ENHANCING MOVIE MARKETING STRATEGY THROUGH PREDICTIVE ANALYSIS OF AGE-BASED VIEWERSHIP

Ch. Devi[1], B. Swanika[2], G. Lalithakumari[3], B. Sai Pravallika[4], G. Grace Sharoon[5]
[1]Assistant Professor, Department of Computer Science and Engineering, Vignan 's Institute of Engineering for Women, Visakhapatnam, India.
[2-5]Department of Computer Science and Engineering, Vignan 's Institute of Engineering for Women, Visakhapatnam, India.

ABSTRACT
The movie is one of the integral components of our everyday entertainment. One of the most prominent and rapidly expanding sectors in the world, the movie industry attracts viewers of all ages. Recent research has shown that only a small number of movies are commercially successful. Uncertainty in the sector has created immense pressure on the film production stakeholder. Movie producers and academics continue to believe that it is essential to have some expert systems that can reasonably estimate a movie's likelihood of success prior to its production. We proposed a system to predict the popularity of the upcoming movie among different audience groups and give advertisements according to the audience group. This would help the advertisement commercials and the TV channels to stream their advertisements. The audience has been divided into four different age groups: junior, teen, middle age, and senior. IMDb data that is available to the public was used in this study. This study highlights the potential of predictive and prescriptive data analytics in information systems to support industry decisions.
KEYWORDS: Audience groups, IMDb data, expert systems, information systems.

INTRODUCTION

Our research problem proposed a system to predict movie success at an early stage of movie production and performs movie Target audience prediction. In order to create the recommendation system, we merely took into account five major aspects of the movie, including its genre, cast, director, keywords, and description. The results of this work may lower the risk associated with the film industry. We have divided the popularity of a movie into six classes super-duper hit (SDH), super hit (SH), hit (H), above average (AA), average (A) and flop (F). The target audience estimation module is then built in the third and final phase. The audience has been divided into four different age groups: junior, teen, middle age, and senior. Our project is basically divided into four phases data collection, data cleaning, data visualization and model building. In the data collection process, we are collecting the data from publicly available Internet Movie Database (IMDb) data using web scraping. Web scraping is a computerised technique for gathering large amounts of data from websites.

In the second phase i.e., data cleaning in this process we got audience rating in the form html classes and the whole cast will be divided into actor 1, actor 2, actor 3 and actor 4; runtime of the movie gets converted min to hours; we have also cleaned the budget estimate and gross worldwide. For the target audience prediction first, we need the movie recommendation in which we will be using the collaborative filtering for the recommendation.

After the recommendation results, we will be doing the target audience prediction we will be using fuzzy k mean and the cosine similarity. We will be getting the rankings of target audience prediction among the four age groups.

LITERATURE SURVEY

In this section, we have given a detailed survey on the past related works.

[1] Movie Popularity and Target Audience Prediction Using the Content-Based Recommender System.

The recommendation system is primarily divided into three parts, collaborative filtering (CF), content-based filtering (CBF), and hybrid filtering. CF is a method that can refine things that a user could want based on answers by similar users. It searches a broad group of people and gets a smaller circle of users with tastes comparable to a particular user. It looks at the things they like and connects them to form a ranked list of suggestions. The recommended things must, in some application scenarios, be content-wise comparable to a reference item, for example, for suggestions of similar items. Also, content information allows the period of better descriptions, which is becoming frequently crucial in fair and open recommender systems. Systems that use content-based recommendations make use of the meta-data of visual or textual objects. Movie recommendation using CBF is one of the widely used research paradigms. A content-based movie recommender has been proposed where users with and movie features are used.

[2] Identifying Audience Attributes - Predicting Age, Gender and Personality for Enhanced Article Writing

We divide the related work into three categories of classification. Gender, age and personality. In field of machine learning, gender classification has been studied with various types of features, classifiers, and corpus.

All individuals have their own unique characteristics and thus shows a certain pattern of behaviour or a way of thinking. To predict author attributes, a model is built using learning algorithm based on a set of training data, each of which consists of a series of a features and labels.

Classifier

We will examine three different classifiers, Linear Regression, Naive Bayes and support Vector Machine (SVM).

- Feature

In the field of text classification, many features based on bag-of words model are widely used. Texts are broken into words and represented with words and their frequency disregarding word order and grammar.

- Corpus

We used 3,226 entries of blog posts collected from blogger.com in August 2004, 2,000 blog posts] from public repositories such as Netlog, and 2,469 texts produced by participants who took the Big5 test for gender, age and personality classification respectively.

[3] Predicting movie success with machine learning techniques: Ways to improve accuracy.

In their previous studies on predicting the box-office performance of a movie using machine learning techniques have shown practical levels of predictive accuracy. Their works are technically- and methodologically-oriented, focusing mainly on what algorithms are better at predicting the movie performance. However, the accuracy of prediction model can also be elevated by taking other perspectives such as introducing unexplored features that might be related to the prediction of the outcomes. In this paper, they examined multiple approaches to improve the performance of the prediction model. First, they developed and added a new feature derived from the theory of transmedia storytelling. Such theory-driven feature selection not only increases the forecast accuracy, but also enhances the interpretability of a prediction model. Second, they used an ensemble approach, which has rarely been adopted in the research on predicting box-office performance. As a result, their proposed model, Cinema Ensemble Model (CEM), outperforms the prediction models from their past studies that use machine learning algorithms. They suggested that CEM can be extensively used for industrial experts as a powerful tool for improving decision-making process.

[4] Movie Recommendation System Using Sentiment Analysis from Microblogging Data

- Collaborative, Content-Based, and Hybrid Filtering
- Various recommendation system approaches have been proposed in the literature for recommending items. Optimizing the recommendation system is an ill-posed problem. Research have proposed several optimization algorithms like Gray wolf optimization, artificial bee colony, particle swarm optimization and genetic algorithms. The combination of the hybrid cluster and optimization

technique showed better accuracy in movie prediction compared with movie prediction by the existing frameworks.

- Each user or product has a profile created by the content filtering method to describe who they are. Content-Based strategies requires gathering external information that might not be available or easy to collect. A content-based movie recommender has been proposed where users with and movie features are used. A content-based movie recommender system built a recommender system by taking into account many movie attributes, such as movie genre, actor and director names, and others. It has been suggested to create a hybrid movie recommendation system that combines collaborative filtering with sentiment analysis (CF).

A. Sentiment Analysis

Sentiment analysis is a technique to computationally identifying and categorizing people's opinions expressed in the form of reviews or survey is positive, negative, or neutral. Sentiment analysis has been used TextBlob1 library to calculate the polarity and subjectivity of the review sentences. Past research has primarily focused on analysing the user-generated textual reviews and categorized the user reviews into positive or negative classes. Online evaluations now frequently include slang, emotions, and other everyday language to better capture readers' opinions. Hutto and Gilbert proposed a valence-aware dictionary and sentiment reasoner (VADER) algorithm that is used to parse the user reviews and analyse them using a rule-based model to calculate the sentiment score of the tweets. To better capture readers' opinions, online reviews increasingly regularly use slang, emoticons, and other common terminology. The VADER method's results outperformed other sentiment analysis methods in terms of performance. In this methodology, sentiment analysis was also applied to posts on microblogging sites.

[5] Weight based movie recommendation system using k-means algorithm.

The movie recommendation system that has been developed depends on five movies attributes to make a recommendation. It does so by balancing the five attributes in a proper way and then providing the user with a recommendation. In this recommender system could be used in two different approaches. Either a user will go to a web page and will input some attributes like the genre, actor, year and rating. It means that the type of genre that the user likes, or the actor that they prefer, or the year which they would like and the rating of the movie they would want to see, will be input by them. Data pre-processing is a data mining technique that involves transforming raw data into an understandable format.

PROPOSED SYSTEM

- In our proposed work, at first the data is collected from the publicly available Internet Movie Database (IMDb) data using Web Scraping. Let us now see the algorithm for the data collection that is nothing but the Web Scraping.

Web Scraping with Python: BeautifulSoup Library

Steps involved in web scraping:

1.Find the URL of the webpage that you want to scrape.

2.Select the particular elements by inspecting.

3.Write the code to get the content of the selected elements.

4.Store the data in the required format.

The popular libraries/tools used for web scraping are:

1.Selenium – a framework for testing web applications.

2.BeautifulSoup – Python library for getting data out of HTML, XML, and other markup languages.

3.Pandas – Python library for data manipulation and analysis.

Steps involved in web scraping:

Step 1: Find the URL of the webpage that you want to scrape.

Step 1.1: Defining the Base URL, Query parameters.

Step 2: Select the elements by inspecting.

Step 3: Write the code to get the content of the selected elements.

Step 4: Store the data in the required format.
Step 4.1 Create a dictionary format with column name as keys and column values as list which we scrape.
Step 4.2 storing to a pandas data frame.
Step 4.3 Writing the content of the data frame to a CSV file.
Therefore, this is the first step of our project which is completed.
Now let us see the proposed system architecture.

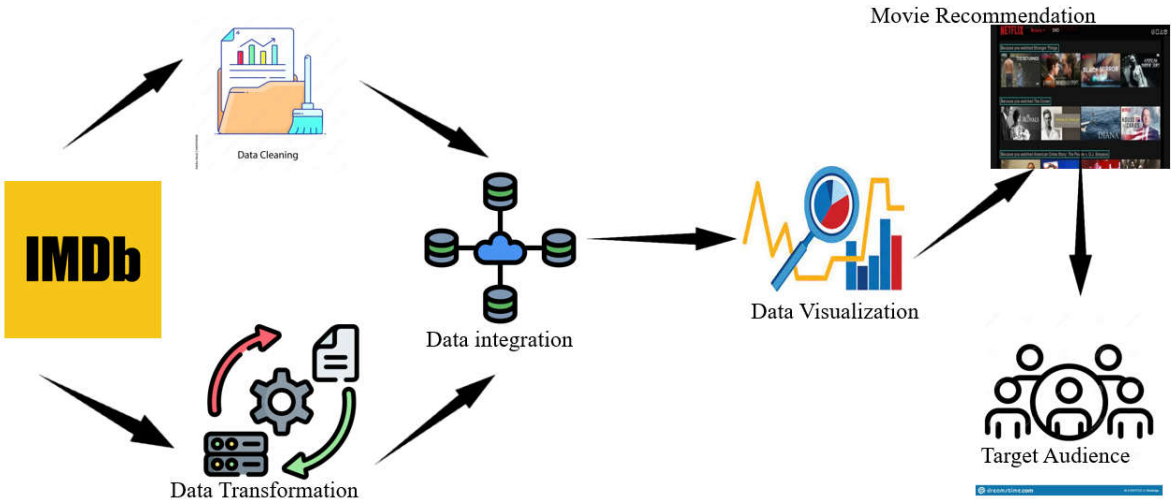


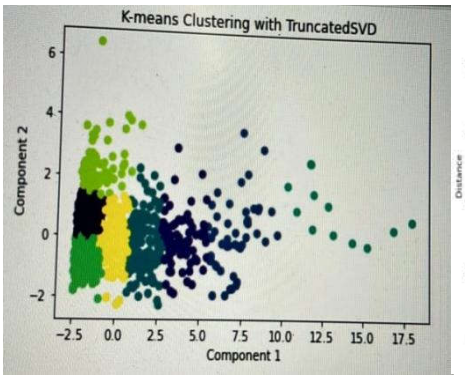
FIGURE 1. Proposed System Architecture

- As we have already seen the steps of data collection. So, the next step is data cleaning and data transformation. By comparing with the raw data that is collected we transformed some fields like release year where we have the data in the form of brackets for example – (1994) which is transformed into just 1994. The next field is audience rating which is in the form of html format which is transformed into the only rating.
- The next which is transformed is runtime in the data collection dataset it is in the form of minutes but in data transformation we have changed in the form of hours and minutes. The next field is votes and cast in the votes we have inserted commas in the entries and in the cast, they are divided into actor 1, actor 2, actor 3 and actor 4.
- Then we have taken another dataset of user reviews in which we have fields regarding age, budget estimated and Gross worldwide. The next step is the data integration in which we have combined everything that is the data cleaned dataset and user reviews dataset. For the better understanding the next step is the data visualization that is we can compare between various data in the form of visualization which would be easier for the better understanding.
- The next module is the movie recommendation for getting those recommendations we are using the collaborative filtering first based on those results we can use content-based filtering too. What is extra we have added from the existing system is we have specified even more in the form of each age group as male and female. After the movie recommendation is done the next module is the target audience prediction in which we would give the rankings of each group in the highest to lowest and the category of the advertisements which would be displayed in the output. We have done the comparative analysis of k means, hierarchy clustering, fuzzy c means, Agglomerative, DBSCAN and Random Forest Classifier. As we have eight clusters and applied thesealgorithms, and we got the accuracy and their results. By comparing the results, we have identified that k means got the better results with less cluster overlapping.

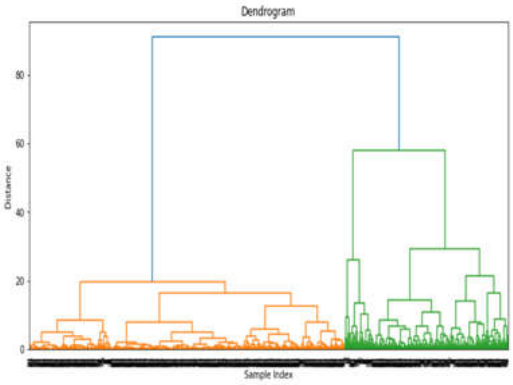
So, these are the accuracy results and the graphs after applying these algorithms.
Table 1. Compare different machine learning model using several evolution parameters.

Machine Learning Model	Accuracy out of 0.5	Accuracy (in %)
K Means	0.36	72%
Fuzzy C Means	0.22	44%
Hierarchy Clustering	0.22	44%
Agglomerative	0.21	42%
DBSCAN	0.08	16%

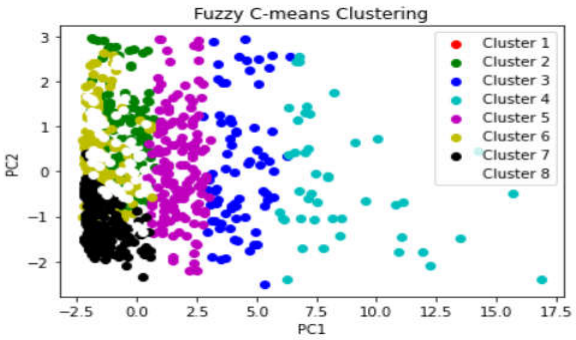
The mean silhouette over 0.5 is considered as a “good” clustering solution. By comparing all of these algorithms we are not getting the output as we wanted.
So, the output which we wanted is the top 3 age groups of the particular movie which we have entered.



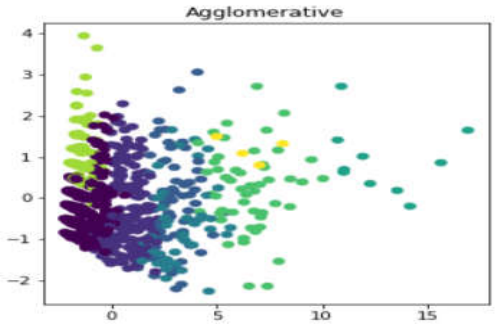
A) K Means



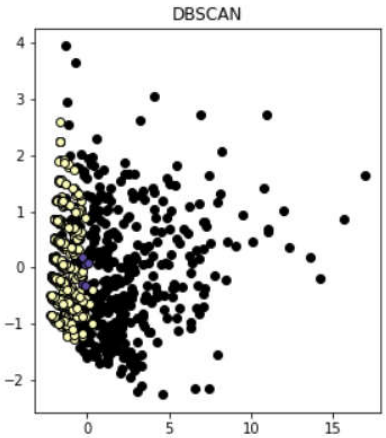
C) Hierarchy Clustering



B) Fuzzy C Means



D) Agglomerative



E) DBSCAN

FIGURE 2. Comparative results of all algorithms

Random Forest Algorithm

Random Forest is a popular machine learning algorithm used for both classification and regression tasks. It is an ensemble method that combines multiple decision trees and makes predictions by taking the average of the predictions made by each individual tree.

The Random Forest algorithm works by creating a large number of decision trees, each of which is trained on a random subset of the training data and a random subset of the features. This randomness helps to reduce overfitting and improve the generalization of the model.

During the prediction phase, each decision tree in the Random Forest makes a prediction, and the final prediction is obtained by taking the average of these predictions. This averaging process helps to smooth out the predictions and reduce the impact of individual noisy or biased decision trees

CONCLUSION

A substantial amount of financing is consumed in every box-office movie. However, most movies fail to achieve success. Earlier, the most significant number of works have been done on post-production or post-release forecast. The estimate does not influence as the investor has already consumed their funds on the film production. The pre-production or early production stage forecast needs high accuracy and the best time to ensure investment. Our approach to focused movie popularity and finding out the target audience of an upcoming movie is very much unique. In our approach, we have used a recommendation system to find similar movies from a given movie and use similar movies for forecasting purposes. We have divided the audience group into four groups and even we divided the audience group in gender wise. According to the given recommendation movie we will be giving the rankings of each given age group and according to that we will be giving the advertisements according to the highest ranging in the target audience group. The proposed system is an excellent tool for the movie industry. The audience group could be divided according to age and according to the demography or profession of the audience. That will be much easier for targeting and promoting an upcoming movie.

REFERENCES

1. Sandipan Sahu 1, Raghvendra Kumari, Mohd Shafi Pathan2, Jana Shafi 3, Yogesh Kumar 4, and Muhammad Fazal Ijaz5, (Member, IEEE) “Movie Popularity and Target Audience Prediction Using the Content-Based Recommender System”. Received March 14, 2022, accepted April 7, 2022, date of publication April 18, 2022, date of current version April 26, 2022.
2. Raad Bin Tareaf, Philipp Berger, Patrick Hennig, Christoph Meinel, “Identifying Audience Attributes - Predicting Age, Gender and Personality for Enhanced Article Writing”. All content following this page was uploaded by Raad Bin Tareaf on 15 June 2018. DOI: 10.1145/3141128.3141129.
3. K. Lee, J. Park, I. Kim, and Y. Choi, ‘‘Predicting movie success with machine learning techniques: Ways to improve accuracy,’’ Inf. Syst. Frontiers, vol. 20, no. 3, pp. 577–588, Jun. 2018.
4. S. Kumar, K. De, and P. P. Roy, ‘‘Movie recommendation system using sentiment analysis from microblogging data,’’ IEEE Trans. Computat. Social Syst., vol. 7, no. 4, pp. 915–923, Aug. 2020.
5. Md. Tayeb Himel, Mohammed Nazim Uddin, Mohammad Arif Hossain, and Yeong Min Jang, “Weight based movie recommendation system using k-means algorithm,” 978-1-5090-4032-2/17 , 2017 IEEE.
6. F. Gedikli, D. Jannach, and M. Ge, ‘‘How should i explain? A comparison of different explanation types for recommender systems,’’ Int. J. Hum.- Comput. Stud., vol. 72, no. 4, pp. 367–382, Apr. 2014.
7. Y. Yao and F. M. Harper, ‘‘Judging similarity: A user-centric study of related item recommendations,’’ in Proc. 12th ACM Conf. Recommender Syst., Sep. 2018, pp. 288–296.
8. T. K. Paradarami, N. D. Bastian, and J. L. Wightman, ‘‘A hybrid recommender system using artificial neural networks,’’ Expert Syst. Appl., vol. 83, pp. 300–313, Oct. 2017.

9. D. Wang, Y. Liang, D. Xu, X. Feng, and R. Guan, ‘‘A content-based recommender system for computer science publications,’’ Knowl.-Based Syst., vol. 157, pp. 1–9, Oct. 2018.
10. P. B. Thorat, R. M. Goudar, and S. Barve, ‘‘Survey on collaborative filtering, content-based filtering and hybrid recommendation system,’’ Int. J. Comput. Appl., vol. 110, no. 4, pp. 31–36, Jan. 2015.
11. S. Sivakumar and R. Rajalakshmi, ‘‘Analysis of sentiment on movie reviews using word embedding self-attentive LSTM,’’ Int. J. Ambient Comput. Intell., vol. 12, no. 2, pp. 33–52, Apr. 2021.
12. S. M. R. Abidi, Y. Xu, J. Ni, X. Wang, and W. Zhang, ‘‘Popularity prediction of movies: From statistical modelling to machine learning techniques,’’ Multimedia Tools Appl., vol. 79, nos. 47–48, pp. 35583–35617, Dec. 2020.
13. M. T. Lash and K. Zhao, ‘‘Early predictions of movie success: The who, what, and when of profitability,’’ J. Manage. Inf. Syst., vol. 33, no. 3, pp. 874–903, Jul. 2016.
14. X. Su and T. M. Khoshgoftaar, ‘‘A survey of collaborative filtering techniques,’’ in Advances in Artificial Intelligence. London, U.K., 2009.
15. P. Lops, M. Gemmis, and G. Semeraro, ‘‘Content-based recommender systems: State of the art and trends,’’ in Recommender Systems Handbook. Boston, MA, USA: Springer, 2011, pp. 73–105.