iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** IV **Month of publication:** April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4799

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Flick Fortune: Movie Gross Prediction using Data Analysis and Prediction

V. Balamurugan[1]
[1]*Assistant Professor, Department of Computer Science and Engineering, S.A. Engineering College, Chennai, India.*

Abstract*: *Data Mining is one of the rapid growing areas in the development of Industrial automation. Many researches are carried out in order to determine the future result of any business such as stock market. The appropriate data mining technique helps the decision makers to arrive upon a valid decision. Our proposed system is the implementation of probabilistic prediction of data from the extensive analysis. We coherently apply the proposed system onto the movie prediction such as Gross collection, Rating, feedback and success ratio. We use web scrapper code with TMDB API to fetch attributes of the movies such as actor name, director, musician, producer and budget. We employ association rule mining to extract the probabilistic prediction from the user query. We further use AJAX for efficient and optimistic data traffic between the client and the server. Experimental results show that our approach is best performing the data analysis under several test strategies
Keywords: Data Analysis, Prediction, Data mining, Association rule mining, AJAX.

I. INTRODUCTION

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data [1]. One of the most important applications of data mining is the analysis of transactional data. Capturing the co-occurrence of items in transactions was first proposed by Agrawal et al. [2]. For example, given a transactional database of a supermarket, we may have {milk,bread} bought together with support of 20%. It means that 20% of all transactions contain milk and bread together. Databases which originate from transactions in a supermarket, bank, department stores and, etc., are all inherently related to time. These are called temporal databases which are databases that contain time-stamping information [3]. One important extension to frequent pattern mining is to include a temporal dimension [4]. For example, milk and bread may be ordered together in 80% of all transactions between 7 and 9 A.M. while their support in the whole database is 20%. In fact, interesting patterns are often related to the specific period of time; therefore, the time during which they can be observed is important.

From the above, we may conclude that different patterns can be discovered if different time intervals are considered. Discovering such patterns that are held during the time intervals may lead to useful knowledge. The discovery of such association rules has been discussed in the literature. In this context, sequential association rules [5], time intervals for association rules [6], [7] and calendar-based association rules [4], [8] are some interesting studies in recent years.

For this work, data mining process was used to extract patterns and trends which can be beneficial in predicting movies success. The data mining techniques were applied to a movie database, but before the mining techniques could be used, the data went through the cleaning and integration process. Data mining deals with discovering trends and patterns in a given data [1].

Due to the powerful data mining techniques and predictions, this approach was used for movie success prediction. Movie success prediction is important because it involved significant time and investment. For this reason, it is important for the shareholders to have fewer uncertainties involved. They can achieve this very well using data mining techniques. Movie success predictions, trends and variable dependence can very well be determined using data mining. Movie success prediction is also significant for the movie watchers who need to know in advance the quality and success rating of a movie before monetary resources can be utilized for a movie.

We have provided a useful model in this study which can lower chance of failure and can provide the stakeholders with confidence and a visible prediction of success. There are various variables which were studied to provide a movie success prediction. Some of these variables included budget, actors, director, producer, set locations, story writer, movie release day, competing movie releases at the same time, music, release location and target audience. The goal of this mathematical model is to provide a precise prediction of success, hence providing confidence to stakeholders in their investments.

II. REVIEW OF THE LITERATURE

In 2004, Saraee, White and Eccleston performed analysis of online movie resource of over 390,000 movies and television shows [2]. In 2006, Sharda and Delen worked with predicting financial success of movies even before the movie is released [3]. Classification

approach is used where the movies were categorized from flop to blockbuster. Facts and relationships among alternatives can be made by making use of data mining. Some of the factors considered were movie budget and movie popularity relationship, movie cast and movie success relationship. This work helped discover important findings. However, due to copy right, there was a challenge involved accessing the data. In 2009, Zhang and Skeina worked on utilizing news analysis to make movie predictions [4]. It was determined that using news data resulted in performance as good as using the IMDB data. Even better performance was achieved using both IMBD data and news data. In 2010, Asur and Huberman worked on predicting outcomes based on social media content [5]. The movie success prediction was based on social media success count, and historical data. The predictions can be made about new movies using this study. However, success prediction cannot be made before the movie is released. In 2015, Lash and Zhao proposed a way to predict decisions about movie investments [6]. This work provided help with investment decision making early in movie production. Historical data was utilized for this work. Some of the features of this work were matching ”who” with ”what” and ”when” with ”what”. The profit was calculated mainly based on box office revenue. However, for many movies, there are other sources of revenue, for example, merchandise.

Recommendation system is utilized for contents such as news, articles, movies, books and music. After the emergence of e-commerce, recommendation system development, predicting preferred manufactures and offering useful information, is fashionable to promote consumption [1]. For representative examples, Amazon and Netflix are related to e-commerce field. Manufactures are recommended with satisfaction in reasonable time by their own contents recommendation system [2]. For recommendation system, data-mining techniques that are included pattern recognition and information filtering methods have been applied to develop recommendation system. Information filtering methods are mostly researched among related methods. The research direction of recommendation system is divided by content-based recommendation and collaborative filtering [3][4]. And, it is easy excessively to characterize and restrict if system is usually following similar item with user preference [5]-[7]. On the other hand, collaborative filtering is predicting user preference by analysis of related user’s propensity. Related user’s propensity is obtained to analyse other user’s information based on similarity to target user. In this paper, data about movies and user is offered by Netflix which is the company renting movies by online system. And, we propose method which is predicting movie rating about user. Prediction system about movie rating is proposed by analysis of personal propensity.

III. PROBLEM DEFINITION AND MOTIVATION

To develop a probabilistic model that yields prediction and analysis of the Flicks across the globe, also produces different representational patterns from their attributes. Proper data mining approach should be integrated with it. TMDB API with AJAX should be incorporated. IMDb is a popular website that provides rating of a movie or a tele-serial across the globe. But it doesn’t have any provision in data representation, analytics and prediction for the desired user query. There is no system available that takes input query w.r.to Movies/tele-serials from the user and processes it using proper data mining techniques and represent the desired pattern with probabilistic model such as Hidden Markovian Model (HMM). The system would be useful to society in such a way that, if applied to different problematic areas, it would predict the expected result using proper data mining approaches such as Association Rule mining with graphical representation, provided the corresponding input query.

In our work, we have developed a mathematical model which is used to predict the success and failure of upcoming movies depending on certain criteria. Our work provides advantage in that strong correlations were found between different criteria and movie success rating. Unlike the related work discussed, our work can be used to predict movie success even before it is released. Our work makes use of historical data in order to successfully predict the ratings of movies to be released.

IV. PROPOSED MODEL AND IMPLEMENTATION

A. Data pre-processing

The analysis method for personal propensity is usually normalizing personal information such as age, sex and occupation. This kind of method can be caused security problems by personal information. And user’s satisfactions are lower than expected because user’s data also wasn’t perfectly matched with personal taste and individuality. We were developing movie rating prediction system based on personal propensity with previous watched movie records for user satisfaction. The structure of proposed system is shown in Figure. 1. All data is consisted of text-type. Data from the Netflix is evaluated information of 480,000 users about 17,770 movies. Rating points are from 1 to 5. And probe and qualifying data are offered to evaluate system performance.

Training data is information of movie evaluation and base information of movie rating prediction system. But training data isn’t structurally suited to analyze personal propensity because it is classified by movie. So, we can reorganize classification by user as a pre-processing. Reorganization of training data based on user is shown in Fig. 2. The mv_i is expressed as Movie ID and i’s range is

from 1 to 17,770. We can know about evaluation of each user through the 17,770 movie files. The u_k is expressed as User ID and k's range is from 1 to 2,649,429 and 480,189 users are distributed among this range. Probe and qualifying data among data from Netflix are offered to evaluate system performance. Evaluation items of data, target user's rating point about target movie, are predicted.

Information of target user and movie through probe data are extracted. And filtering process is used to identify reliability of user information based on watched movie rating point about target user. Filtered users are evaluated by means of mean value of each user's movie rating records and unfiltered user are evaluated by means of collaborative filtering to analyze personal propensity. Related user group is created by similarity with target user. Exactly, Target user's rating point about target movie is predicted by similarity calculation of the two. And fuzzy inference process is executed to make up for the collaborative filtering.

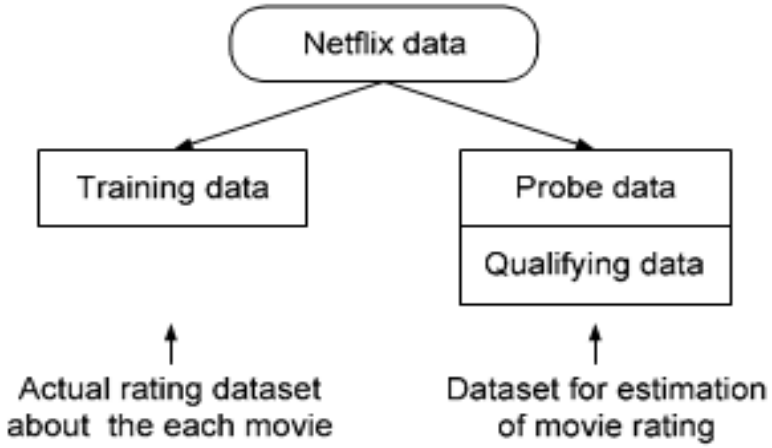


Fig. 1. Data structure by Netflix

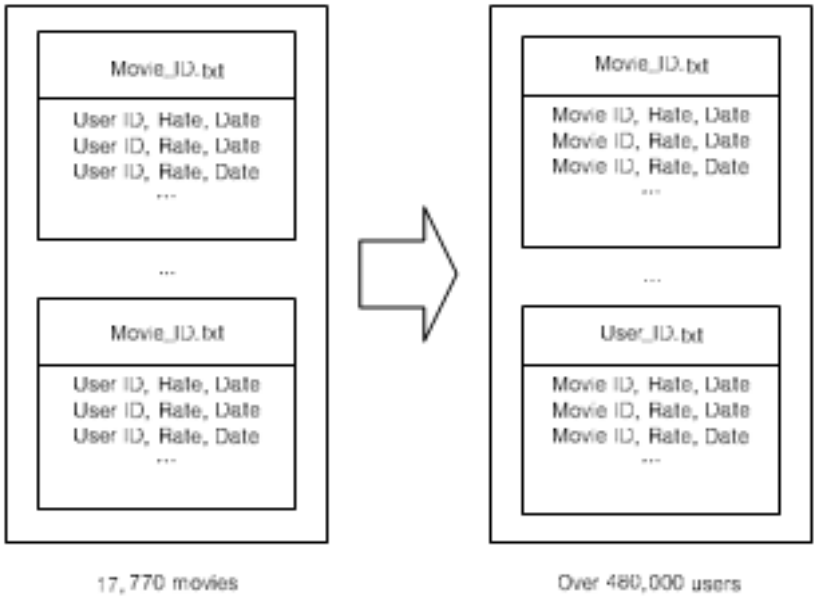


Figure 2. Data set processing

B. The FlickFortune Algorithm

Clean, integrate and transform simulation data. Find X2 analysis between genres and ratings of the movies. Find X2 analysis between movie actors and movie ratings. Find X2 analysis between movie actors and movie genres. Find the correlations from the respective X2 analyses above. Predict success rating from the correlations between various movie criteria.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

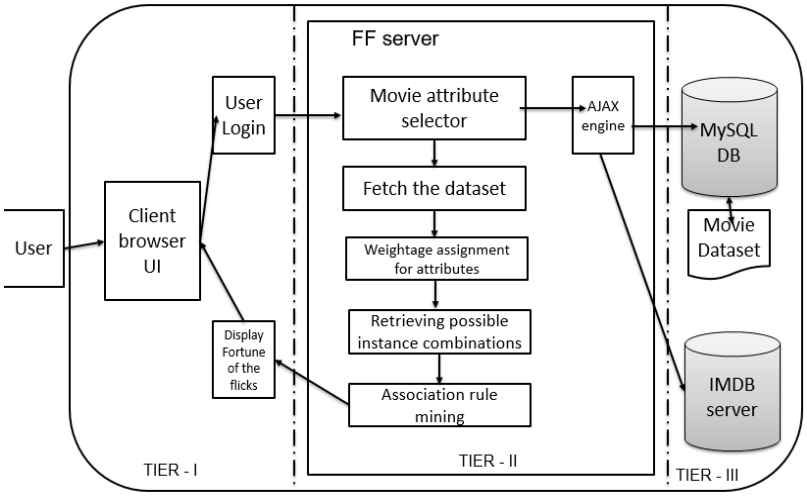


Figure 3 FlickFortune Architecture

Information of target user and movie through probe data are extracted. And filtering process is used to identify re- liability of user information based on watched movie rating point about target user. Filtered users are evaluated by means of mean value of each user's movie rating records and unfiltered user are evaluated by means of collaborative filtering to analyze personal propensity. Related user group is created by similarity with target user. Exactly, Target user's rating point about target movie is predicted by similarity calculation of the two. And fuzzy inference process is executed to make up for the collaborative filtering.

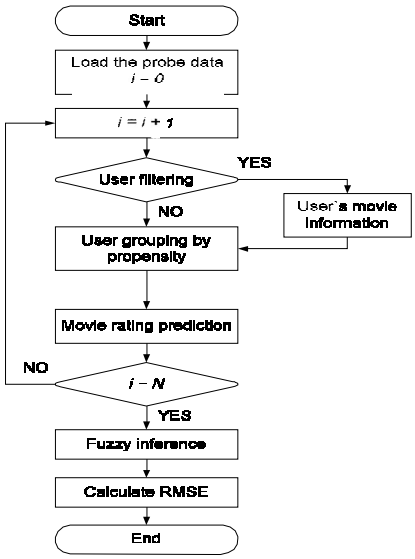


Figure 4.Flow chart of Flick Fortune

C. User Filtering

Reliability of watched movie information is very important for this research. But meaningless information is existed in the training data. For example, a user rates points about 17635 movies per day in the training data from Netflix. This data can be easily affecting to system performance degradation. So, user filtering is required to have reliability of related user group. In this paper, there are two conditions to be filtered. First, the users who is exceeded the mean number of rated movies per day are filtered, that is, users who rated over 200 movies per day are filtered.

$$portion = \frac{Max(n_{u,d})}{n_{u,i}}$$

nu, i is the total rated number of movie (i) by user (i), nu, d is the rated number of movies by user per day. Personal propensity about filtered user is analyzed by means of previous movie information without comparison with other user's propensity. That is, the mean value about all rated movies by user is used to analyze propensity.

D. Personal propensity Based User Grouping

Unfiltered users are subject to member of related user group to analyze personal propensity. Related user group is organized as user-movie profile, Fig. 4. There is user's rate about movie item in the user-movie profile and filled blanks present watched movie item. Movie i's range is from 1 to 17,770 and user k's range is from 1 to 2,649,429. Related user group is created by movie rating analysis between target user and related users through the user-movie profile. Target user is u_3 and target movie is mv_4. Related users are selected by similarity with target user among watched information by target user. Related users are u_1, u_4 and u_k-1.

The version of this template is V2. Most of the formatting instructions in this document have been compiled by Causal Productions from the IEEE LaTeX style files. Causal Productions offers both A4 templates and US Letter templates for LaTeX and Microsoft Word. The LaTeX templates depend on the official IEEEtran.cls and IEEEtran.bst files, whereas the Microsoft Word templates are self-contained. Causal Productions has used its best efforts to ensure that the templates have the same appearance.

Causal Productions permits the distribution and revision of these templates on the condition that Causal Productions is credited in the revised template as follows: "original version of this template was provided by courtesy of Causal Productions (www.causalproductions.com)".

V. EXPERIMENTAL RESULTS

Simulation data was used for the movie database. The data was cleaned, integrated, and transformed before the data mining techniques were applied. The proposed system implements the association rule mining to find the prediction from the given dataset. The web service is must to fetch the movie dataset from the IMDB website, even though we have the dataset of downloaded movie data so far. The IMDB API helps fetching the movie data from the IMDB server. The training dataset is preprocessed, i.e., 6000 records were found in the server, after preprocessing it has become 4900 records. AJAX is implemented to efficiently retrieve data from the server in order to maintain the latency at both the ends (client and server). The home page of the project contains different attributes of a movie, displayed in drop down box, which has to be populated from the IMDB server through AJAX web service. Asynchronous requests are made from the project to the IMDB server to fetch movie attributes such as actors, directors, etc. Upon the display of all attributes, the user will choose the required combination of attributes. The dataset is preprocessed and then displayed in the web browser. AJAX asynchronous task runs in the background that retrieves the movie attributes such as actors, directors, album image, genre, gross, budget, etc. Since Javascript takes care of client side validation and displaying, the home page ets displayed in elegant manner. The IMDB (International movie Database) server contains all the attributes of all the movies of about 8000 flicks. Web scrapper code scraps the data from the IMDB server in AJAX and stores in MySQL server. Once the possible set of attributes are retrieved from the IMDB server, our system predicts the fortune of the given combination of attributes such as profit/loss, gross, budget, etc. We perform pre-processing of dataset obtained from the IMDB server and assign a weightage value for each attributes. Depending upon the attribute that user selected, each attribute will be assigned a score (weightage). Perform association rule mining for the possible set of attributes with values preprocessed. The average values of all gross of actors separately, to display when the possible combination is not found. After the possible attributes and their values are retrieved, they will be mined and the result will be displayed for predicted gross, etc. The graphical representation of year wise actors' performance, actor wise performance will be done with a graph using Graphics library in PHP. Whenever a new attribute is selected, the corresponding result will be displayed by fetching the data from the IMDB server and the values are computed and displayed in browser. AJAX asynchronous background task takes responsibility of making the web pages more dynamic and data availability is more. The system is more efficient in terms of network parameters such as latency, bandwidth, etc.

In this study, the mathematical model developed to predict the success and failure of the upcoming movies involved finding correlation between various attributes using X2 analysis. Correlation is a measure of dependence between two variables. The correlation can be negative or positive. A positive correlation indicated that the two variables increase or decrease in parallel, whereas, a negative correlation indicated that the two variables change in opposite directions.

A. X2 Analysis: Genres vs. RATINGS

Correlation between genres and ratings was analyzed first. The X2 results are shown in the table below.

TABLE I x^2 analysis: genres vs. Ratings

Ratings/Genres	Romance	Comedy	Other	Total
1	2(9.8)	0(4.8)	16(3.4)	18
2	0(0)	0(0)	0(0)	0
3	3(10.8)	16(5.33)	1(3.7)	20
4	19(12.5)	4(6.13)	0(4.34)	23
5	25(15.8)	4(7.7)	0(5.47)	29
Total	49	24	17	90

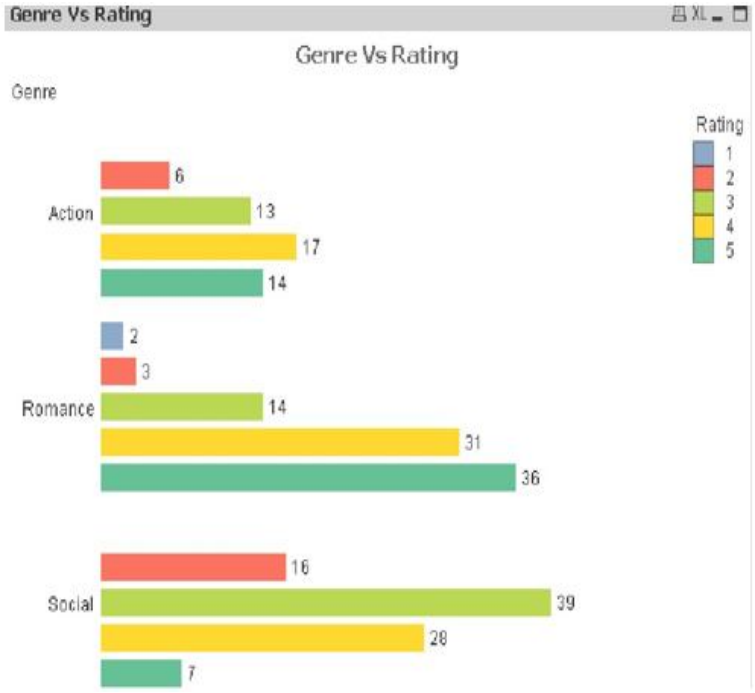


Figure 5. Genre vs rating

Expected frequencies are calculated as Count(Genres)xCount(Ratings)/n
X2 = 6.2 + 5.6 + 3.38 + 5.35 + 4.8 + 21.3 + 0.74 + 1.81 + 3.4 + 2.1 + 4.34 + 5.47 = 64.39.
Degrees of freedom = (4)(2) = 8. From observing the chi-square table, the p valve is very low, so we can reject the null hypothesis that ratings and genres are independent and conclude that the two attributes are strongly correlated. This means that movie genres are predicted to have specific ratings.
As shown by the bar graph below, most of the romance movies have the rating of 5, most of the action movies have rating 4, and most of the social movies have rating 3.

B. X2 analysis: Actors vs. Ratings
Correlation between actors and ratings was analyzed next. The X2 results are shown in the table below.

TABLE II
X^2 ANALYSIS: ACTORS VS. RATINGS

Ratings/Actors	Shahrukh	Rajni	Total
≤>33	1(4.2)	5(1.8)	6
	13(9.8)	1(4.2)	14
Total	14	6	20

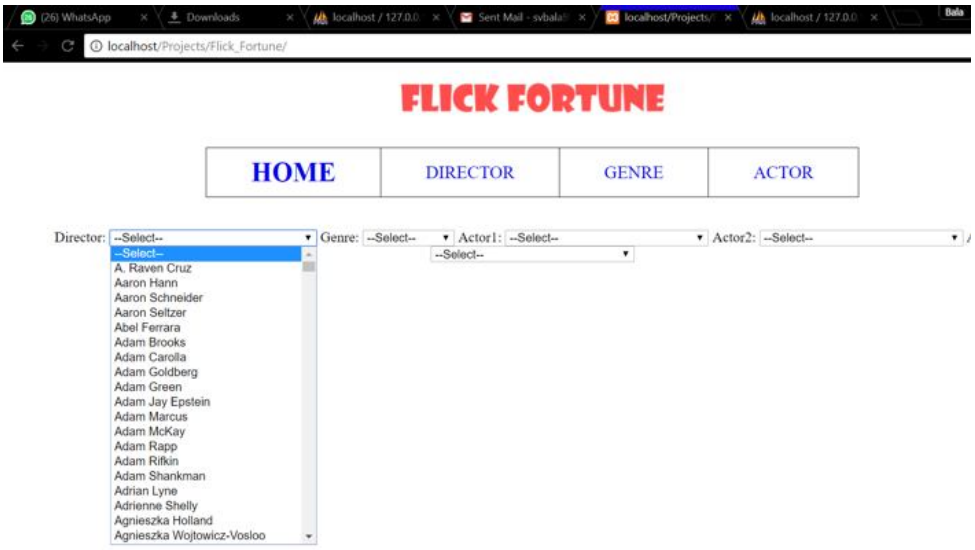


Figure 6. Home page

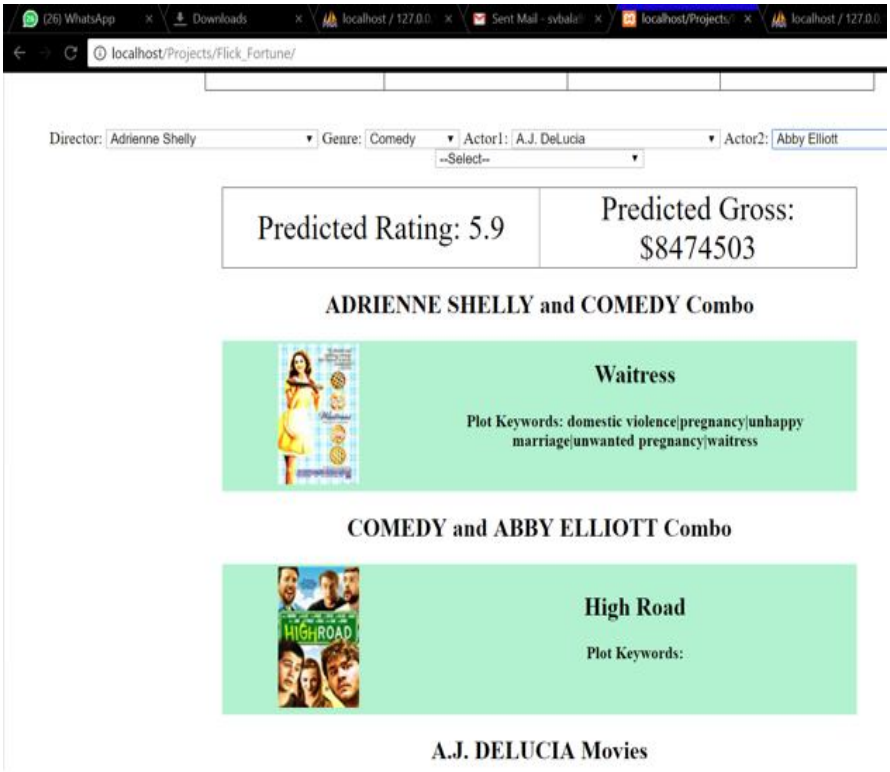


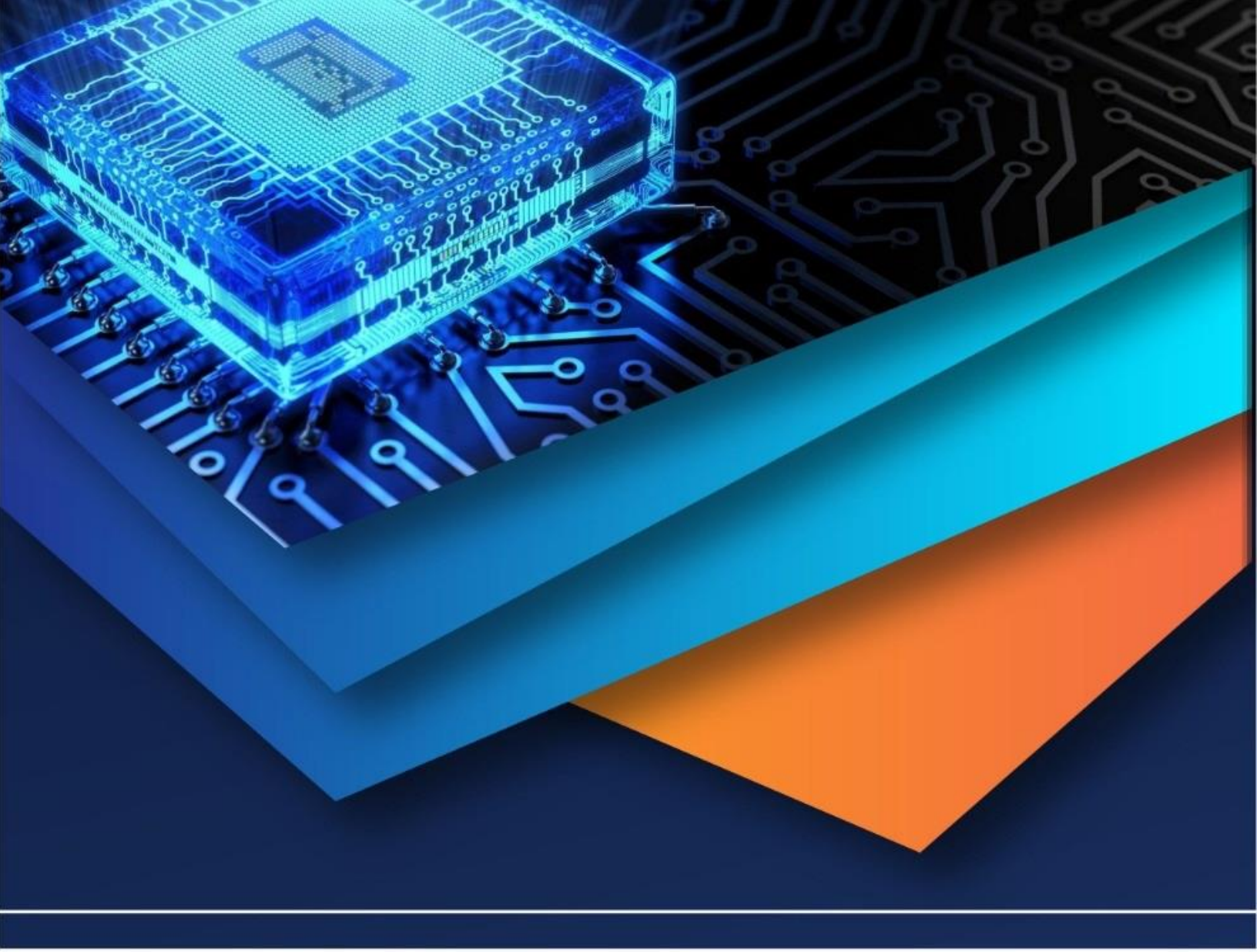
Figure 7. Prediction analysis

VI. CONCLUSION AND FUTURE ENHANCEMENT

FlickFortune is a novel framework in predicting the results from the training data set. There are many areas that need the prediction technique efficiently in order to make decisions in prior. Our proposed system helps in such cases, provided the input training dataset is error free and pre-processed. The association rule mining and FP growth tree are helpful in mining frequent patterns. Therefore our proposed system will be helpful for those who need prediction system with accuracy. The proposed work frames a mathematical model with data mining for prediction from the input data set. This model can be employed for all the societal needs where prediction on data is required. In future work, the big data analytics will be brought into picture in order to process huge volumes of data.

REFERENCES

[1] Abdulsalam, Shailendra Singh, Atif Alamri, "Mining Human Activity Patterns from smart home big data for health care applications", 27.7.2017, Advances of multi sensory technologies and services for healthcare in smart cities, IEEE ACCESS, Volume 5, 2017.
[2] Archana Gahlaut, Tushar, Prince Kumar Singh, "Prediction analysis of risky credit using Data mining classification models", 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017, IEEE.
[3] Fen Miao, Nan Fu, Yuan Ting, Zhang, "A novel continuous blood pressure estimation approach based on data mining techniques", IEEE journal of biomedical and health informatics, IEEE Access, Volume 21, No.6, NOV 2017, pp-1730 – 1741
[4] Gilang Firmanuddin, Suhono H. Supangkat, "City analytic development for modeling population using data analysis prediction", International Conference on ICT For Smart Society (ICISS), 2016 IEEE.
[5] Huigui Rong, Zepeng Wang, Hui Zheng, Chunhua Hu, "Mining efficient taxi operation strategies from large scale geo-location data", 20.7.2017, IEEE ACCESS 2017 special edition for Intelligent systems for IoT.
[6] Kren, Kos, Zhang, 2017, "Public Interest Analysis Based on Implicit Feedback of IPTV Users", IEEE Transactions on Industry Informatics, Volume: 13, Issue: 4, Aug. 2017, Date of Publication: 18 April 2017, DOI: 10.1109/TII.2017.2695371
[7] Lic. Lilian Judith, "Machine learning algorythms for analysis and data prediction", IEEE 37th Central America and Panama Convention (CONCAPAN XXXVll), 2017.
[8] Lilli Tong, Viting Wang, Fan Wen Xiaowen Li, "The research of customer loyalty improvement in telecom industry based on NPS data mining", IEEE ACCESS on network and security, 2017 – pp.260 – 268.
[9] Mazher Ghorbani, Masoud Abessi, "A new methodology for mining frequent itemset on temporal data", IEEE transactions on engineering management, Volume 64, NOV 2017, pp-566 – 578
[10] Po-Yuan Yang , Jinn-Tsong Tsai, Jyh-Horng Chou, "Prediction analysis of oxygen content in the water for the fish farm in southern Taiwan, "International Conference on System Science and Engineering, 2017.
[11] Praman Deep Singh, Anuradha Chug, "Software defect prediction analysis using machine learning algorithms", 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence, 2017 IEEE.
[12] Qing Yang , Guoqiang Ji , Wang Zhou, "The correlation analysis and prediction between mobile phone users complaints and telecom equipment failures under big data environments", 2nd International Conference on Advanced Robotics and Mechatronics (ICARM), 2017 IEEE.
[13] Shen Ren, Lin Han, Zengxiang Li, Bharadwaj Veeravalli, "Spatial-temporal traffic speed bands data analysis and prediction", IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2017 IEEE.
[14] Weijing Qi, Qingyang Song, Xiajoe Wang, "trajectory data mining based routing in DTN – enabled VANETs", IEEE Access on Security and privacy for vehicular networks, October 2017, Volume 5, 2017, pp.24128 – 24138
[15] Zhiquiang Ge, Zhihuan Song, Steven Ding, Biao Huang, "Data mining and analytics in the process industry – Role of machine learning", IEEE Access on data driven monitoring, fault diagnosis and control of cyber physical systems, September 26, 2017, pp-20590 – 206



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY