Predict Upcoming Movie Rating using IMDB Dataset

G. K. M. IFATULLAH
Student Id: 012123027

A Project

in

The Department

of

Computer Science and Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Science in Computer Science and Engineering

United International University

Dhaka, Bangladesh

Summer 2015

Approval Certificate

This project titled "**Predict Upcoming Movie Rating Using IMDB Dataset**" submitted by **G. K. M. IFATULLAH**, Student ID: **012123027**, has been accepted as Satisfactory in fulfillment of the requirement for the degree of Master of Science in Computer Science and Engineering on 22-Jan-18.

Board of Examiners

1.

_____ Supervisor

PROF. DR. MOHAMMAD NURUL HUDA,
Professor,
Dept. of CSE, UIU

2.

_____ Examiner

SWAKKHAR SHATABDA,
Associate Professor & Undergraduate Program Coordinator,
Dept. of CSE, UIU

Declaration

This is to certify that the work entitled "**Predict Upcoming Movie Rating Using IMDB Dataset**" is the outcome of the research carried out by me under the supervision of **Porf. Dr. Mohammad Nurul Huda**, Professor of Dept. of CSE, United International University, Bangladesh.

_____

G. K. M. IFATULLAH,
ID: 012123027,
MSCSE, UIU

In my capacity as supervisor of the candidate's project, I certify that the above statements are true to the best of my knowledge.

_____

PROF. DR. MOHAMMAD NURUL HUDA,
Professor,
Dept. of CSE, UIU

Abstract

Predicting a movie's box office and viewer's expectation success is a crucial issue since it does not always depend on its nature and condition only. External component such as competing movies of release date, time, occasion & festival of the year, weather and even political situation influence the success as these component impact the box office sales for the moving opening sales. Essentially, it is a venture whether a movie will be a success or failure. Our objective is to build the application to solve this problem by reducing the number of movies those lose money domestically and internationally. Our application desire to bring to light attractive delightful and entertaining, pairs of attributes that will fare well at the box office such as unique actor, actress, director and writer combinations. Our initial step for this project was to prevent studios, producers from making bad, rotten and unprofitable movies. Making use of our application, studios will be able to look into what virtue and characteristic make a successful movie before production begins. Another possible application is to advice investor make judgement on whether cast combination of the movie will be good or bad and profitable or not.

Acknowledgement

A lot of effort and study have been put to make this report. This would not have been possible without the genuine support and assistance provided by the people whom I approached during the various stages of writing this report. First of all, thanks to Allah SWT to have an opportunity by building the guide line for film industry including my involvement & activities within such a potential project. I'd like to express gratitude to my academic supervisor Prof. Dr. Mohammad Nurul Huda, Professor, Department of CSE, United International University; for his advice, counseling, direction and help. Without his guidance & methodology of works, this report would not have been possible.

I also must thank to all the faculty members of United International University for all of their direct and indirect encouragement and assistance in this work, my batch mates and friends who have provided me with their valuable suggestions throughout the entire semester. Their cooperation, suggestion, guidance and sincere encouragement played significant role throughout my working period.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Predicting the success of movie is a crucial problem in data mining and predictive analysis. A variety of way have been proposed to address real world prediction problems issuing in various domains. And there are vast movie data sources available online. Those are very suitable to decide which upcoming movie will be popular and successful. A movie success prediction has a lot of use for production and distributor houses to plan their resources. For example, a Bollywood studio, that expects that its upcoming movie to be highly successful will rent more theatre rooms in advance and that will increase more revenue if the prediction going to be true. If it rent less theatre rooms then all viewers will not able to watch the movie in it's opening weekend. Similarly, theatre owners can prevent loss by predicting lower rated movies to release in less theatre. If the process of data mining· and prediction was actually implemented by entities in the showbiz industry, it would be an improvement step toward ensuring the success of movies that the investor can invest in. The movie success is predicted on the basis of ratings from millions of user available in a consolidated dataset that we have used is IMDB and IMDB PRO (AMAZON DATASET). IMDB and IMDP PRO are the massive in the domain of rating and reviewing movies by using a Bayesian equation in which the ratings of multiple users is fed as the input for the movie and the result is a comprehensive and accurate rating that depicts the user's view of the movie. Based on IMDB and IMDB PRO dataset our application hopes to categorize new and interesting pairs of attributes that will fare well at the box office such as unique actor, actress director and writer combinations. Our early motivation for this project was to prevent studios from making unprofitable, bad and rotten movies. By making use of our application, studios will be able to look into what aspects make a profitable movie before it's production begins. But astonishingly accurate mechanism to predict future tends is very difficult. In this day's Web has turned into a major platform for information exchange, thus becoming a mirror of the real world, millions of volunteer post latest news on the websites and blogs those can influence on upcoming movies hive. Same way YouTube trailers hits count and also Facebook,

Twitter and other social network users and blogs comments can also influence a movie to success or failure. It is hard to tackle the problem of movie success prediction with the help of a different methodology, analytics, in order to get more insight into complex data structures. Analytics is a developing field of computer science, interlinking the areas of visualization and analytical methods (like, statistics, machine learning), in order to create deeper perceptions into datasets. Often typical machine learning models can not represent multipart data well enough to make good predictions, also it can be a very challenging task to identify important features. So there are lots of data mining and prediction challenges, those are not included in our current project. The main challenge of our project is measured by some indicators such as actor, actress, director, writer, estimated budget and genres of the movie from the IMDB and IMDB PRO dataset to predict the rating and estimated collection.

Chapter 2

Background and Literature Review

There are several works related to this project. An enormous concern in analysis and prediction of movie success become more popular in 2006 when Netflix declare prize money of about 1 million USD to the best team for who improvise their movie rating algorithm cinematch.

Google has an application system· that work on search volume of the movie trailer. After calculation and analysis on search volume it can estimate an opening weekend revenue of a new released movie.

Research regarding trendsetters (clark & zboja & Goldsmith, 2007) is often shows the concept of social network analysis. It relates the association of news trend and social media, expert user's options.

The features used and proposed by other groups (Oghína et al, 2012, El Assady et al, 2013 Jager et al, 2013, Yaffeng Lu and Maciejewski,2013) firstly works as group then they are separated into two groups movie to analysis the internal features and all external features of the movie set. First group· work on features, such as the actor, directors, genre of the movie and so on, which determine· the movie itself and can brought correlation with its quality and success of the movie.

The second group works on external behavior and real world conditions, which do not power the movie itself, but massively impact for its reception, popularity and success, hence the revenue. These features consist· of the release date of the year, whether there is holiday or not, time of the year and other movie releases at the same time, the weather conditions, marketing campaign, audience movie anticipation and many more.

Awad, Delarocas and Zhang (2004) analyzed over the data that make impact on movie rating on movie success. They developed statistical models based on movie ratings to estimate forecast revenue. And they also examined the relationship of professional critics and also traditional consumer communication and online word-of-mouth. Finally, they

came to the decision that professional critics, traditional consumer communication and online word-of-mouth has great influence for increasing number of movie viewers. Eliasberge, Elberse and lenders (2006) suggest further research make a correlation to internet resources and movie consumption and also box office revenue.

There are lots of research done and also ongoing. But some research also indicates that movie reviews do not have a significant relationship with box office success. Due to this research, we chose to develop our application without the use of reviews and were able to predict a movie success even before it's production begin with help of right combination sets.

Chapter 3

Methodology

Essentially, it is a venture whether a movie will be a success or failure. Our objective to develop the algorithm to solve this problem by reducing the number of movies that lose money in box office. Our application hopes to categorize new and interesting pairs of characteristics that will fare well at the box office such as unique actor, actress, director and writer combinations. Another possible application is to assist investor make judgements on whether cast combination of the movie will be good and profitable or not. In this chapter firstly we will discover our application design specification then we describe the detail implementation and technique process as well as testing.

3.1 SYSTEM DESIGN SPECIFICATION

3.1.1 DATA FLOW DIAGRAM

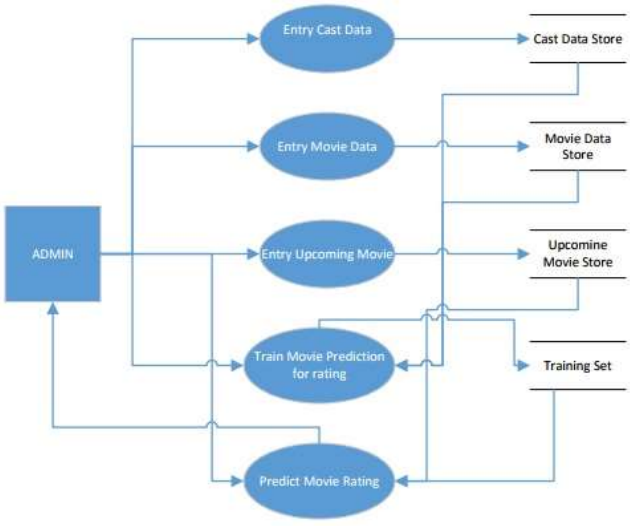

Figure 3.4.1.1 Data Flow Diagram

3.1.2 USE CASE DIAGRAM

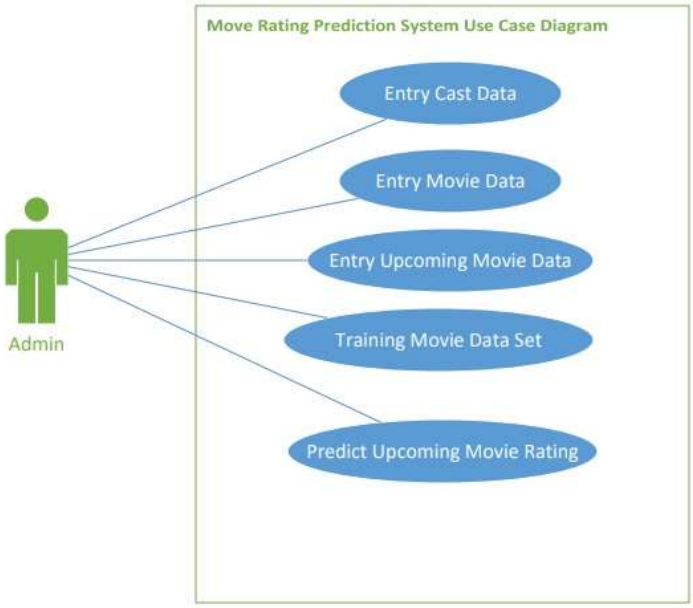

Figure: 3.4.2.2 Use Case Diagram

3.1.3 Class Diagram

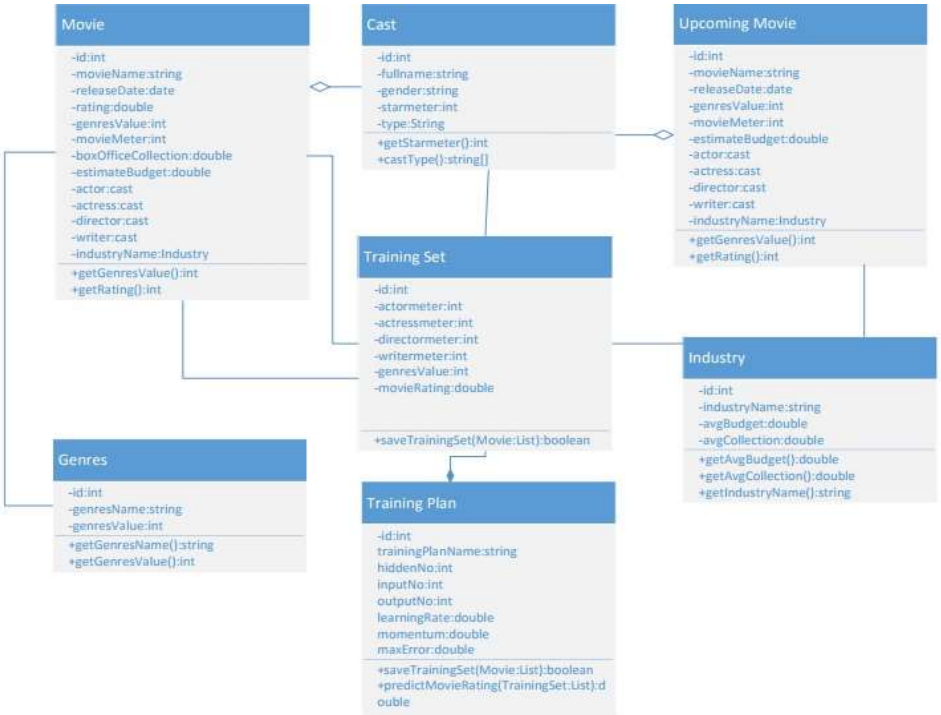

Figure: 3.4.3.1 Class Diagram

3.1.3 Time Sequence Diagram

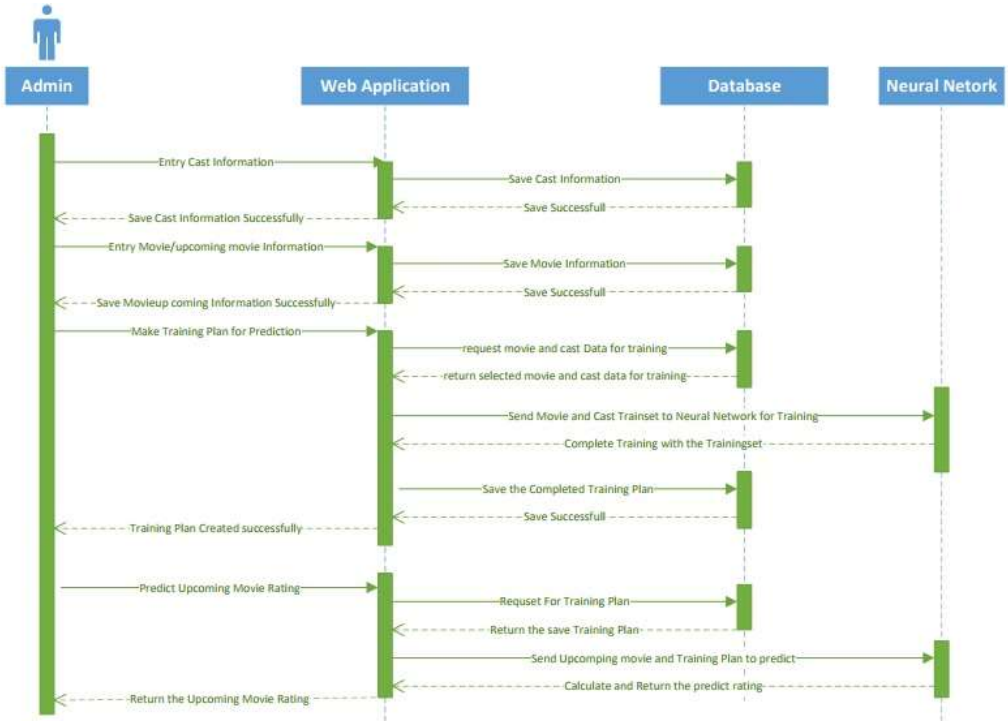

Figure 3.4.4.1 Time Sequence Diagram

3.2 DATA COLLECTION

For our project there was no public dataset available which had all of the attributes that we need for predicting success. So I initially tried different options for obtaining data from several APIs. But I failed for undocumented and unauthorized access of API. Finally, I created a new dataset by collecting relevant information from IMDB and IMDB PRO site which is an amazon dataset. IMDB is huge movie dataset open for all user. It has information related to movie ratings, movie genres type, actor, actress, director and writer. But it has not information regarding movie budget's, opening weekend collections, and local collection and worldwide gross collection which are very important attributes to determine the movie success and failure. But we can get this information from the other site IMDB PRO. This is paid site where we can get the all detailed information of a movie. Here, I show some picture how we take data from IMDB and IMDB PRO site.

Figure:3.2.1 Actor Profile form IMDB PRO site

This figure shows the web page of a cast information of IMDB PRO site. It consists of lots of information of casts like date of birth, height, nick name, award collected in their carrier and their popular movie information. From here we collect the cast information. Every cast like actor, actress, director and writer have this star-meter value which I used as main indicator for prediction. Here I want to show by some figure that how I pick movie information the site.

Figure: 3.2.2 Movie Profile from IMDB PRO site

Figure: 3.2.3 Movie Profile from IMDB site

Here two figure shows about the same movie. First figure: 3.1.2 have the information about budget, opening weekend collection and also gross collection of the movie. And the second Figure: 3.1.3 contain information about rating and cast like actor, actress, director and writer. Both figure have the data of movie genres like action, comedy or drama.

3.3 DATA PROCESSING

After collection the data from IMDB and IMDB Pro site, we stored corresponding cast and movie data in our application database. Analysis the data, we have finding that maximum star-meter of cast is like 100000 and maximum budget of a Bollywood movie is 1000 cores. So we applied following formula to process the data compatible for our developed neural network.

Cast Star-meter conversation formula:

Process cast meter = 100000- (original star-meter value) / 100000

Process estimate budget = budget of movie / 1000

For genres value we have make genres value table. This values are similar to Linux file system permission. For example, a movie genre is drama and action then it's genres value will be some of both value like 12. And a movie with genres of romance and comedy then genre value will be 48.

Process genres value = (Sum of genres value)/100

ID	Genres	Value
1	Horror	1
2	Thriller	2
3	Drama	4
4	Action	8
5	Comedy	16
6	Romance	32

Table: 3.3.1 Table of Genres Value

3.4 TECHNIQUES AND CALCULATIONS

A common approach for predictive analysis requirement is by using neural networking algorithms. Backpropagation is one of these algorithms. In backpropagation, we send the input channels values through the hidden layer as well and propagates and back propagates until it reached to desire outputs. At this phase, it compared with loss and error function then push back in reverse direction of the layers. Layer node's weight are adjusted to reduce the error so that it can reached to the optimize accuracy. These error value will help to improve the data which use for training so that we get a new vector of input with minimize the loss of output accuracy. The backpropagation is very effective to improve a training dataset. Our application neural network is showing in following figure.

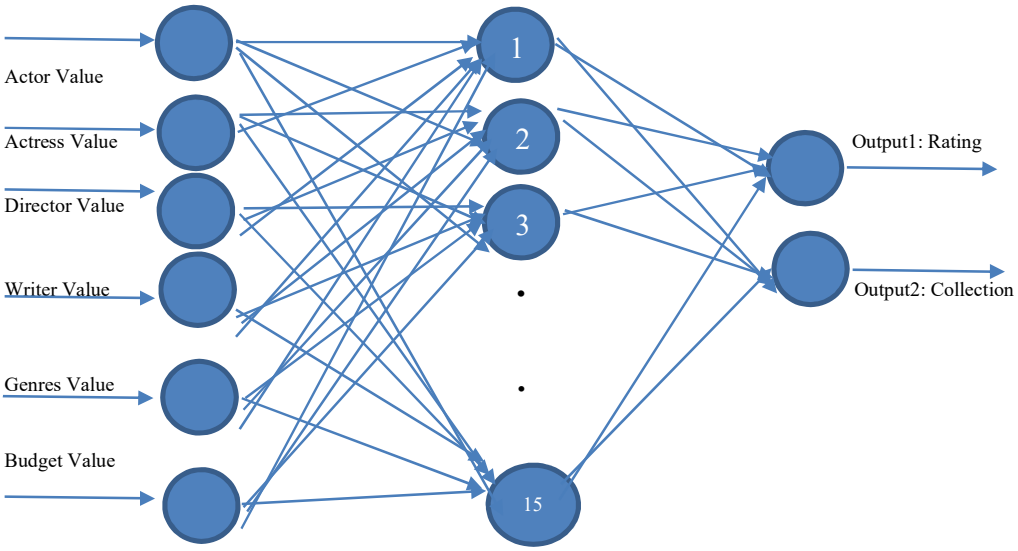

Figure: 3.4.1 Application Neural Network Design.

In our designed neural network, we push actor, actress, director, writer, genres and estimated budget's processed data as input. We can take 10 to 20 node in our system as a hidden layer. And we can put dynamic value for learning rate, momentum and error rate in our system. Then we applied sigmoid function and backpropagation method to adjust with the desire output by minimizing the error rate. We used our processed data that are showing in following table.

Input1:Actor Value	Input2:Actress Value	Input3:Director Value	Input4:Writer Value	Input5:Genres Value	Input6:Budget Value	Output1:Rating Value	Output2:Collection Value
0.85	0.83	0.78	0.78	0.36	0.03	0.76	0.18
0.89	0.87	0.75	0.75	0.40	0.40	0.52	0.74
0.90	0.97	0.76	0.76	0.38	0.04	0.68	0.11
0.91	0.81	0.75	0.75	0.08	0.93	0.33	0.89
0.91	0.86	0.79	0.79	0.16	0.08	0.51	0.31
0.91	0.97	0.61	0.61	0.20	0.13	0.65	0.08
0.91	0.93	0.79	0.79	0.12	0.10	0.57	0.22
0.91	0.87	0.00	0.00	0.24	0.30	0.39	1.00
0.97	0.93	0.88	0.88	0.20	0.02	0.84	0.08
0.97	0.87	0.65	0.65	0.12	0.07	0.86	1.00

Showing 1 to 10 of 31 entries Previous 1 2 3 4 Next

Table: 3.4.1 Part of Processed Input and Output Vector Table of Trainset.

Now we provide our processed input and output data to designed network to training. When it completes the training then we saved it for future prediction. Here we can make our training plan with different movies sets and then test the prediction with that different plan.

3.5 IMPLEMENTATION

Some of the user interfaces, internal includes and their life cycles are to be discussed in this chapter

3.5.1 CAST ENTRY PAGE

From this interface, we can entry our collected data of cast like actor, actress, director and writer from IMDB and IMDB PRO site.

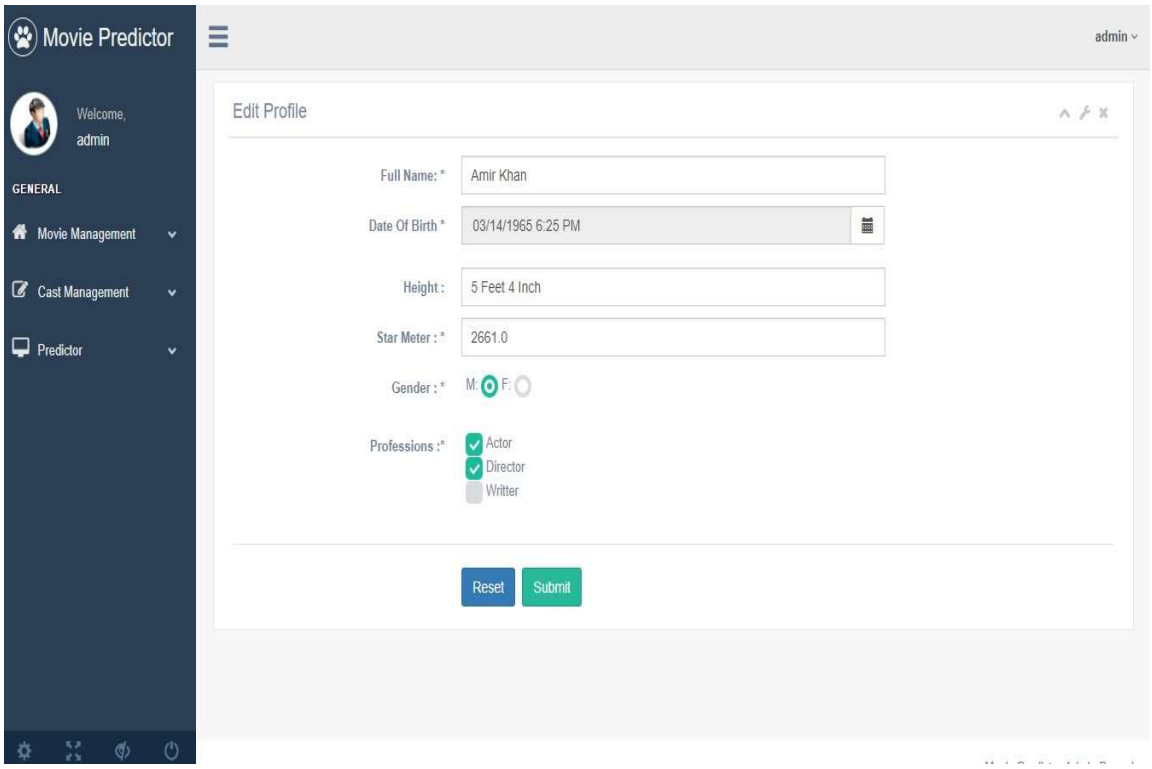

Figure:3.5.1.1 Cast Entry Page of application.

3.5.2 MOVIE ENTRY PAGE

From this interface of our application, we add movie datasets to our database from collected data from IMDB and IMDB PRO site.

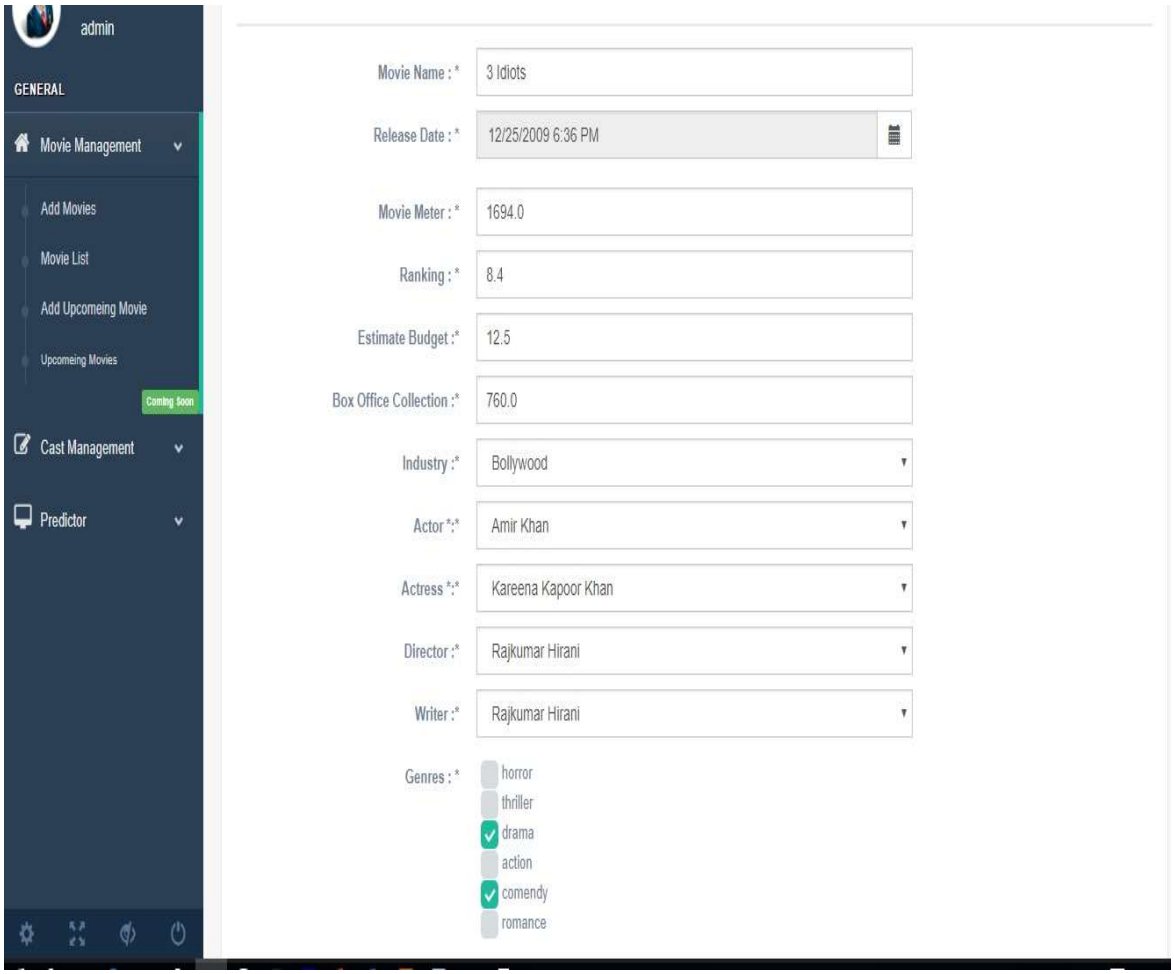

Figure: 3.5.2.1 Movie Entry Page

3.5.3 UPCOMING MOVIE ENTRY PAGE

This page is for entry the data for upcoming movie. We can make entry for a dummy movie that a production house plan to make with actor, actress combination. And this upcomeing movie we can predict with prediction page of our system.

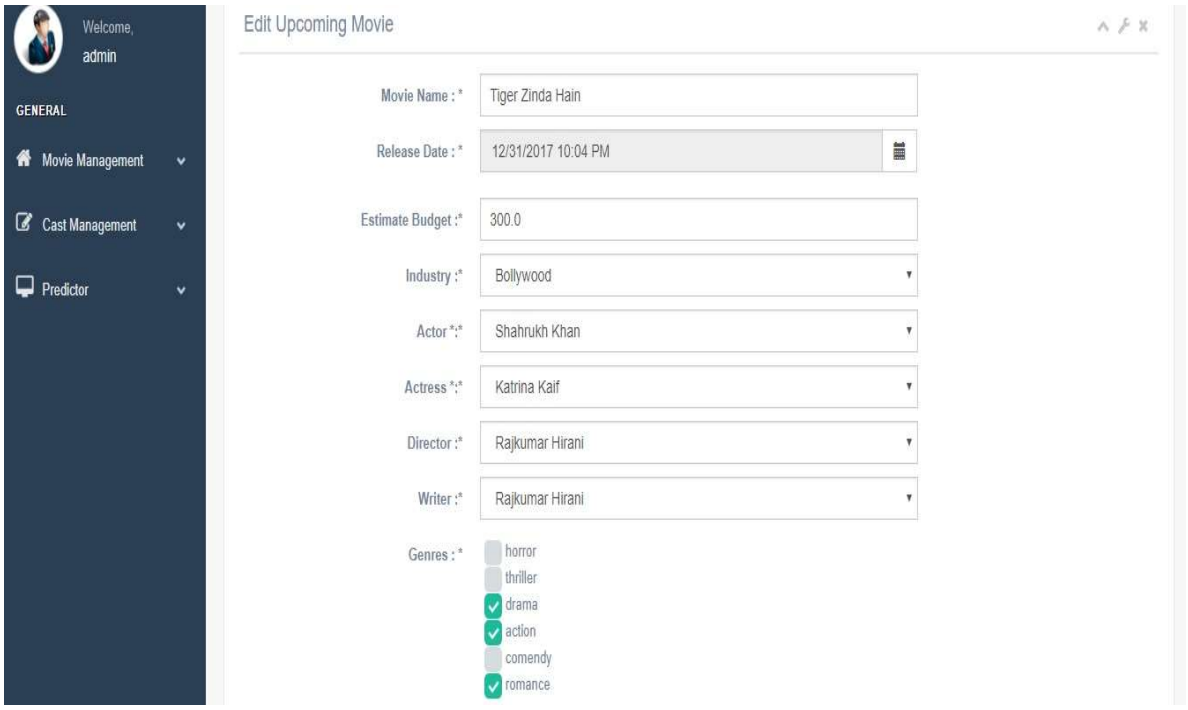

Figure:3.5.3.1 Upcoming movie entry page

3.5.4 TRAINING PLAN PAGE

In this page, we can make training plan where we can set number of node in input layer and output layer for our neural learning network. If we need to predict only for rating of movie then we can choose five input node and one output node. Five input nodes are actor value, actress value, director value, writer value and genres value. We get rating value as output in this case. And we have to set 6 input node and two output node to predict rating as well as box office collection. Then we have to put estimated budget value as extra input node and here we get second output node which it shows about the estimated box office collection We can mention our hidden layer node size also. For hidden layer we can set value 10 to 20 as number of hidden node. We can put learing rate, momentum, and how much error rate we can accept to our neural network learning.

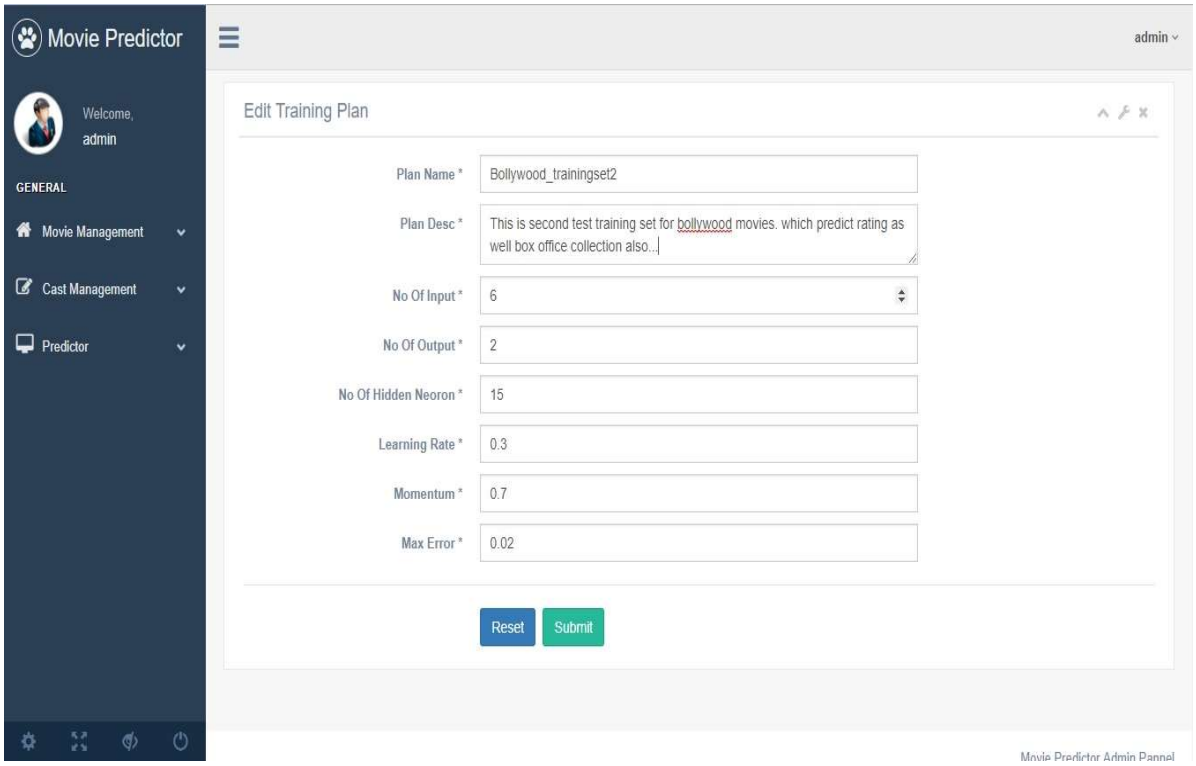

Figure: 3.5.4.1 Training Plan creation page

3.5.5 TRAINING PLAN INTIALIZE

From this page we can intialize the training plan with move set for training. And also flush the traingset and initalize a new trainingset as well.

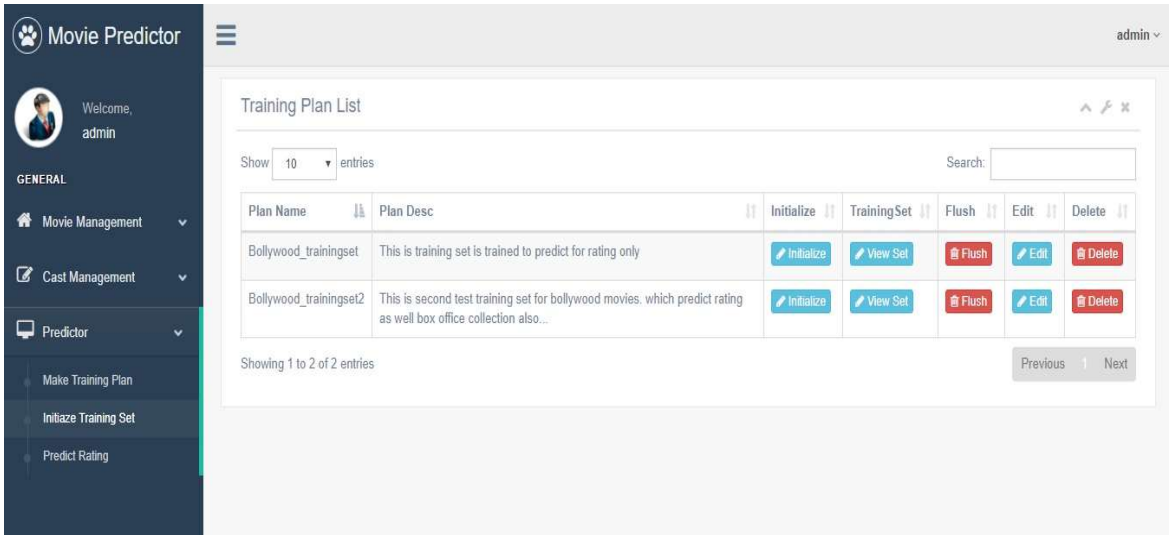

Figure 3.5.5.1 Training Plan Initialize page

3.5.6 PROCESSED TRAINING SET PAGE

This page we processed our movie dataset to compatable dataset for neural network training set.

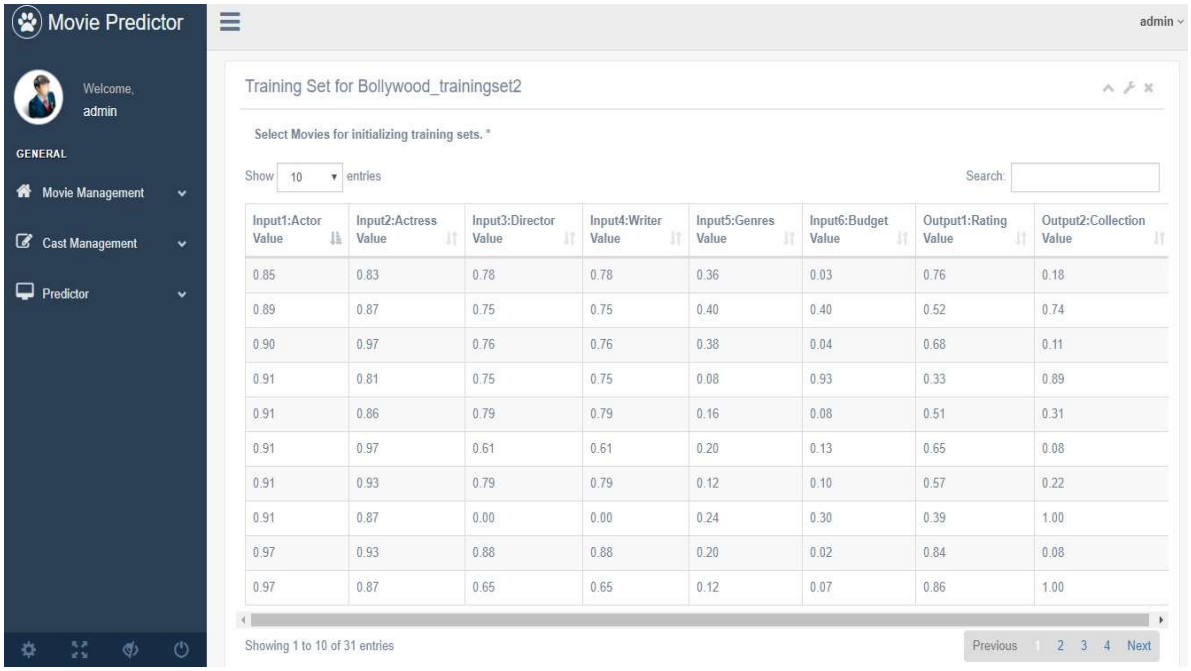

Figure: 3.5.6.1 Processed Training Set

3.5.7 PREDICTION PAGE

From here, we can predict our upcomening movies rating and estimate collection from our chosen trainingset.

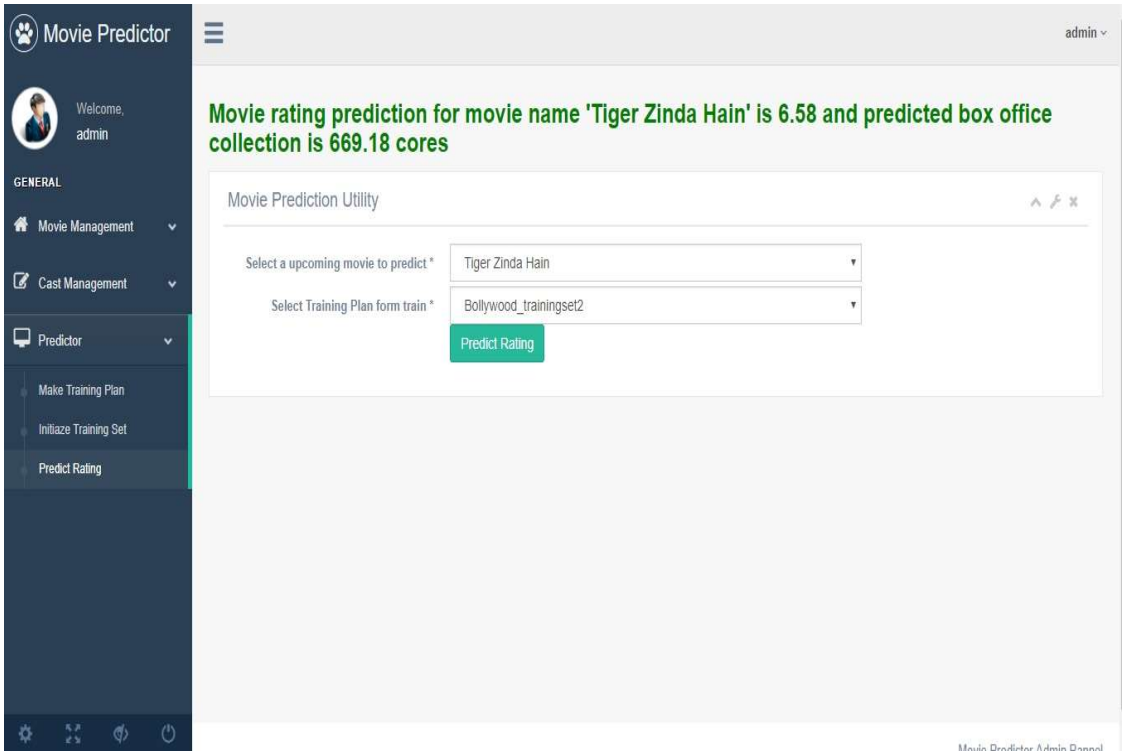

Figure: 3.5.7.1 Prediction Result page

3.6 TESTING

3.6.1 SOFTWARE TESTING

The process of evaluating software during the development process or at the end of the development process to determine whether it satisfies specified business requirements.

Validation Testing ensures that the product actually meets the actual needs. It can also be defined as to demonstrate that the product fulfills its intended use when deployed on appropriate environment.
It answers to the question, are we building the right and quality product?

3.6.2 VALIDATION TESTING - Workflow:

Validation testing can be best demonstrated using V-Model. The Software/product under test is evaluated during this type of testing.

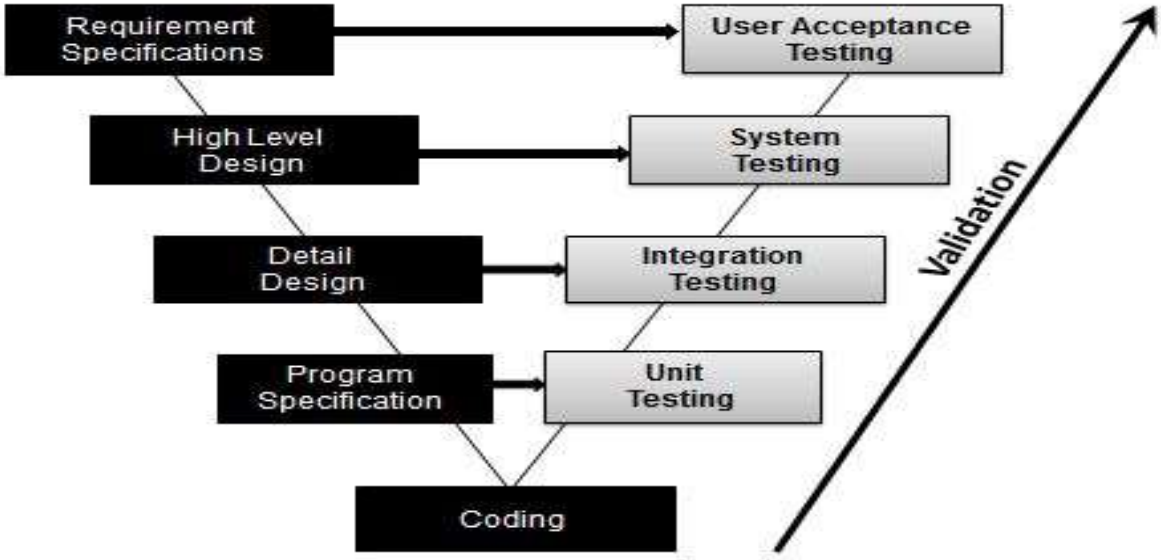

Figure:3.6.2.1 Validation testing workflow

For this testing several test cases are made to serve the validation testing purpose

3.6.3 TEST CASE TABLE OF SYSTEM

ID	Topic	Related Requirements	Test Category	Pass/Fail	Remarks
1	Login	With proper user & pass	Validation	Pass	N/A
2	Add/Edit Cast	Proper post and Integration	Functional	Pass	N/A
3	Add/Edit Movie	Proper post and Integration	Functional	Pass	N/A
4	Add/Edit Upcoming Movie	Proper post and Integration	Functional	Pass	N/A
5	Make Training Plan	Proper post and Integration	Functional	Pass	N/A
6	Initiate Dataset to Plan	Integration	Functional	Pass	N/A
7	Predict Movie Result	Training Plan and upcoming movie	Neural Learning	Pass	N/A

Table: 3.6.1.1 Test case table for the application

Chapter 4

Results

We run our training with 50 cast record and 30 movie set from IMDB and IMDB PRO. Multiple test cases were run to perform testing to verify the results with that of the IMDB and see how well our algorithm has fared. We have tested with a recently released movie “Tiger Zinda Hain” which has the actor Salman Khan and actress Katrina Kaif. The obtain output of this movie is very close to IMDB rating and also box office collection till now. Our system prediction rating is 6.58 and box office collection is like 669.18 cores whereas real life rating form IMDB rating is 6.6 and box office collection is until now is around 664 cores. So our prediction result is very much acceptable. In below figure we are showing what we are getting from our developed application.

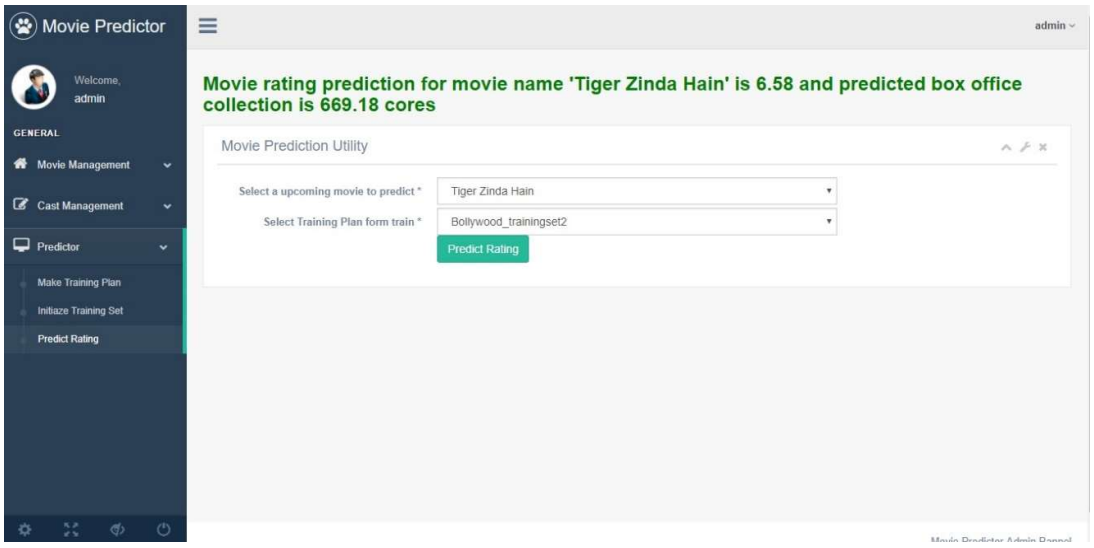

Figure:4.0.1 Movie Prediction Result from the application.

Another movie named “ZERO” of actor Shah Rukh Khan and actress Katrina Kaif which will relase in this year. This movie is directed by Anand L. Rai. We predict with our threshold value with our learned neural network and found the rating of the movie as 6.16 out of 10. And Estimated collection showing in our application is about to 790.99 cores.

Chapter 5

Conclusion

From the experimental analysis, we could comprehend that the prediction of movie success is certainly possible with high percentages of accuracy. So by using our prediction engine of our application, production house can evaluate before it production begins. And production house can decide that movie is worth investing or not and accordingly make their decisions. IMDB data are for whole movie industry of the world. But if could classified the data by industry wise, then we could make a more accurate segmented dataset for movie prediction. We can expect that with valid accurate data and fuzzy logic is an effective means of categorizing predictions and adds more accuracy and dynamicity to the system. In case of multiple cast like more actor, actress, director and writer this current application is not suitable· enough. Here we can just one actor, actress and director as input. But many movie datasets have more than one actor, actress and other crew. In this case we have to make input layer dynamic so that we can feed any number of movie cast for training as well prediction. Although there are a few cases where the obtained result does not match with the expected result, they can be considered negligible when compared with the larger percentage of accurate results. For future work, we can add trailer hit count and social media hive, news trending and release date, vacation of years to this system to get the system more mature for prediction.

References

[1] Steven Yoo, Robert Kanter, David Cummings TA, Andrew Mass, “Predicting Movie Revenue from IMDB Data”, pp. 1-0 ,2015

[2] Neoroph, “Predicting the result of football match with neural networks”, 2016 [Online] Available: http://neuroph.sourceforge.net/tutorials/SportsPrediction/Premier-%20-League%20-Prediction.html [Accessed: 15-Dec-17]

[3] Nevon Project “Movie Success Prediction Using Data Mining” [Online] Available: http://nevonprojects.com/movie-success-prediction-using-data-mining/ [Accessed: 10-Dec-17]

[4] Wit Press “A Data Mining Approach To Analysis And Prediction Of Movie Ratings” [Online] Available: https://www.witpress.com/elibrary/wit-transactions-on-information-and-communication-technologies/33/14248 [Accessed: 05-Dec-17]

[5] Steven Miller, “How to Build a Neural Network” [Online] Available: https://stevenmiller888.github.io/mind-how-to-build-a-neural-network/ [Accessed: 28-Nov-17]

[6] Ijariit, “Prediction of Movie Success for Real World Movie Data Sets” [Online] Available: https://www.ijariit.com/manuscripts/v3i3/V3I3-1228.pdf [Accessed: 18-Dec-17]