment analysis, and risk modelling.

B. Movie Quality Prediction

Movie quality prediction is the approximate prediction of the qualitative aspects of a movie. Such predictions have important applications. Box office successes do not always guarantee conceptually strong movies or good acting. Moreover, many individuals have inadequate information about an upcoming movie, which may lead them to search for additional information before deciding whether to watch that movie. The prediction tool helps in avoiding wastage of their money and time.

A movie has several contributors such as directors, actors, music directors, lyricists, screenplay, camera, lighting, etc. However, a movie is not affected by all these criteria equally. Determining the aesthetic nature of movies has become cumbersome due to the availability of many factors that may or may not affect a movie. Hence, a challenging task is to identify the characteristics that actually affect the movie quality.

II. RELATED WORKS

A. Early Work

In 1709, Roger De Pilates, a French art critic was one of the first ones to try to elaborate characteristics of quality or beauty. He decomposed painting into the four basic characteristics namely drawing, composition, expression and colour. Each of them was given a rating from zero to 20 for a painting and then these ratings were aggregated to give a combined rating of the painting. The concept of decomposing the entire judgement into these characteristics and explaining the way to rate them individually and thereby, eventually aggregating them was extended to many art forms including movies. [3]

The research was conducted regarding whether awards affect the economic success of movies. In one such research, it was predicted initially that movies that won awards like Oscars or Golden Globe were much successful than other nominated movies. However, success may precede Oscars as competing movies are released much before the awards. To check if awards influenced success, it was important to account for revenue generated after the Oscars. An example of ‘Singing in the Rain’ is considered. It is a movie released in 1952 which appears in prominent lists like ‘100 greatest movies by 1500 leaders from American Film Community’,’100 best movies of all time’ etc. which was not nominated at all, whereas ‘The greatest show on earth’ was given an Oscar though it appears nowhere on the lists today. Hence, it was determined that awards and prizes are bad predictors of the quality of movie. [3]

B. Recent Work

Different approaches to movie success have been used. A movie can be reviewed using two parameters like positive (+) for good and negative (-) for bad. These parameters can be extended to be more explanatory like very negative (--), somewhat negative (-), very positive (++) and somewhat positive (+). This was a binary classification done using random forests. [2]

Social media plays an important role in determining how well a movie has been received or criticized. Movies are always subjected to comments and discussions among a vast audience with varied opinions. On studying the correlation of the social media structure with box office revenues, it can be predicted whether a movie will be nominated for the Oscars. [4]

Predicting the profitability of a movie at the early stages of film production to support movie decisions can be very crucial to prevent losses due to wrong decisions. This can be done by considering questions like "who" has been casted, "what" the story of the movie is, "when" it will release and other features. [5]

Visual-interactive approaches have been studied for predictive analysis, aiming to enhance the prediction and enable vaster user groups to use these predictive tools. The quality of such predictions depends heavily on the model used and the training data. Furthermore, background knowledge needs to be taken into account during the generation of models and analysis. Visual analytics uses visualizations to capture such knowledge and allow the users to steer the analysis process to improve the prediction eventually. [6]

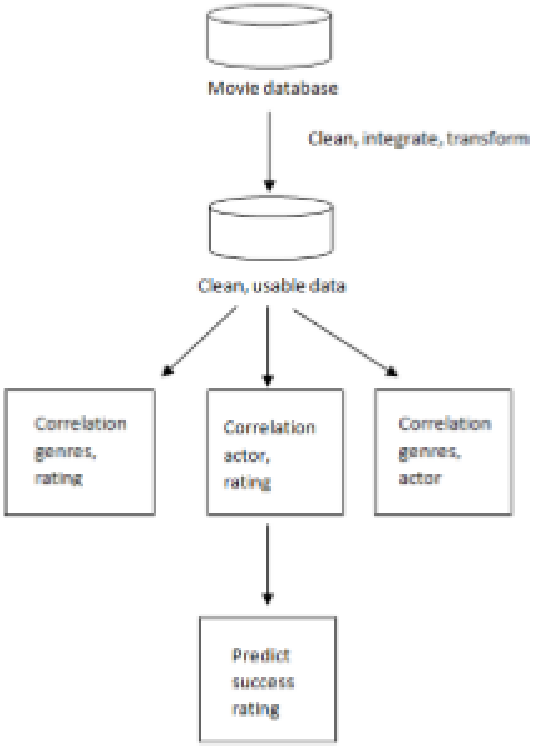
Sentiment analysis uses deep learning techniques to classify movies based on user reviews as positive and negative. In this, the sentiments of the people are used to classify the movie, but the reviews are purely personal. Suggestions given by one person may not be liked by another. Reviews suggested by the majority are taken into consideration. Twitter reviews are considered and abnormal posts that do not match the pattern of the general trend are flagged as abnormal. [7]

In another study, film revenues are thought to be dependent on various independent variables. On estimating the equation, we can gain some insights into the significance of every autonomous variable to explain the fluctuations in revenues for a number of movies. Multiple regression analysis is a statistical method to fit observed data in an estimating equation. Coefficients in regression analysis are generated for all autonomous variables in the equation. Revenue changes are represented by coefficients resulting due to a small variation in the respective independent variable. Larger the statistical fit, larger is the variation within the dependent variable which means the revenues are explained and hence, higher is the accuracy of equation. [11]

III. METHODOLOGY

Movie success prediction in recent times uses data mining, which enables exploration of patterns and trends in a given set of data and to identify relationships among various variables. This leads to sequential events like, classifications, clustering and hence predicting the future events. These predictions become helpful to movie stakeholders who venture to invest their resources in the creation of the movie to reduce economic risks. Success here is hugely determined based on revenues collected by the movie after its release. [8]

The proposed model used the concept of correlation between different parameters that would best suit the interest of the stakeholders. A correlation was found between (i) genres and ratings and (ii) actors and ratings. The mathematical model of correlation shows a measure of dependence between two variables. It can be either a positive (the parameters move towards in the same direction) or negative (they move together in the opposite directions). After finding the correlation values, the p was found very low for some and high for others. A low value determines a correlation between parameters. [8]


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

Random Forest is the most widely used classification algorithm, proficient in both regression and classification. It can classify accurately, large amounts of data. An ensemble learning technique for regression, classification and many other tasks, it operates by building a number of decision trees at the time of training and outputs that class which is the mode of all the classes or mean prediction of the singular trees.

This algorithm is a combination of many decision trees. Every tree depends on the random vector value sampled independently and having the same distribution for every tree in the "forest." Each tree is grown to the greatest extent possible.

For growing these ensembles, vectors are often randomly generated which govern the growth of every tree present in the ensemble. Bagging is an example where a tree is grown by doing a random selection from the training set. We can also use random split selection where the split is chosen at each node from K best splits randomly. New training sets were generated by randomizing the training set outputs. [9]

The recurring element in these procedures is that for kth tree, a vector Θ_k is generated randomly, which is independent of past random vectors Θ_1,... Θ_{k-1} though having the same distribution. The tree is grown by making use of the training set as well as Θ_k, that results in a classifier $h(x, \Theta_k)$ where x is an input vector. The dimensionality and nature of Θ depend on its usage in the building of the tree. [9]

The advantages of this bagging are that (i) it has good accuracy (ii) relatively robust to noise and outliers and (iii) provides estimations for strong correlation and strengths and also (iv) corrects the problem of over fitting the training data. It is best suited to represent movie features. [10]

IV. CONCLUSION

The purpose of doing this research was to analyze various methodologies that are used for predictive analytics. Many approaches are studied and a number of different predictions have been made based on different problems. For carrying out predictive analysis, it is clear from the research that the most accurate results are obtained using the random forest algorithm. It is also clear that the studies conducted, aimed to determine those features of movies that would increase the box office success. However, we inferred that box office success is not the most important factor to ensure if a movie is conceptually appealing. It is important to determine if the concept is getting lost in the movie world of glamour and money. We were intrigued by the idea of applying data analytics to predict the conceptual quality of movies before their release by using the categories that can best describe them. Such an analysis would enable moviegoers to make better decisions about watching certain movies and also save their time and effort in researching about an upcoming movie.

REFERENCES

[1] K. Yessenov, S. Misailovic, "Sentiment Analysis of Movie Review Comments" Spring 2009 final project, pp: 1-17.

[2] H. Pouransari, & S. Ghili “Deep Learning for Sentiment Analysis of Movie Reviews”,2014, From: http://cs224d.stanford.edu/reports/PouransariHadi.pdf

[3] Victor Ginsburgh, “Awards, Success and Aesthetic Quality in the Arts”, Journal of Economic Perspectives—Volume 17, Number 2—Spring 2003—Pages 99–111

[4] Krauss, Jonas; Nann, Stefan; Simon, Daniel; Fischbach, Kai, 2008 “PREDICTING MOVIE SUCCESS AND ACADEMY AWARDS THROUGH SENTIMENT AND SOCIAL NETWORK ANALYSIS”, researchgate.net

[5] Michael T. Lash & Kang Zhao (2016) Early Predictions of Movie Success: The Who, What, and When of Profitability, Journal of Management Information Systems

[6] Mennatallah El-Assady, Wolfgang Jentner, Manuel Stein, Fabian Fischer, Tobias Schreck, and Daniel Keim, “Mennatallah El-Assady, Wolfgang Jentner, Manuel Stein, Fabian Fischer, Tobias Schreck, and Daniel Keim”.

[7] Piyush Gupta, Atul Sharma, Jitender Grover, “Rating based Mechanism to Contrast Abnormal Posts on Movies Reviews using MapReduce Paradigm”.

[8] Javaria Ahmad, Prakash Duraisamy, Amr Yousef, Bill Buckles, “Movie Success Prediction Using Data Mining”, IEEE – 40222.

[9] LEO BREIMAN(2001), Random Forests, Kluwer Academic Publishers.

[10] Rijul Dhir, Anand Raj,(2018) ” Movie Success Prediction using Machine Learning Algorithms and their Comparison “, First International Conference on Secure Cyber Computing and Communication (ICSCCC)

[11] Barry R. Litman, “Predicting Success of Theatrical Movies: An Empirical Study”.

[12] Byeng-Hee Chang & Eyun-Jung Ki (2005) Devising a Practical Model for Predicting Theatrical Movie Success: Focusing on the Experience Good Property, Journal of Media Economics, 18:4, 247-269, DOI: 10.1207/ s15327736me1804_2

[13] Richard Socher, Alex Perelygin, Jean Y.Wu, JasonChuang, ChristopherD. Manning, Andrew Y.Ngand Christopher Potts,” Recursive Deep Models for Semantic Compositionality Over a Sentiment Tree bank”

[14] Hitesh Parmar, Sanjay Bhanderi, Glory Shah, “Sentiment Mining of Movie Reviews using Random Forest with Tuned Hyperparameter”

[15] Xiaohui Yu, Member, IEEE, Yang Liu,Member, IEEE, Jimmy Xiangji Huang, Member, IEEE, and Aijun An, Member, IEEE, (2019), Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 4.