Movie Success Prediction using Data Mining

Conference Paper · October 2019

CITATIONS
0

READS
1,258

1 author:

Ankit Kharb
VIT University
3 PUBLICATIONS **1** CITATION

SEE PROFILE

MOVIE SUCCESS PREDICTION USING DATA MINING

AMAN CHAUHAN(19MCA0142), DEVESH KUMAR(19MCA0116), ANKIT(19MCA0071)

Abstract

In this project for Movie Success Analysis and Prediction we apply data mining technique using **R** software to predict success or failure of a movie based on several attributes, and criteria on which the success of any movie can depend. The proposed work aims to develop a system based on data mining techniques that will help in predicting the success or failure of a movie thereby reducing certain level of uncertainty of future of a movie . It can be done by using past available dataset by using several sources such as IMDB database that can be downloaded from internet. An attempt is made to predict the future of movie for the purpose of business certainty or simply a theoretical condition in which decision making (success of movie) is without risk, because the decision makers has all the information about the exact outcome of the decision, before he or she makes the decision .Some criteria that help us to calculate movie success or failure includes budget, actors, writer, movie releasing date, competing the movie releasing at the same time, music, audience interest. We know making of any movie involves a huge amount of investment thus movie prediction can help and can play a very important role in the movie industry. For this project we gather a series of interesting facts and relationships using various techniques of data mining. In particular, we concentrate on some attributes relevant to the success prediction of movies, such as whether any particular actors or actresses are likely to help a movie to succeed, and we also use many other attributes and each attribute has some criteria on the basis of weightage has been given and then any prediction is made based on that records.

Keywords: Data Mining, Movie, R Software

Introduction

Movies are the best way of for entertaining people for a long time. It is the best way to enjoy but everything depend upon the type of movies and their success rate. That's why only few movies get success and higher ratings among other movies produced by the industry in a year. Movie revenue is totally based on various components of the movies such as the actors working in the movie, past success of those actors, film critics review, rating for the movie, release year, time, date etc. of the movie. We don't have any formula that can help us to provide analysis for predicting how much revenue a particular movie is generating because of these multiple components and criteria. But by analyzing the revenues generated by previous movies, a model can be built which can act helpful to predict the expected revenue for a particular movie. Various stakeholders such as actors, producers, directors etc. can use these predictions to make more informed decisions. Historical data of each component such as actor, actress, and director, composer that influences the success or failure of a movie is given due a weightage. This proposed work aims to develop a model based upon the data mining techniques that may help in predicting the success of a movie in advance thereby reducing certain level of uncertainty. As we all know in today's world the movie is the biggest source of entertainment and also the movie making for the business purpose. The main aim of any business is profit and we all know either it is a movie making business or any other business, all we want is profit and success. For this we need technology through which we can predict the success of the movie. If we will be able to predict the future of the movie in correct manner then it will be easy for the businessman to make changes in their project by improving the contents of the movie so that they can get higher revenue from it.The best and efficient way to find the detailed information about almost movie is IMDB A huge amount of data, which contains much valuable information about general trends of movies and people's interest is available there. Data mining techniques enable us to uncover information which will both conform and disprove common assumptions about movie and by this we can be able to predict the success of the movie before its release For this we are here using R-Software for predicting the movie success rate to do this first we get data from various sources like kaggle.com , and after that we generate the data set. After getting data we are generating the training and test data set. In that data set we generate a training set to build the model, while a test set is there to validate the model built.On the basis of generated model,the prediction have been done and results have been generated. By using that result that we get from the model we can easily predict that the movie is going to be super-hit, blockbuster or flop.

Literature Survey:-

•In paper [1], predictive model for the box office performance of the movie were represented by factors derived from social media and IMDb. Accordingly ,we have identified the following patterns in there:- (1) the popularity of leading actress is an important factor to the success or failure of a movie, and the combination of a past successful genre of a movie and a sequel movie is another pattern for the success of a movie, (3) a new movie which is not in the popular genre and also an actor with lower popularity could be a pattern for the Flop of the movie. It was seen that sentiment analysis and view and comment were not identified as relevant in our analyzation. Author understands that it is related to how weights are assigned to each of the attribute. In addition, our prediction is for the movies yet to be released.Future work to improve models will include further refinement of Neutral class and characterization of movie box office performance in terms of net profits and profit ratios.
•In paper[2],the author have arrived up with a model to find the success rating of upcoming movies based on certain results which can be derived from the already released movies. As per their model, we concluded that one factor was the movie genre which determined the success rating of the movie. It was also seen that the movie success depended on the cast of the movies. There was strong correlation between actors and genres which indicated that some certain actors always tend to work in certain genres only. The actors and genres then define success rating of movie. A limitation of the work is that it focused on the Bollywood movies only. In future, [2] will expand their model to include Hollywood movies also.
•In paper[3], author predicted movie based on online reviews because it is a trendy approach for the public to share their views and also it is more useful in business intelligence. [3] author used a lot of variety of algorithms for predicting the sales performance of the movies using sentiment information mined from reviews and tweets analyzed. Our main work is use of S-PLSA model. Using S-PLSA author summarized sentiment information from online review and the tweets, they have used ARSA model for predicting sales performance of the movies using sentiment analysis and past box office performance. They further classified the reviews into positive or negative or neutral after that and set a simple metric PN ratio and set the threshold value to predict the success or failure of movies, i.e. Hit, Success or Flop.
•In paper[4], the author shows success of a movie depends on different aspects associated.Yet, predicting the performance of a movie is some crucial part of the decision-making among the stakeholders of the movies.Over the time several techniques are proposed to predict the performance of a movie The existing works mainly concentrated on predicting performance of a movie by applying some criteria sentiment analysis on comments collected from several sources like YouTube, news, blogs.Without the sentiment analysis from the comments, other data like number of viewers, number of likes and dislikes are totally omitted in the existing works However, these data can play a vital role in predicting the performance of a movie in the box office.So in this work, author prepared a dataset which contained the data that was previously not taken serious.Later on, we performed an experiment by applying data mining techniques on the prepared dataset to identify some most suitable technique for predicting gross result of a movie The experiment showed that Linear regression most suitable method for predicting the gross income of a movie. As for future work, it was recommend performing an experiment by adding the data from different sources.

Dataset description:-

The dataset requirement for our project is fulfilled through kaggle repository from here we downloaded the dataset and used as the input. This dataset consists of 651 rows with 32 columns.The dataset we get is preprocessed so we need not to pre-process it. Our first task for this assignment is to select which variables we have to include in our work. It would be easier to start with eliminating variables that will not be of use for our model. Uniform Resource Locators or commonly known as URLs provide an easy way to find more information for each movie but will not provide information whether a movie is popular or not. Runtime or length of the movie in minutes is not a key ingredient for popularity of the movie. Runtime would be probably a good predictor of movie genre. Animation and Documentaries are generally shorter than the feature films. The title of a movie is what a audience remembers when a movie is popular but it is not what makes movie popular. However, this is not the case when it comes to actors, actresses or directors. Movie audience turn into fans when an certain actor,or the director captures their imagination and become the key determiner whether subsequent movie from the same actor or director is must see. Let us now focus our attention towards choosing our response variables. Since, we are working on development of model and

prediction, the requirement of our data is numerical. Thus, the attribute selected here must be an attribute containing the numerical values.

Since we are doing linear regression in place of logistic regression, there is a need for us to analyze the central tendency and quintiles of all these attributes. The table below show the acquired outputs:-

SVM1.R | Movie_prediction_code.R* | movies

Filter

	title	title_type	genre	runtime	mpaa_rating	studio	thtr_rel_year	thtr_rel_month	thtr_rel_day	dvd_rel_y
1	Filly Brown	Feature Film	Drama	80	R	Indomina Media Inc.	2013	4	19	
2	The Dish	Feature Film	Drama	101	PG-13	Warner Bros. Pictures	2001	3	14	
3	Waiting for Guffman	Feature Film	Comedy	84	R	Sony Pictures Classics	1996	8	21	
4	The Age of Innocence	Feature Film	Drama	139	PG	Columbia Pictures	1993	10	1	
5	Malevolence	Feature Film	Horror	90	R	Anchor Bay Entertainment	2004	9	10	
6	Old Partner	Documentary	Documentary	78	Unrated	Shcalo Media Group	2009	1	15	
7	Lady Jane	Feature Film	Drama	142	PG-13	Paramount Home Video	1986	1	1	
8	Mad Dog Time	Feature Film	Drama	93	R	MGM/United Artists	1996	11	8	
9	Beauty Is Embarrassing	Documentary	Documentary	88	Unrated	Independent Pictures	2012	9	7	
10	The Snowtown Murders	Feature Film	Drama	119	Unrated	IFC Films	2012	3	2	
11	Superman II	Feature Film	Action & Adventure	127	PG	Warner Bros. Pictures	1981	6	19	
12	Leap of Faith	Feature Film	Drama	108	PG-13	Paramount Home Video	1992	12	18	
13	The Royal Tenenbaums	Feature Film	Comedy	110	R	Buena Vista Distribution Compa	2002	1	4	
14	School for Scoundrels	Feature Film	Comedy	100	PG-13	MGM	2006	9	23	
15	Rhinestone	Feature Film	Comedy	111	PG	20th Century Fox	1984	6	20	
16	Burn After Reading	Feature Film	Drama	96	R	Focus Features	2008	8	27	
17	The Doors	Feature Film	Drama	140	R	Sony Pictures Home Entertainment	1991	3	1	
18	The Wood	Feature Film	Drama	106	R	Paramount Pictures	1999	7	16	
19	Jason X	Feature Film	Horror	91	R	New Line Cinema	2002	4	26	
20	Dragon Wars	Feature Film	Drama	90	PG-13	Sony Pictures Home Entertainment	2007	9	13	
21	Fallen	Feature Film	Drama	124	R	Warner Home Video	1997	6	1	
22	The Gleaners and I	Documentary	Documentary	82	Unrated	Zeitgeist Films	2001	4	6	

Figure 1:- Shows the summary information of dataset

FLOW DIAGRAM OF IMPLEMENTATION

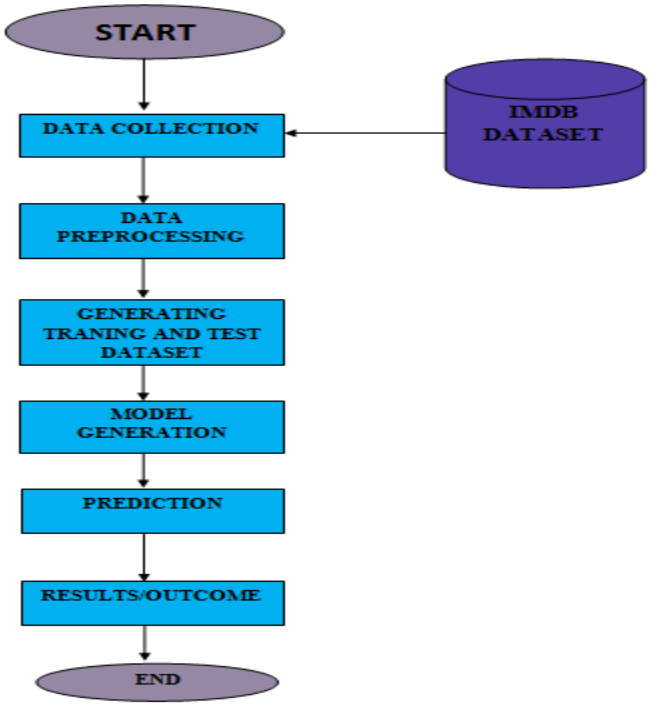


Figure 2

Data Collection:

The IMDB(Internet Movie Database) is an database of information related to movies, television shows, online streaming videos. It include cast, production, crew, writers, directors, producers, plot summery, audience and critics reviews and ratings The crude IMDb dataset is organized so that the greater part of its information is sorted out. For example, the majority of the approximately most of movies picture appraisals are organized in the content document evaluation. In such a way, some kind of cleaning, mixing and pre-

processing is probably going to be required so as to utilize the information with the end goal of information mining.

Data Pre-processing:

Before applying data mining technique, pre-processing methods like converting raw data into understandable format, data partitioning and other techniques for attribute selection must be applied. Some times in real world scenario, data is often incomplete, inconsistent or lacking in certain trends and may like to contain many errors. As the data is taken from IMDb it is first required to be pre-processed. Pre-processing is the crucial phase of the data mining as it mainly focuses on the working of the algorithm and also filtering out the useful data from the raw data.To overcome the missing values scenarios central tendency methods is used i.e. mean and median and the duplicate items are removed.

Generating Training and Test Dataset:

Training dataset is a set of attributes used to fit the parameters of the model. The current model has run with the training dataset and produces a result, which is then compared with the *target*, for each input vector in the training dataset. Based on the result of the comparison and specific learning algorithm being used, the parameters of the model are adjusted. The model fitting can include variable selection and parameter estimation. Finally, the test dataset is a dataset that is used to provide an evaluation of a *final* model fit on the training dataset.

Data Analysis:

Data Analysis is an process of cleaning, transforming, and modelling the data to discover some useful information for business decision making. All attributes are analyzed on the basis of different factors that help the user to gather most accurate outcome for further stages. The Analysis can be done on movie success prediction on the basis of following factors: IMDb rating, IMDb no.of up votes and down votes, critics ratings, critics score, audience rating and audience votes.
An actor, actress, or directors who has won the Oscars, iifa, or any other awards can be a great factor of predicting the movie success percentage As awards impact the image of the cast of the movie in the mind of the audience, critics.

Model Generation:

Modelling is a simplified mathematically formalized way to make predictions from the dataset available. Representing a quantity by an average and a standard deviation is a very simple way to form the statistical modelling. In this project we are using correlation along with correlation matrix plot to generate our model. The test result validate the quality and validate the model. After assessing the model it came out that the models meet the business initiatives or not.

Figure 3:- Below snapshot shows the developed training and testing dataset

```
> training
# A tibble: 650 x 32
   title title_type genre runtime mpaa_rating studio thtr_rel_year thtr_rel_month
   <chr> <fct>      <fct>   <dbl> <fct>       <fct>          <dbl>          <dbl>
 1 Fill~ Feature F~ Drama      80 R           Indom~          2013              4
 2 The ~ Feature F~ Drama     101 PG-13       Warne~          2001              3
 3 Wait~ Feature F~ Come~      84 R           Sony ~          1996              8
 4 The ~ Feature F~ Drama     139 PG          Colum~          1993             10
 5 Male~ Feature F~ Horr~      90 R           Ancho~          2004              9
 6 Old ~ Documenta~ Docu~      78 Unrated     Shcal~          2009              1
 7 Lady~ Feature F~ Drama     142 PG-13       Param~          1986              1
 8 Mad ~ Feature F~ Drama      93 R           MGM/U~          1996             11
 9 Beau~ Documenta~ Docu~      88 Unrated     Indep~          2012              9
10 The ~ Feature F~ Drama     119 Unrated     IFC F~          2012              3
# ... with 640 more rows, and 24 more variables: thtr_rel_day <dbl>, dvd_rel_year <dbl>,
#   dvd_rel_month <dbl>, dvd_rel_day <dbl>, imdb_rating <dbl>, imdb_num_votes <int>,
#   critics_rating <fct>, critics_score <dbl>, audience_rating <fct>, audience_score <dbl>,
#   best_pic_nom <fct>, best_pic_win <fct>, best_actor_win <fct>, best_actress_win <fct>,
#   best_dir_win <fct>, top200_box <fct>, director <chr>, actor1 <chr>, actor2 <chr>,
#   actor3 <chr>, actor4 <chr>, actor5 <chr>, imdb_url <chr>, rt_url <chr>
> testing
# A tibble: 1 x 32
  title title_type genre runtime mpaa_rating studio thtr_rel_year thtr_rel_month
  <chr> <fct>      <fct>   <dbl> <fct>       <fct>          <dbl>          <dbl>
1 Pris~ Feature F~ Drama     102 R           New W~          1988              3
# ... with 24 more variables: thtr_rel_day <dbl>, dvd_rel_year <dbl>, dvd_rel_month <dbl>,
#   dvd_rel_day <dbl>, imdb_rating <dbl>, imdb_num_votes <int>, critics_rating <fct>,
#   critics_score <dbl>, audience_rating <fct>, audience_score <dbl>, best_pic_nom <fct>,
#   best_pic_win <fct>, best_actor_win <fct>, best_actress_win <fct>, best_dir_win <fct>,
#   top200_box <fct>, director <chr>, actor1 <chr>, actor2 <chr>, actor3 <chr>,
#   actor4 <chr>, actor5 <chr>, imdb_url <chr>, rt_url <chr>
```

After generating, we got training dataset with 650 rows and 32 attributes and, testing dataset with 1 row and 32 attributes

Data Analysis:

In data analysis, all selected attributes are analyzed on the basis of different factor that help us to gather most accurate outcome for further stages. Selected features for analysis are as follows:, imdb_num_votes, critics_score, critics_rating, imdb_rating audience_score and audience_rating. On the basis of these attributes, we are generating various visualized graphs for analyzing the best possible attribute among these for further predictions.

An actor,or director who won an oscar award is a acceptable motivation to analyze the movie success. Movie which has won an oscar has the same weight as an actor or a director in making some movie popular. So, in our analysis we are generating some scatter plot to show differences between number of oscar won by particular actor and director.

Figure 4:- Distribution of critics_score and imdb_rating on scatter plot

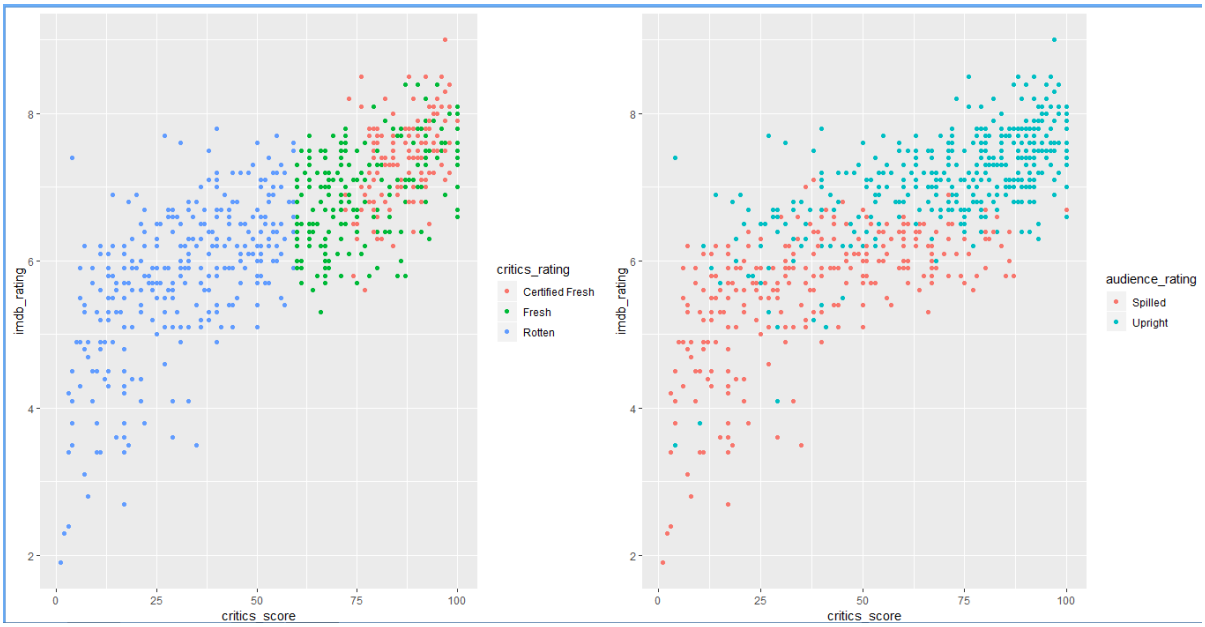


Figure 5:-Showing imdb_rating, imdb_num_votes, critics_score and audience_score data on histogram

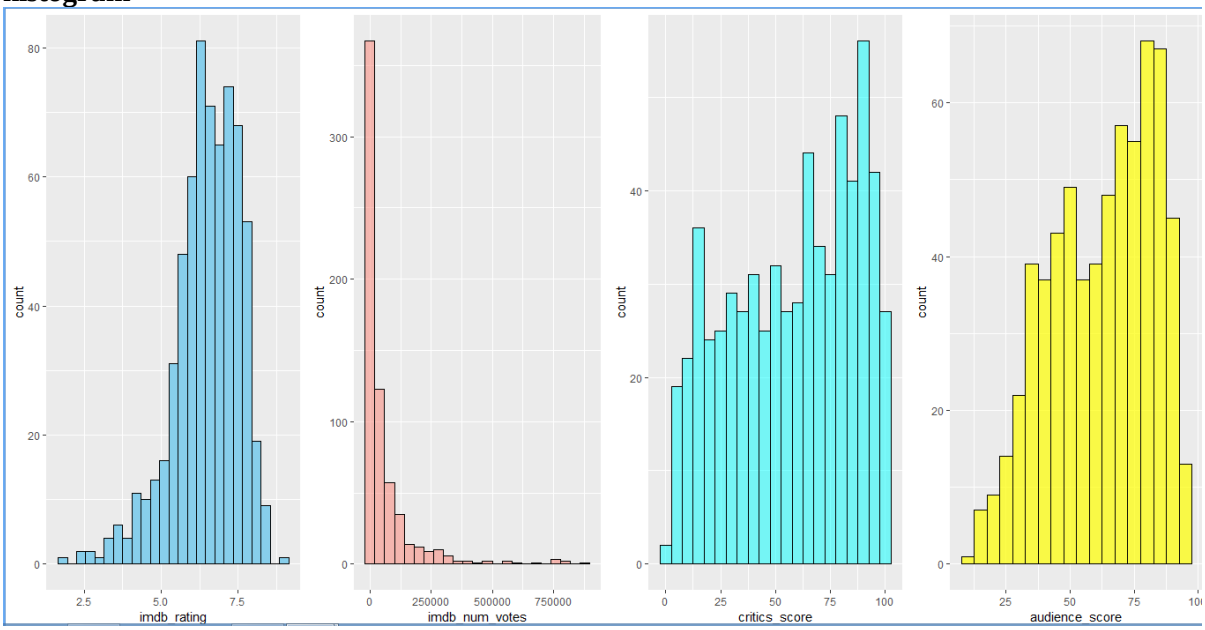


Figure 6:- Tabular representation of oscar win by actor, actress and director

	no	yes
At.least.one.Oscar	479	171
best.actor	557	93
best.actress	578	72
best.director	607	43

Figure 6:- Scatter plot showing the above data for oscar win

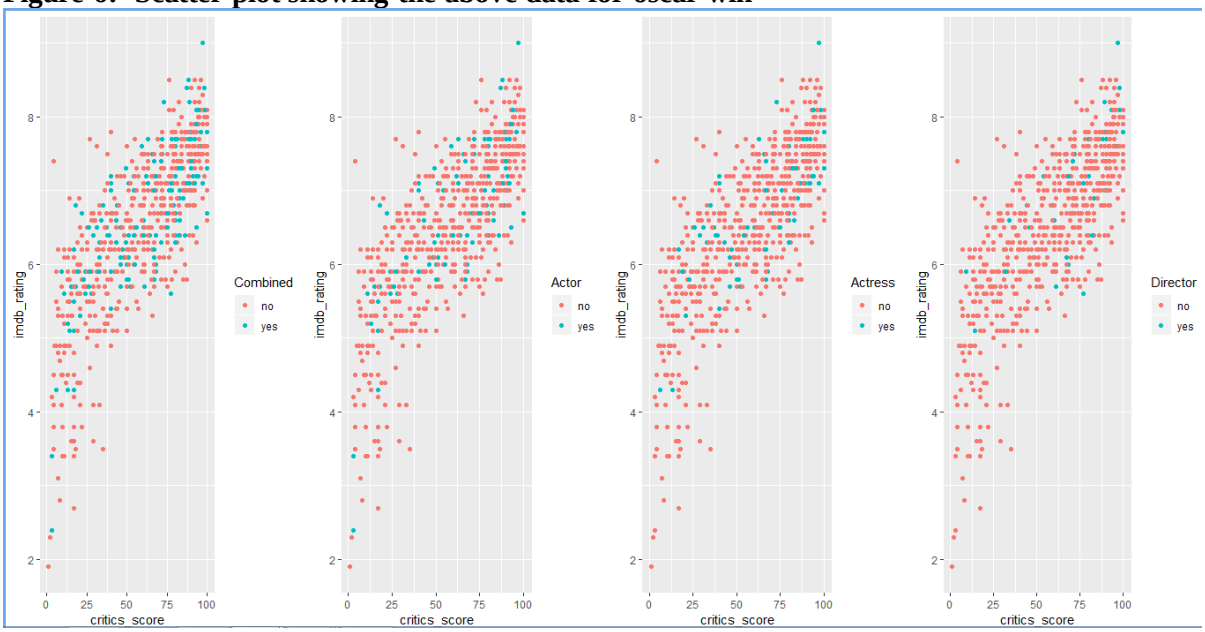


Figure 7:- Tabular representation of movie nomination and winning for oscar

	no	yes
combined	627	23
nominations	628	22
wins	643	7

Figure 8:- Scatter plot showing the above data for oscar nomination and winning

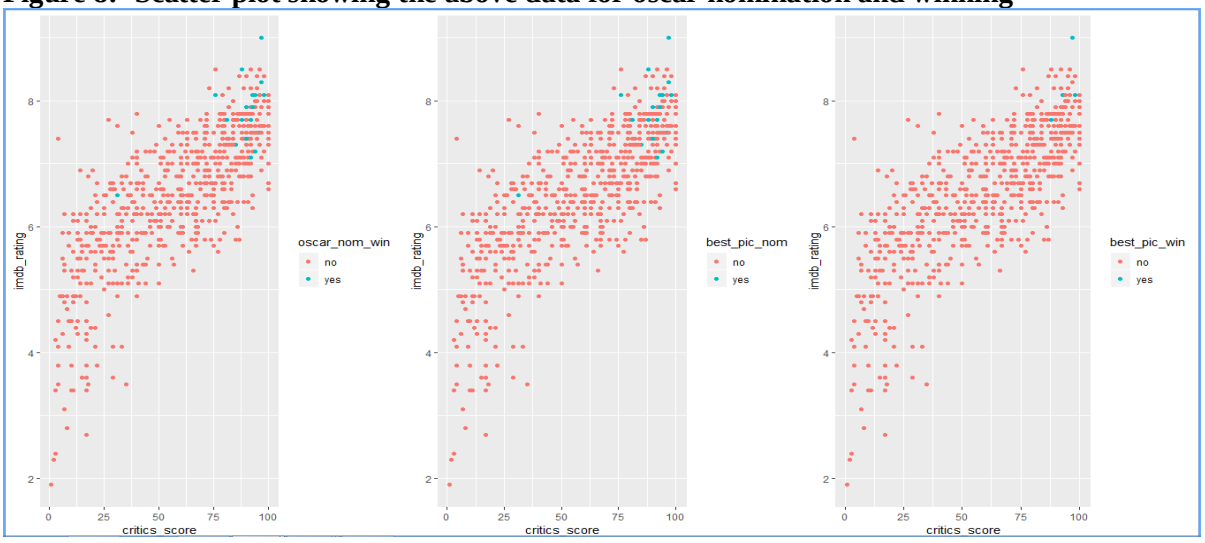


Figure 9:- Barplot and Scatter plot for year, month and day basis movie release data

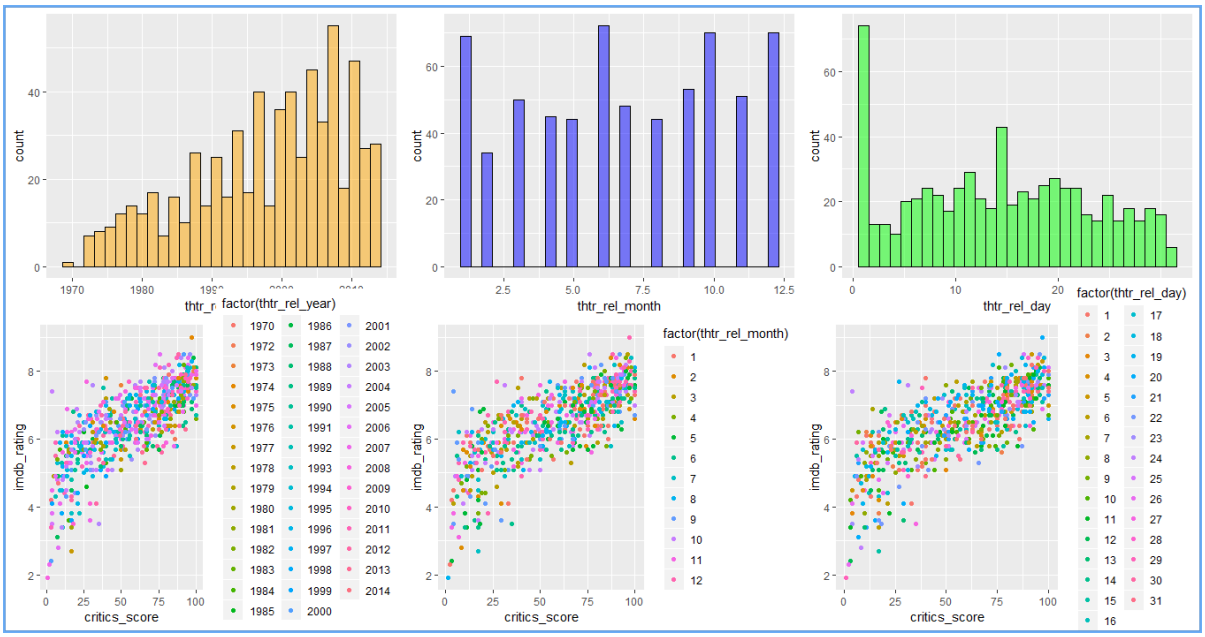


Figure 10:- Barplot and Scatter plot for year, month and day basis movie dvd release data

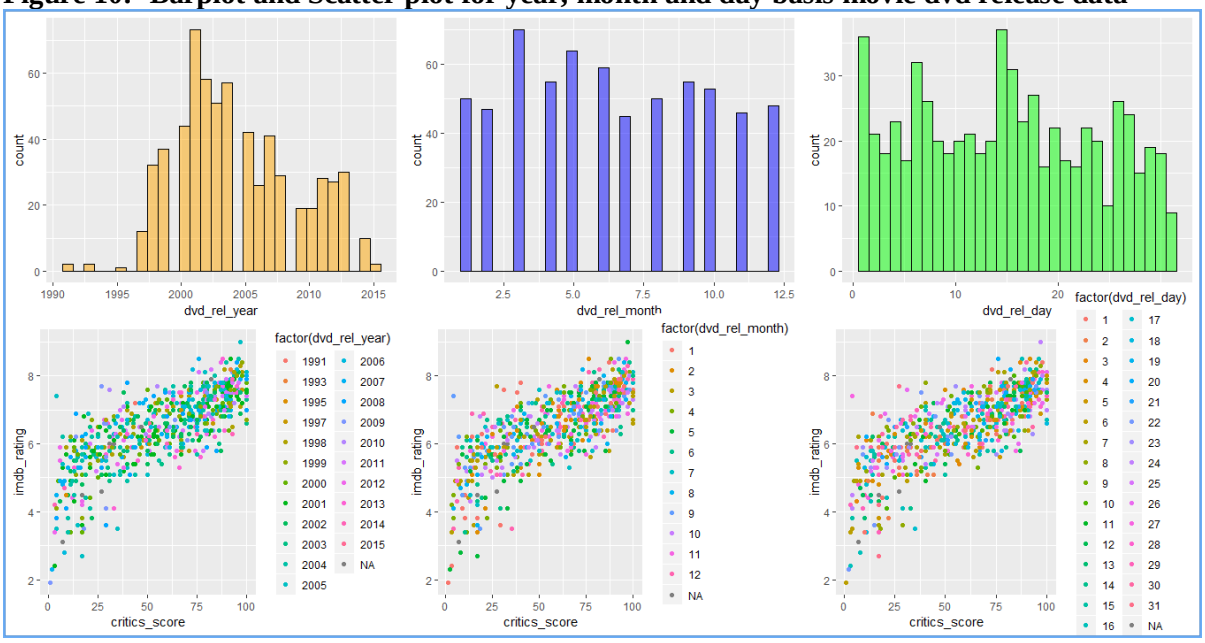


Figure 11:- Scatter plot along with line chart to show critics_score and audience_score

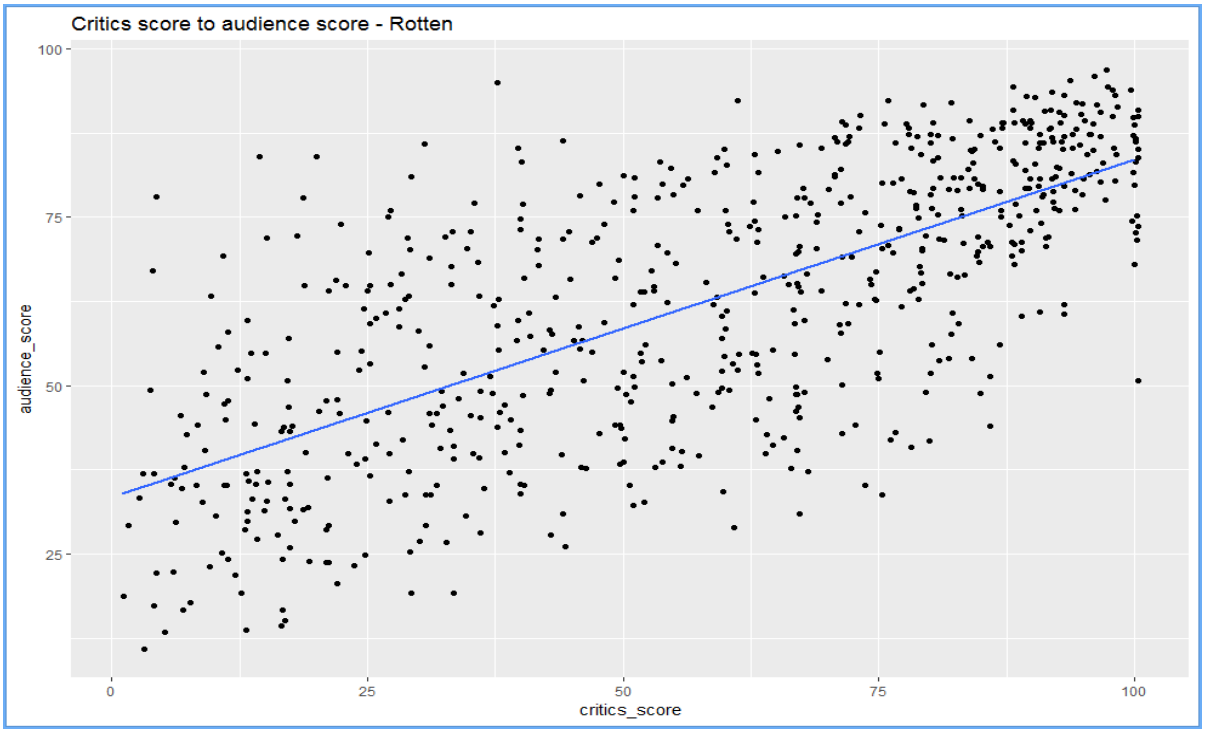
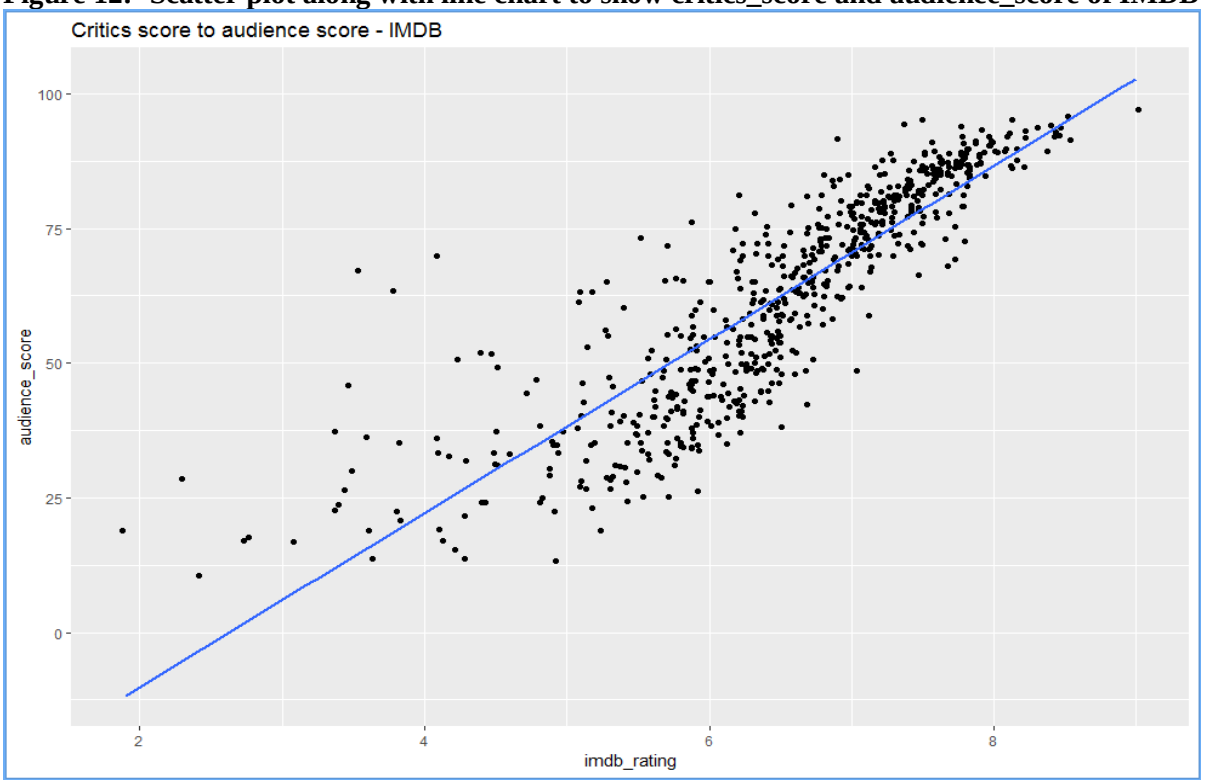


Figure 12:- Scatter plot along with line chart to show critics_score and audience_score of IMDB



Model Generation:

In simple terms, modelling is a simple and mathematically-formalized way to approximate the reality.The statistical model is the mathematical equation that is used. Representing a quantity by an average and a standard deviation is a very simple form of statistical modelling. In this project we are using correlation along with correlation matrix plot to generate our model. We also using linear regression for model generation.

Figure 13:- Correlation matrix plot of selected attributes

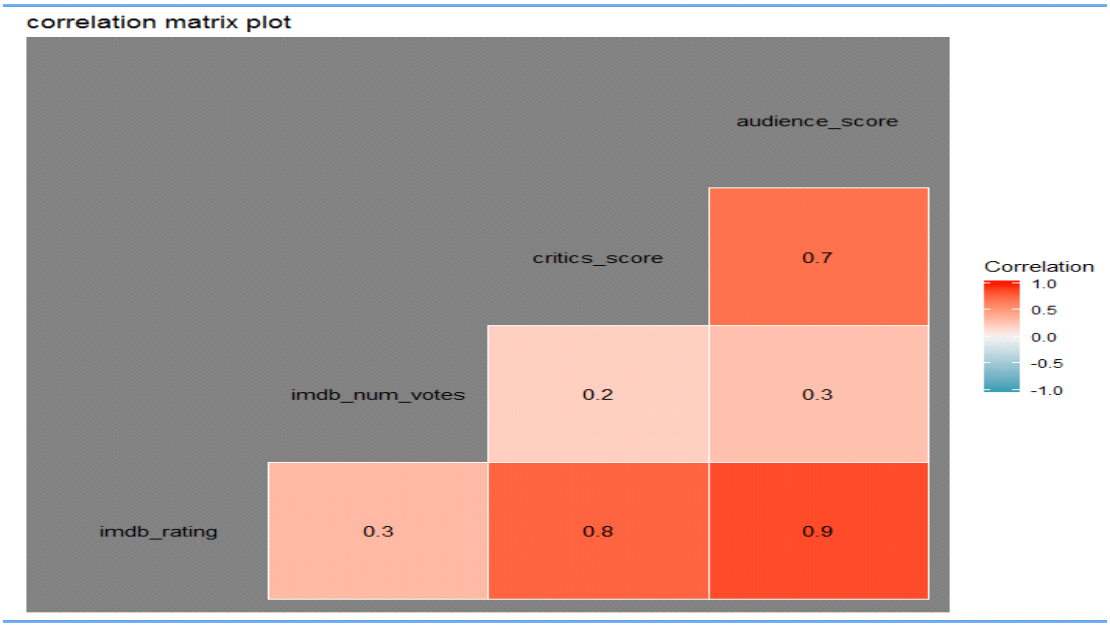


Figure 14:- Correlation between imdb_rating and audience_score, critics_score

```
> fit2 <- lm(imdb_rating ~ audience_score + critics_score, data = flm2)
> summary(fit2)

Call:
lm(formula = imdb_rating ~ audience_score + critics_score, data = flm2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.52039 -0.19919  0.03143  0.30586  1.22849

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6422058  0.0625817   58.20   <2e-16 ***
audience_score 0.0347913  0.0013412   25.94   <2e-16 ***
critics_score  0.0117924  0.0009538   12.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4902 on 647 degrees of freedom
Multiple R-squared:  0.7966,    Adjusted R-squared:  0.796
F-statistic:  1267 on 2 and 647 DF,  p-value: < 2.2e-16
```

Figure 15:- Correlation between imdb_rating and audience_score, critics_score, oscar

```
> fit4 <- lm(imdb_rating ~ audience_score + critics_score + oscar, data = flm2)
> summary(fit4)

Call:
lm(formula = imdb_rating ~ audience_score + critics_score + oscar,
    data = flm2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.49737 -0.22072  0.01678  0.29978  1.26035

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6122334  0.0629596  57.374  < 2e-16 ***
audience_score 0.0349077  0.0013333  26.182  < 2e-16 ***
critics_score  0.0115817  0.0009503  12.188  < 2e-16 ***
oscaryes       0.1325564  0.0435311   3.045  0.00242 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4871 on 646 degrees of freedom
Multiple R-squared:  0.7995,    Adjusted R-squared:  0.7986
F-statistic: 858.6 on 3 and 646 DF,  p-value: < 2.2e-16
```

Prediction:

Prediction is to identify data points purely on the description of another related data value. Prediction is done on the basis of the available attributes and variables of the movie prediction. By using the prediction, we can drive the relationship between a thing we know and the things we want to predict. Prediction also evaluates that the current data is matching with the desired outcomes or not.The prediction in data mining is known as Numeric Prediction. Generally, Regression analysis is used for prediction. Here, we give movie name as one of input along with critics score and then it will predict the audience score. If the audience score is nearer to old data then and it can fit in the range of lower limit and upper limit of predicted value, then we can say that our movie is successful or hit or flop.

Figure 16:- Prediction of Movie audience score

```
> #Prediction-1
> film <- data.frame(title ="Disaster Movie", critics_score = 1)
> predict(model, film, interval = "prediction", level = 0.95)
       fit      lwr      upr
1 33.93695 5.616396 62.2575
>
> #Prediction-2
> film2 <- data.frame(title ="Hellraiser - Bloodline", critics_score = 25)
> predict(model, film2, interval = "prediction", level = 0.95)
       fit      lwr      upr
1 45.97145 17.70844 74.23446
```

Here, we can see that we have got prediction1 of 33.94 audience score with 95% confidence level that our score will be between 5.62 and 62.26. Well, actual audience score of this movie is 19 We can see that predicted audience score is greater than actual audience score So, we can say that movie "Disaster Movie" is more successful than its actual performance.

Similarly, we can see that we have got prediction 2 of 45.97 audience score with 95% confidence level that our score will be between 17.71 and 74.23. Well, actual audience score of this movie is 47 We can see that predicted audience score is less than actual audience score So, we can say that movie "Hellraiser-Bloodline" is not a successful movie.

Conclusion:

In this project we have tried to determine if there is any association between different attributes presented in our dataset. Here, our main aim is to find association between numeric type attributes that is used as scoring system and how we can use this association for the prediction. As a result, we found that critics score is in strong positive relationship between critics score and audience score. And we can also conclude that critics score are best predictor of audience score. Thus, we can predict our movies success on the basis of critics score. In future, we can add many attributes as our predictors and can build enhanced model for that attributes to perform the prediction. Here, we assume that if we have movie gross

score and movie net profit along with manufacturing cost of the movie, then we can build a better and more strong model for movie success prediction.

References:-

[1] Krushikanth R Apala ; Merin Jose ; Supreme Motnam ; C -C Chan ; Kathy J Liszka ; Federico de Gregorio" Prediction of Movies Box Office Performance Using Social Media", 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining(ASONAM 2013), PP 1209- 1214

[2] Javaria Ahmad ; Prakash Duraisamy ; Amr Yousef ; Bill Buckles" Movie Success Prediction Using Data Mining", 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), PP 1- 4

[3] Susmita S Magdum ; J V Megha" Mining Online Reviews and Tweets for Predicting Sales Performance and Success of Movies", International Conference on Intelligent Computing and Control Systems (ICICCS)PP 334-339

[4] Md Shamsur Rahim ; A Z M Ehtesham Chowdhury ; Md Asiful Islam ; Mir Riyanul Islam" Mining Trailers Data from YouTube for Predicting Gross Income of Movies",2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC),PP 551-554

[5]A Data mining Technique for Analyzing and Predicting the success of Movie
K Meenakshi, G Maragatham, Neha Agarwal and Ishitha Ghosh
Department of Information Technology, SRM Institute of Science and Technology

[6] Movie success prediction using data mining

Publisher: IEEE Javaria Ahmad ; Prakash Duraisamy ; Amr Yousef ; Bill Buckles

[7] Movie Success Prediction Using Data Mining
Antara Upadhyay, Nivedita Kamath, Shalin Shanghavi, Tanisha Mandvikar, Pranali Wagh
B E Student, Asst Professor
Department of Information Technology, Shah & Anchor Kutchhi Engineering College, Mumbai, India

[8] Prediction of Movie Success using Sentiment Analysis of Tweets
Vasu Jain,Department of Computer Science,University of Southern California,Los Angeles, CA, 90007

[9] Movie Success Prediction
PROJECT REPORT ,Rakesh Parappa | U01382090 | CS660