GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Movie Rating Prediction Analysis

A Report for the Final Evaluation of Project 2

Submitted By

Rohit Joshi

(1613101589,16SCSE101579)

in partial fulfillment for the award of the degree

of

Bachelor of Technology

IN

Computer Science And technology

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Under the Supervision of

MR. DEEPENDRA RASTOGI , AST. PROFESSOR

APRIL/MAY-2020

TABLE OF CONTENTS

1.Abstract

Predicting the success of a movie before its release has far been a huge point of concern for directors and producers alike, especially after they factor in real world data and the occurrence of unforeseen circumstances.

Since the seentities invest massively in movies,they need some kind of reassurance that the movie will be successful and reapinmassive profit in terms of financial returns and credibility. In this paper,wetalk about a prediction engine we have implemented that uses LOGISTIC REGRESSION, LINEAR REGRESSION, to categorize a movie as successful or not as well as using our own algorithm to calculate a target variable which is the IMDB score of a movie based on various parameters.

Our engine has proved to be successful from the technical unit testing performed on it with theuse of multiple iterations of input values. There are various related paper which I studied to do this research. There search paper which I studied are given in the reference. The number of reference paper taken to write this paper is more than twenty. Here we are applying relationship between IMDB and duration time of movie with the help of logistic regression and linear regression. The tool which we used for the establishing of relationship between the seare -JUPYTER. And the libraries which are used are numpy, pandas, matplotlib, seabborn, sklearn, OS.The customer side of in this research paper is that customer easily able to know that which movie is hit, flop or average according to relationship with duration and IMDB rating. So if movie is hit or average customer go to watch the movie and if movie is flop then customer may go or not.

2. Introduction

2.1 Overall Description

A movie is also called a film or motion picture is a combination of still images, when displayed on a screen behaves as an illusion of moving images. It is because of the phi phenomenon. The process of movie making is both an industry and an art. Movies are a great source of entertainment and people are crazy about movies. Movie Industry produces hundreds of movies every year of different genres such as animation, war, comedy, thriller horror etc. Most of the time, people are not sure about which particular movie to look for so that their spare time is utilized in entertainment.

There are many online platforms that keep track of movies like Internet Movie Database (IMDB) which provide information about movies such as actors, directors, budget, as well as user ratings and comments which provide a fair information about the movie. Internet movie database (IMDB) is the number one consumer site of movies. It contains information about programs, films and television including financial information, biographies, user rating, cast, reviews, crew, actors, directors, summaries etc. It has database of approx. 60 million registered users and 6.6 million personalities with 3.4 million movie and episodes titles.

2.2 Purpose

“Hollywood is the land of hunch and the wild guess” .Thousands of movies are released every year. According to a study, movie industry in the United States generate revenue up to 10 billion dollars. Each movie cost 100 million approx. but still there is a great deal of uncertainty that the movie will do

business or not. Movie industry is a big business, which can give profits or loss up to several millions dollar. It will be a hit or flop this give rise to the movie prediction problem. A lot of research has been done on prediction of movies. Most of them include user ratings on different movies, whereas some of them use social media (e.g. YouTube, Twitter etc) for prediction. However, less work has done on using movies attributes such as crew, dates etc. to predict movies. The amount of data available about the movies over the internet makes its serious candidate for data mining, knowledge discovery and also machine learning. Most of the work done on movies relating to its ratings and reviews over the internet

2.3 Project Scope

. Prediction of a movie is of great importance to industry; movie makers are still never sure about whether there movie will do business or not; when they should release the movie and how to advertise it.

1) Original IMDB score 2) Algorithmically calculated IMDB score 3) Fuzzy categorization of the movie's success 4) Percentile of movies it falls under within that category
5) Accuracy range of our algorithm
6) Fuzzy Success

2.4 Literature Survey

Neural Networks have been extensively used in forecasting and prediction studies, and so it has been also employed for predicting the success and failure the movies also. However, Hadavandi *et al [*4] have used integration of genetic fuzzy systems and artificial neural networks for stock price

forecasting .Stock market prediction is regarded as a challenging task in financial time-series forecasting. The central idea to successful stock market prediction is achieving

best results using minimum required input data and the least complex stock market model. They used stepwise regression analysis (SRA) to determine factors which have most influence on stock prices. In the next stage divided raw data into k clusters by means of self-organizing map (SOM) neural networks. Finally, all clusters were fed into independent GFS models with the ability of rule base extraction and data base tuning. They evaluated capability of the proposed approach by applying it on stock price data gathered from IT and Airlines sectors, and compare the outcomes with previous stock price forecasting methods using mean absolute percentage error (MAPE). Results show that the proposed approach outperforms all previous methods, so it can be considered as a suitable tool for stock price forecasting problems. An organization has to make the right decisions in time depending on demand information to enhance the commercial competitive advantage in a constantly fluctuating business environment. Therefore, estimating the demand quant i try for the next period most likely appears to be crucial. Efendigil et al [5] have suggested a decision support system for demand forecasting wit h artificial neural network and Neuro fuzzy logic. Neural networks have been used in predicting the traffic flow. Chan et al [6] have deployed neural networks based upon exponential smoothing method for predicting the traffic flow. Reuter and Muller [7] have developed artificial neural network for forecasting of fuzzy time series. Forecasting box office revenue of a movie before its theatrical release is a difficult and challenging problem. In a study conducted by Zhang et al, [8] a multi-layer BP neural network (MLBP) with multi -input and multi output was employed to build the prediction model. All the movies were divided into six categories ranged from “blob” to “bomb” according to their box office incomes, and the purpose is to predict a film into the right class. The selections of the input variables were based on market survey and their weight values were determined by using statistical method. As to the design of the neural network structure, theoretical guidance and plentiful experiments were combined to optimize the hidden layers‟ parameters which include the number of hidden layers and their node numbers. Then a classifier with dynamic thresholds was used to standardize the output for the first time, and to improve the robustness of the model to a high level. Finally, a 6-fold cross-validation experiment methodology was used to measure the performance of the prediction model. The comparison results with the MLP method showed that the MLBP prediction

model achieves more satisfactory results, and it is more reliable and effective to solve the problem. A Bayesian belief network (BBN), which is known as a causal belief network, was constructed [9] to investigate the causal relationship among various movie attributes in the performance prediction of box-office success. Subsequently, sensitivity analysis was conducted to determine those attributes most critically related to box- office performance. Finally, the probability of a movie's box-office success was computed using the BBN model based on the domain knowledge from the value chain of theoretical motion pictures. The results confirmed the improved forecasting accuracy of the BBN model compared to artificial neural network and decision tree.

Asur and Huberman [10] in their paper have demonstrated how social media content can be used to predict real -world out comes. In particular, they have used the chatter from Twitter.com to forecast box-office revenues for movies. They have shown that a simple model built from the rate at which tweets are created about particular topics can outperform market-based predictors. Some studies have also been carried out and have been reported in ne was median and social media (YouTube), but no academic record for these studies is available. One such study was carried out by researchers of Indian Institute of Management Ahmadabad and other one was carried in China [11].The present study will use neural network based machine learning algorithm for predicting movie success.

3.Proposed Model

he entire study was conducted in a sequential manner as listed below:

Collection of data pertaining to parameters under study (Actor, Actress, Producer, Director, Writer, Music

Director , Time of Release and Marketing Budget)

i. Data processing for assigning weights and calculating thresholds

ii. Input & Target pattern formation

iii. Architecture design and Neural Network Learning

iv. Performance Analysis

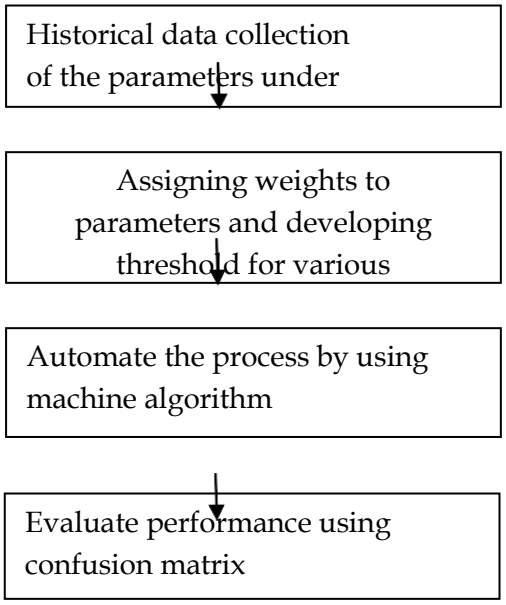


Fig1. Basic Flow of the research process

Step – 1 Since our research work will be driven mainly based on the observations from the historical data pertaining to the parameters under study it is anticipated that dataset will be non linear in nature. And will have data point's distribution in such a manner that it will be difficult to find the hyper plane. So, will be using SVM for feature extraction.

A_h is number of hit movies delivered out of the last 10 released movies of the concerned actor

$$\frac{\sum Ash}{10} = Asw$$

Where As_W is weight assigned to the actress.

A_h is number of hit movies delivered out of the last 10 released movies of the concerned actress

Similarly weights will be assigned to other factors (Director, Producer, Music Director, and Writer).

Only two types of weights will be assigned for the Time of Release. Movies released during holiday season (Diwali, Dushera, Id, New Year, Christmas etc.) will be assigned a weight of 0.9. While the movies released during other time will be assigned weight 0.7.

For assigning weights to the parameter marketing budget, a base value of Rs.10 crores will be taken. Weights will then be assigned as per the following formula:

$$\frac{\sum MB}{10} = MBw$$

Where MB_W is weight assigned to the Marketing budget.

$\sum$ MB is the sum total of all the expenses incurred on promotion of the movie (TV advertisements, events, launch on you tube etc.)

However for our research work, all the weights and biases are set to small real random values between 0 &1

Step2.

Table – 1 Attributes/Parameters under study

S.No	Attribute	Description	Mathematical Expressions
1	Actor	Leading Actor and status of his last 10 movies	$\frac{\sum A_h}{10} = A_w$
2	Actress	Leading Actress and status of his last 10 movies	$\frac{\sum A_{sh}}{10} = A_{Sw}$
3	Director	Director and status of his last 10 movies	$\frac{\sum A_d}{A_d\ 10} =$
4	Producer	Producer and status of his last 10 movies	$\frac{\sum A_p}{10} = A_p$
5	Music Director	Music Director and status of his last 10 movies	$\frac{\sum A_m}{A_m\ 10} =$
6	Writer	Writer and status of his last 10 movies	$\frac{\sum A_w}{A_w\ 10} =$

7	Marketing Budget	Base value of Rs.10.00 crores	∑ MB _____= MBw 10
8	Time of Release		Release during holiday season =0.9 Release during other time =0.7

Step 3:- Selection of classifier for the above dataset: Dataset in our case is non-stationary as well as non-linear, so we will be using neural network for understanding the pattern and predicting the success of movie. Movie business in India is a billion dollar industry, employing thousands of people as such success or failure of a movie can have profound effect on the stake holders. It is therefore of prime interest of the stake holders to know how the movie will fare. The present study aims to develop a model for the same. Many machine learning algorithms are available; however for in this study we will be using Artificial Neural networks utilizing supervised learning.

Step – 4 Evaluation of the Results: Model so developed will be tested for actual prediction of the movies and efficacy will be accordingly established.

4. Implementation

4.1 Coding Part

Algorithm 1

```
import pandas as pd
import os
data=pd.read_csv(r'C:\Users\sony\Desktop\dataset\movie1.csv')
data1=pd.DataFrame(data)
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
data1.columns
Index(['Unnamed: 0', 'Title', 'Year', 'Genres', 'Language', 'Country',
       'Content Rating', 'Duration', 'Aspect Ratio', 'Budget',
       'Gross Earnings', 'Director', 'Actor 1', 'Actor 2', 'Actor 3',
       'Facebook Likes - Director', 'Facebook Likes - Actor 1',
       'Facebook Likes - Actor 2', 'Facebook Likes - Actor 3',
       'Facebook Likes - cast Total', 'Facebook likes - Movie',
       'Facenumber in posters', 'User Votes', 'Reviews by Users',
       'Reviews by Crtiics', 'IMDB Score'],
      dtype='object')
from sklearn.cross_validation import train_test_split
train_test_split
<function sklearn.cross_validation.train_test_split(*arrays, **options)>
X_train, X_test, y_train, y_test = train_test_split(x1,y1, test_size=0.33, random_state=1)
from sklearn.linear_model import LogisticRegression
logmodel=LogisticRegression()
```

```
logmodel.fit(X_train,y_train)
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score
accuracy_score(y_test,prediction)
```

Algorithm 2

```
X1=new_data["Year"].values
Y1=new_data['IMDBScore'].values
mean_x=np.mean(X1)
mean_y=np.mean(Y1)
n=len(X1)
n
100
numer=0
denom=0
for i in range(n):
    numer+=(X1[i]-mean_x)*(Y1[i]-mean_y)
denom+=(X1[i]-mean_x)**2
b1=numer/denom
b0=mean_y-(b1*mean_x)
pprint(b1,b0)
0.17336291942344018 -330.42087157372345
ss_t=0
ss_r=0
for i in range(n):
y_pred=b0+b1*X1[i]
```

```
ss_t+=(Y1[i]-mean_y)**2 ss_r+=(Y1[i]-y_pred)**2

r2=1-(ss_r/ss_t)

print(r2)


-6.140174122189605

-6.230447652509172

from sklearn.linear_model import LinearRegression

from sklearn.metrics import mean_squared_error

X1=X1.reshape((n,1))

reg=LinearRegression()

reg=reg.fit(X1,Y1)

Y1_pred=reg.predict(X1

                    )

mse=mean_squared_error(Y1,Y1_pred)

rmse=np.sqrt(mse)

r2_score=reg.score(X1,Y1)

print(rmse)

0.7680779398355558

print(r2_score)

0.001881849330636065
```

4.2 TRAINING DATASET

4.2.1 YEARS ATTRIBUTE(X1)

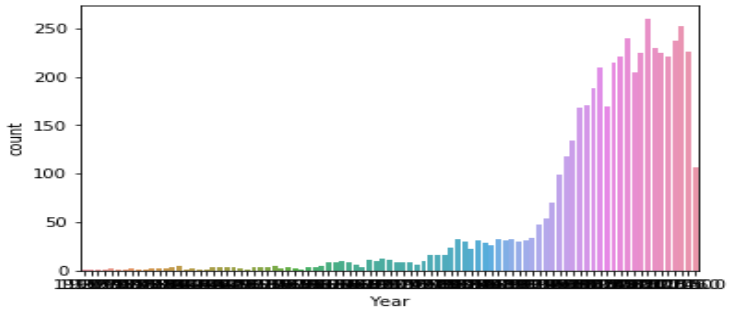
1916., 1920., 1925., 1927., 1929., 1929., 1930., 1932., 1933.,
1933., 1934., 1935., 1936., 1936., 1937., 1937., 1938., 1938.,
1939., 1939., 1939., 1940., 1940., 1940., 1940., 1940., 1941.,
1942., 1942., 1943., 1944., 1945., 1945., 1945., 1945., 1946.,
1946., 1946., 1947., 1947., 1947., 1948., 1948., 1948., 1949.,
1949., 1950., 1951., 1951., 1951., 1952., 1952., 1952., 1952.,
1953., 1953., 1953., 1953., 1954., 1954., 1954., 1954., 1954.,
1955., 1955., 1956., 1956., 1956., 1957., 1957., 1958., 1959.,
1959., 1959., 1960., 1960., 1960., 1961., 1961., 1961., 1961.,
1961., 1962., 1962., 1962., 1962., 1962., 1962., 1962., 1962.,
1963., 1963., 1963., 1963., 1963., 1963., 1963., 1963., 1964.,
1964.

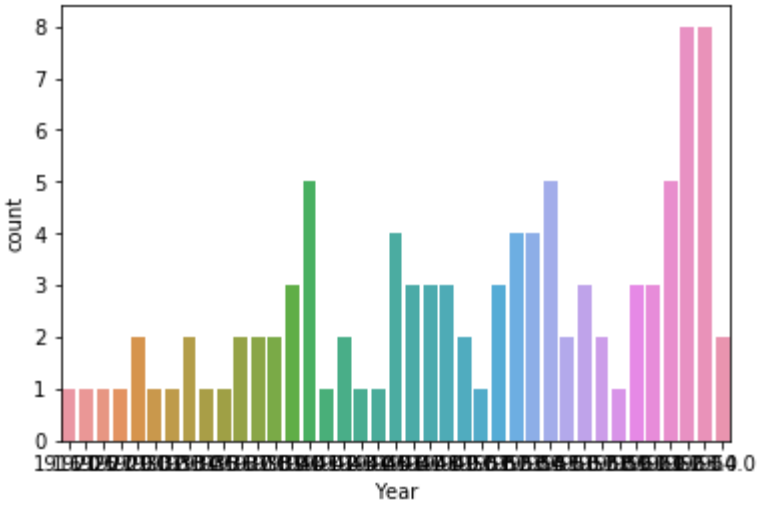
4.2.2 IMDB RATINGS(Y1)

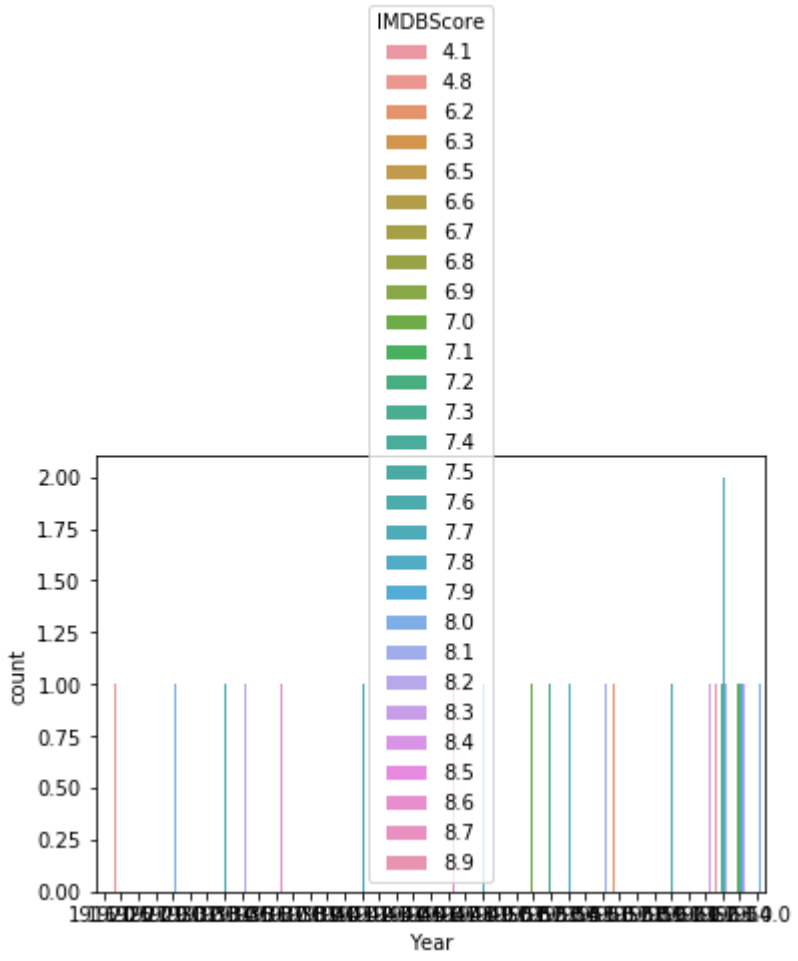
8. , 4.8, 8.3, 8.3, 8. , 6.3, 7.8, 6.6, 7.7, 6.5, 8.2, 7.8, 8.6,
7.1, 7.7, 7.8, 7. , 8. , 8.2, 8.2, 8.1, 7.1, 7.8, 7.5, 8.2, 6.5,
7.8, 7.4, 8.6, 7. , 6.5, 7.6, 7.1, 8. , 7.5, 6.9, 8.6, 8.1, 7.4,
7.7, 6.2, 7.1, 7.8, 7.1, 7.2, 7.4, 7. , 8. , 7.2, 7. , 7.2, 8.1,
8.3, 6.7, 7.8, 7.1, 6.7, 6.8, 7.2, 7.2, 8.2, 8.7, 6.6, 8.1, 7.2,
6.2, 7.4, 6.8, 8.9, 8.2, 8.1, 6.2, 8.3, 7.6, 7.9, 8.5, 8.3, 8.3,
7.3, 8. , 7.4, 7.6, 7.3, 8.4, 7.7, 7.7, 4.1, 7.8, 8.1, 8.4, 6.8,
7. , 6.9, 7.5, 7.9, 7.6, 8.3, 6.8, 8. , 7.7.

5. Output

5.1 DATA VISUALTION







TITLE	YEAR	GENRES	LANGUAGE	COUNTRY	DURATION	BUDGET	IMDB RATING
INTOLERANCE	1916	DRAMA	GERMAN	GERMANY	123	385907$	9
A FARWELL	1920	CRIME	ENGLISH	USA	134	379000$	4
PANDORA	1929	CRIME	FRENCH	FRANCE	145	456337$	9

BAMBIA	**1942**	DRAMA	GERMAN	GERMAN Y	78	564374$	5
ROSE MARY	1968	ROMANTIC	ENGLISH	GERMAN Y	86	454646$	7

FINAL OUTPUT:

-163.42375621409488

-4.892904228092516

-7.487654390527414

-9.339049636519608

-10.834481400969363

-9.68890915061622

-10.845942873162988

-10.364921782231885

-11.11852802270422

-10.388775581570249

-11.340209129197092

-11.121793784439332

-11.226845621177636

```
rmse=np.sqrt(mse)

r2_score=reg.score(X1,Y1)

print(rmse)
```

0.7680779398355558

```
print(r2_score)
```

0.001881849330636065

6.Conclusion

As movies are defined as experience goods with short product lifetime cycles, it is difficult to forecast the demand for motion pictures. Nevertheless, producers and distributors of new movies need to forecast box-office results in an attempt to reduce the uncertainty in the motion picture business. The study intends to develop a model to predict the financial success of a movie.

7. References

[1] https://en.wikipedia.org/wiki/Film , Accessed on August 1st, 2015

[2] https://en.wikipedia.org/wiki/Internet_Movie_Database , Accessed on August 1st, 2015

[3] Darin Im and Minh Thao Nguyen : "PREDICTING BOXOFFICE SUCCESS OF MOVIES IN THE U.S. MARKET ", CS 229, Fall 2011

[4] Jeffrey S. Simonoff and liana R.Sparrow : "Predicting Movie Grosses : Winners and Losers, Blockbusters and Sleepers" .

[5] Ramesh Sharda , Dursun Delen :"Predicting box-office success of motion pictures with neural networks", Expert Systems with Applications 30

[6] Nithin VR, Pranav M, Sarath Babu PB, Lijiya "A Predicting movie success based on IMDB data" International journal of data mining and techniques , Volume 03, june 2014, pages 365-368

[7] Sitaram Asur, Bernardo A. Huberman "Predicting the Future With Social Media", Hp Labs

[8] Andrei Oghina, Mathias Breuss, Manos Tsagkias, and Maarten de Rijke , "Predicting IMDB Movie Ratings Using Social Media", Advances in Information Retrieval ,

[9] Mestya´n M, Yasseri T, Kerte´sz J (2013): "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data".

[10] Mahesh Joshi Dipanjan Das Kevin Gimpel Noah A. Smith: "Movie Reviews and Revenues: An Experiment in Text Regression ", The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics,Pages 293-296

[11] Wenbin Zhang ,Steven Skiena :"Improving Movie Gross Prediction Through News Analysis", Department of computer science stony brook university, 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology – Workshops, Pages 301-304

[12] Khalid Ibnal Asad , Tanvir Ahmed , Md. Saiedur Rahman: "Movie Popularity

Classification based on Inherent Movie Attributes using C4.5,PART and Correlation Coefficient", IEEE/OSA/IAPR International Conference on Infonnatics, Electronics & Vision, Pages 747 – 752

[13] Saraee, M, White, S and Eccleston, J 2004, "A data mining approach to analysis and prediction of movie ratings ", The Fifth International Conference on Data Mining, Text Mining and their Business Applications,, 15-17 September 2004, Malaga, Spain

[14] Jeffrey Ericson & Jesse Grodman : "A Predictor for Movie Success" CS229, Stanford University

[15] Nikhil Apte, Mats Forssell, Anahita Sidhwa :"Predicting Movie Revenue", CS229, Stanford University ,December 16,2011

[16] Steven Yoo, Robert Kanter, David Cummings : "Predicting Movie Revenue from IMDb Data"

[17] Jiawei Han, Micheline Kamber, Jian Pei : "Data mining concepts & techniques", third edition, 2011

[18] http://www.cs.waikato.ac.nz/ml/weka , Accessed on August 1,2015

[19] Saba Bashir, Usman Qamar, Farhan Hassan Khan, M.Younus Javed : "An Efficient Rule-based Classification. of Diabetes Using ID3, C4.5 & CART Ensembles", 12th International Conference on Frontiers of Information Technology, Pages 226 – 231

[20] http://www.metacritic.com/ , Accessed on August 1,2015

[21] http://gim.unmc.edu/dxtests/ROC1.htm , Accessed on August 1,2015 Charles E. Metz , “Basic principles of ROC analysis”, volume 8, issue 4, pages 283-298, Deparmtnet of radiology, university of chicago