**Final Project Report**

**Predicting success of movies - Data Mining**

**By:**

**Sanika Deshpande**

**Student ID : 805600277**

**Varun Krishna**

**Student ID: 805589256**

**CPSC 531-01 (13293)**

**Fall, 2016**

**Professor: Dr. Chun-l Phillip Chen**

**Department of Computer Science**

**California State University, Fullerton**

**Dec 15, 2016**

**Table of Contents**

Abstract ..................................................................................................................................................5

1. Introduction......................................................................................................................................6

1.1 Background ……...........................................................................................................6

1.2 Project Goal and Benefit of this Study..................................................................................6

1.3 Relevance and Significance …………………………………………….............................7

1.4 Assumptions and Limitations………………………………………………………………7

1.5 Definition of Terms………………………………………………….................................8

2.0 Literature Review .............................................................................................................................9

2.1 What is Data Mining? .................................................................................................9

2.2 Data Mining Techniques .............................................................................................10

2.3 Data Mining Algorithms used in this project ….…………………………...…………….10

2.3.1 Linear Regression Model.....................................................................................10

2.3.2 Regression Tree Model……………………………………………………………...12

2.3.3 Correlation Algorithm…………….…………………………………………………13

2.4 System Architecture ……………………………………………………………………...14

2.5 Approach/ Technical Approach ……….......................................................................15

2.5.1 Data Collection Process…………………………………………….........................15

2.5.2 System Algorithm .............................................…………………………………..15

3.0 Methodology....................................................................................................................................17

3.1 Project Methods, Tools and Techniques ............................................................................17

3.2 Design and Functional Requirements…......................................................................18

3.3 Data Requirements....................................................................................................19

3.3.1 Source of Data Collection…………………………………………………………...19

3.3.2 Method of Collection ……………………………………………………………….19

3.3.3 Input ……………………………………………………………………………...…20

3.3.4 Report ……………………………………………………………………………….20

3.4 System Requirements……………………………………………………………………..20

3.4.1 Hardware Requirements……………………………………………………………..20

3.4.2 Software Requirements……………………………………………………………...21

3.5 System Installation and Implementation ………………………………………………..21

3.6 Tools for Testing and Validation of Mining Models……………………………………..21

4. 0 Results and Discussions ................................................................................................................23

5. 0 Conclusion, Implications and Recommendations .........................................................................29

5.1 Conclusion................................................................................................................29

5.2 Implications ……...……………………………………………………………………….29

5.3 Recommendation …………………………………………………………………………30

References....................................................................................................................................31

**List of Figures**

1. Figure 1: Process of Data mining …………………………………………………………..9

2. Figure 2: Linear Regression Model…………….………………………….........................11

3. Figure 3: Regression Tree Representation………………...………………………………12

4. Figure 4: System Arcitecture…………...……………………….........................................14

5. Figure 5: Types of Graphs…………….…………………………………………………...16

6. Figure 6: System Design ………………………………………………………………….18

**ABSTRACT**

Record of producing several hundred movies by American film studios makes the United States of America as a third most prolific producer of films in this world. Predicting the success of upcoming movie is difficult as it depends on many different factors. This paper shows the detailed analysis of the Internet Movie Database (IMDb), a free, user-maintained, online resource of production details for over 5,000 movies, which contains information such as title, genre, cast credits and user's ratings. We have used variety of data mining techniques to show the relationships between different attributes. This paper mainly concentrates on attributes related to user ratings, genre such as having particular actor or actress in movie are responsible for that movie to succeed. Here, we found the most interesting fact after analysing IMDB data that, budget of film is no more related to the well-rated film instead the rating of cast of that film influence more to decide the success or lack of the film. This paper introduces a simple solution of predicting movie success in terms of rating of casts like actor, director and genre. As a result, this approach achieved detailed result of the prediction in terms of positive (super hit), negative (flop) and neutral (hit). Result of this prediction helps audience to decide whether to book ticket in advance or not. It also helps production team for their investment in movies.

**Keywords: IMDB, Internet Movie Database, Data Mining, Movies, Film**

1. **INTRODUCTION**

**1.1 Background:**

The accurate mechanism for prediction of future trends in prediction market is “wisdom of crowds”. Today, it has become very common that the result of any work is analysed by group of ordinary people instead of one expert’s opinion. Predicting the success of movie before its release is only possible after analysing multiple parameters like rating of cast, imdb\_score of film etc. For the audience who have a lot of interest in watching movies and still wait for the reviews of movies, this prediction of success of the movie is very helpful. Result of this prediction will help them to decide whether to watch this movie in cinemas or to book the tickets in advance or not. It is also helpful for the production team, who wants to make a lot of profit from movies only, need some analysis of whether the movie in which they are investing will make that much profit or not.

Data mining is used in many sectors like banking, customer relationship healthcare etc. There are a lot of data mining techniques that can be used for this analysing purpose in many aspects like clustering, classification and prediction.

**1.2 Project Goal and Benefit of this Study:**

To find the detailed information about any upcoming movie, IMDB is an excellent resource. It contains large amount of valuable information about general trends in films. We are using data mining techniques to uncover information for confirming or disproving common assumptions about movies. It helps in prediction of success of movie before its release.

Data sets that we are using for this purpose are IMDB movie data sets. This data set has lots of attributes with various records and values. The relationship between these attributes is used to find out the result of our prediction. This data set mainly concern with the rating of each actor/actresses, rating of director, and type of movie i.e. genre. There are total 5043 records and 16 attributes which we are going to analyse for this prediction. Combination of these records and attributes will help for assuming the different conditions for analysing the data and finding out hidden analysis from it.

**1.3 Relevance and Significance:**

Today, many sectors are using different data mining techniques to achieve their desire goals. Watching movie and giving credits to each of them is the activity that will never end. So data mining in predicting the success of movie can be very useful as in other sectors. Data mining techniques are used to find patterns and anomalies in data sets to analyse large datasets. Supervised learning is mainly used for predicting unknown patterns and information.

Within the movie prediction domain, regression analysis is mainly used. With a movie database containing large amount of information about recent and upcoming movie productions, movie ratings based on user’s likes and dislikes sparks some interest of how movies will get success in the audience’s perspective and at what degree such aggregation of this user rating will help for prediction.

User rates the movie according to cast, budget, and directors of movie. Evaluation of two algorithms linear regression analysis and regression tree analysis is used to further examine the possibility of predicting movie ratings.

**1.4 Assumptions and Limitations:**

The success of movie depends on the ratings of actor/actresses, directors and genre. The combination of some actor/ actresses defines the success of movie. A user rating of particular combination of casts in specific genre is use as a factor of condition for success prediction of movie. This project can be further expanded by considering other factors like gross and budget to check the income of movie and can include this factor predicting success of movie.

**1.5 Definition of Terms:**

**Data Mining:** Extracting information from a data set and transforming it into an understandable structure for further use.

**Regression Analysis**: a statistical process for estimating the relationships among variables.

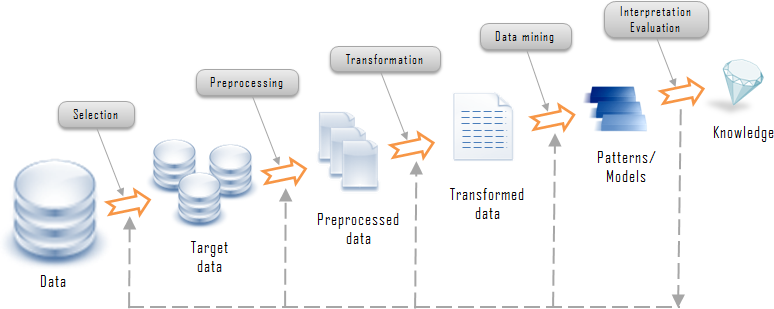
1. **Literature Review**

**2.1 What is Data Mining?**

Data Mining is process of analysing data from different views and extracting the useful information from large datasets. It is also a process of finding correlations or patterns among different fields in large datasets. It allows users to categorize data and summarize identified relationship after analysing it. It is mainly used for prediction allowing users to make pro-active and Knowledge-driven decisions. Data mining answer business questions that were too time consuming in traditional system.

Data mining consists of five major elements:

* Extract, transform, and load transaction data onto the data warehouse system.
* Store and manage the data in a multidimensional database system.
* Provide data access to business analysts and information technology professionals.
* Analyse the data by application software.
* Present the data in a useful format, such as a graph or table.



**Figure 1: Process Of Data Mining**

**2.2 Data mining Techniques :**

1. **Association:**

Relationship between data items in same transaction are identified first and then the pattern is discovered accordingly. It is also known as relation technique.

1. **Clustering :**

The process of separating the objects that have similar characteristics is known as clustering. Here the technique first define the class and then puts objects in that class.

1. **Classification :**

The predefined sets of classes are used to classify the data in different categorized classes. It uses mathematical techniques such as linear programming, neural network and decision tree etc. Classification of data items into different groups have been performed in this process.

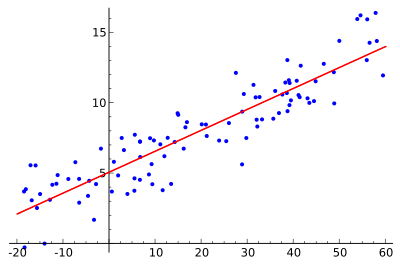
1. **Prediction :**

It is the process that defines the relationship between independent variables or independent and dependent variables. It involves analysing trends, pattern matching and classification. Making future prediction by analysing past events is called as prediction technique.

* 1. **Data Mining Algorithms used in this Project:**

**2.3.1 Linear Regression Model:**

It is process of modelling the relationship between scalar dependent variables and more explanatory variables. The simple linear regression is a case for one explanatory variable. The multiple linear regression algorithm is the case for more than one explanatory variables. In Linear Regression, unknown model parameters of modelled relationship are estimated from data using linear predictor functions. These models are called as linear models. Linear regression is the first type of regression analysis which use extensively in lot of applications. In this regression, models which depend linearly on their parameters are easier to use than other models which are non-linear.



**Figure 2: Linear Regression Model**

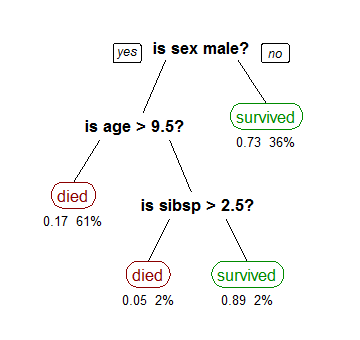
Linear Regression has many practical uses. Linear regression can be used to fit predictive model to observed data set if the result we want is prediction. To identify what redundant values dataset contains, we used linear regression that can be applied to qualify the strength of relationship between data variables.

Use of this algorithm in our project is that we have considered two variables. One is independent i.e. number\_of\_voted\_users and one is dependent i.e. imdb\_score. By modelling the relationship between these two variables, we have predicted the imdb\_score. The predicted result is shown by linear regression graph.

Before this operation has been performed, we have calculated R\_squared and Mean Squared Error (MSE) to find out whether this algorithm is suitable for the data sets we are using for accurate prediction. However, the result of r\_squared is near to 0 and MSE is near to 1, which is exactly opposite of what is required. This clearly shows that this linear regression algorithm is not suitable for our data set.

**2.3.2 Regression Tree Model:**

Decision tree learning uses a [decision tree](https://en.wikipedia.org/wiki/Decision_tree) as a [predictive model](https://en.wikipedia.org/wiki/Predictive_modelling) which maps observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modelling approaches used in [statistics](https://en.wikipedia.org/wiki/Statistics), [data mining](https://en.wikipedia.org/wiki/Data_mining) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning). Decision trees where the target variable can take continuous values (typically [real numbers](https://en.wikipedia.org/wiki/Real_numbers)) are called regression trees. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and [decision making](https://en.wikipedia.org/wiki/Decision_making). In [data mining](https://en.wikipedia.org/wiki/Data_mining), a decision tree describes data.



**Figure 3: Regression Tree Representation**

The goal is to create a model that gives prediction in the form of target value variable based on several input variables. The example diagram shown above has each interior node corresponds to one of the input variables. There are edges to children for each of the possible values of that input variable. Each leaf represents target variable value given the input variables values represented by the path from the root to the leaf.

Before this operation has been performed, we have calculated R\_squared and Mean Squared Error (MSE) to find out whether this algorithm is suitable for the data sets we are using for accurate prediction. However, the result of r\_squared is near to 1 and MSE is near to 0, what is exactly required. This clearly shows that this regression tree algorithm is suitable for our data set.

After implementation of this algorithm on the imdb\_score values, we got the predicted graph and decision tree. This graph describes that the predicted result of imdb\_scores are almost near to the actual imdb\_scores. It clearly states that our prediction is correct.

**2.3.3 Correlation Algorithm:**

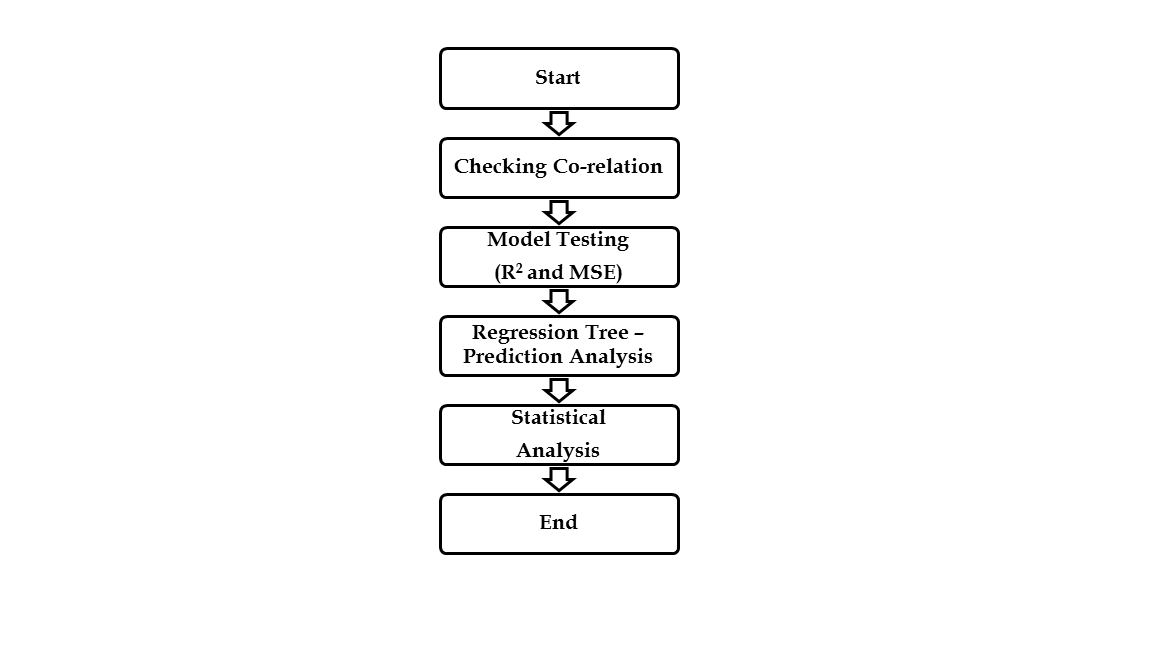
This term is mainly used to analyse the strength of the relationship between two variables. A strong, or high, correlation means that two or more variables have a strong relationship with each other. However, weak or low relationship defines that the two variables are highly related. For possibly distinct random variables X(s) and Y(t) at different points s and t of some space, the correlation function is

C(s,t) = corr( X(s), Y(t))

In the Correlation analysis, we estimate sample correlation coefficient denoted by r. r ranges between -1 and +1 and quantifies the direction and strength of the linear regression between the two variables. The result of correlation can be positive or negative. This sign of correlation indicates the direction of association.

In our project, we have calculated the correlation of each attribute of IMDb data set. The result gives the value of imdb\_score attribute as 1 and the value of num\_of\_votes\_users attribute as 0.59. Other attributes have less value than these two attributes. In conclusion, we states that these two attributes are more related to each other. So, these two attributes should use further for the prediction analysis on IMDb data set. Also we have plotted the graph which shows values of correlation of each attribute from data set.

* 1. **System Architecture**

****

**Figure 4: System Architecture**

Above figure shows the steps of the system we have created. The description of each step is as follows:

**1.** **Checking Correlation:** The system first checks the correlation values for each attribute of IMDb data set.

**2.** **Model Testing:** This step includes R2 calculation and MSE calculation in it. Where, R2 of linear and regression model is calculated as well as MSE of both model is calculated to see whether which is suitable for predicting results using our IMDb data set.

**3.** **Regression Tree Model:** This step defines that regression model is suitable for IMDb data set we have used. And by using this model, prediction of movie has been performed.

**4.** **Statistical Analysis**: It includes the result of statistical analysis performed on IMDb data set.

**2.5 Approach / Technical Approach**

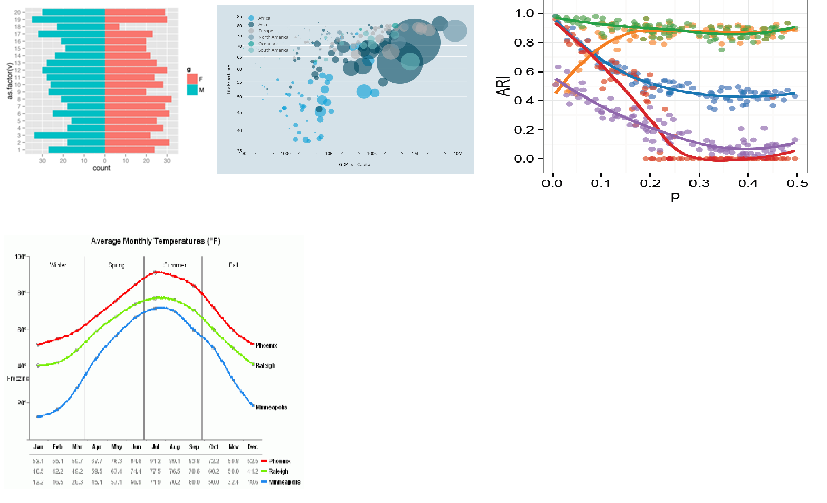
As mentioned earlier the aim of this study is to examine and compare the performance of the movies before it is released in the cinemas. This could help the production team to make necessary requirement before the movie releases. Firstly, we need to find the appropriate dataset which provides the all necessary information that can be used to perform the prediction. Secondly, we have to choose the appropriate algorithm and analysis type to perform the prediction operation.

**2.5.1 Data Collection Process:**

We have obtained the dataset from the kaggle.com which provides the free large number of datasets. The IMDB has provided the 5000 movies’ data which provides the information about the actors, director, budget, facebook\_likes and ratings. For movies, they have provided the title of the movie, the year of release, the number of likes it has received from users, as well as its rating as shown on the IMDb website. For actors, actresses and directors, they have provided the details such as facebook likes of actors, actresses and directors. Finally, the data is grouped according to genre in order to investigate the correlation of genre and the popularity of the respective movies.

**2.5.2 System Algorithm:**

We are using regression analysis which is a statistical process for estimating the relationships among variables. To perform these operations, we need to first clean the dataset that we have obtained as the dataset contains few unwanted columns and few values are missing. After removing the unwanted value from the dataset we obtain the clean dataset that can be used to perform the mining operation. Now, the datasets has to get divided into test data and training data. These data are to be stored in the file with an extension Comma separated values (.csv). Now we test the data by using algorithms like regression analysis. After obtaining the result, we have created visualized graphs. These graphs provide the pictorial representation so that the users can understand the prediction and take necessary decision. GGplot2 is a plotting system for R, based on the grammar of graphics, which tries to take the good parts of base and lattice graphics and none of the bad parts. It handles the many of the fiddly details that make plotting a hassle (like drawing legends) as well as providing a powerful model of graphics that makes it easy to produce complex multi-layered graphics.



**Figure 5: Types of Graphs**

1. **Methodology**

**3.1 Project Methods, Tools and Techniques:**

We have used the Exploratory Data Analysis which is a function in R. Data in R program are often stored in frames, because they can store multiple types of data. This function helps to store the data in the frames before performing any statistical analysis. After the required data is stored in the frames we have created a set of data with respect to the year the movie released and along with its genre. Each movie contains subset which includes the movie\_facebook\_likes, actors\_facebook\_likes, cast\_facebook\_likes and director\_facebook\_likes. We used these data values to perform the analysis and predict whether the movie will be successful or not before it is released. Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analysing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. We are going to use the regression analysis to determine the relationship among the attributes present in movie dataset. To perform all these analysis, we are using R studio, open source software which is integrated development environment for R. R is a programming language used for statistical computing and graphics was founded by JJ Allaire. Along with R features we are also using the other packages such as GGplot which is used for plotting the graph. R markdown is an another package to generate the PDF report.

**3.2 Design and functional requirements:**

**……….**

**Figure 6: System Design**

In the above diagram, we require the year of the movie released so we can group the movies and genres according to the year. The group of each year consists of n number of movies. Then we create a subset for each movie which consists of the facebook likes of the actors, director and cast. After obtaining the values we will perform the mean operation to find out the average likes of the actors and others received and we will store those values in the vector. Further the process will continue until it gets the results from all the movies. Once the process has been completed we would perform the comparison among the values and we will find out the results which tells which movie is going to be successful. In the same order we would perform series of operations on the data using regression analysis to predict the other result such as successful genre of the year, average IMDB rating by year.

* 1. **Data requirements:**

**3.3.1 Source of data collection:**

In order to properly evaluate the relevant algorithms on successful movie predictions a relevant dataset has to be considered for the experiment. As previously described, the datasets used within the experiment must include a number of relevant attributes for predicting the success of a movie prior to its release, such as genre, budget, actors starring the movie and most importantly some sort of overall rating. Likewise, the datasets must also include a sufficient number of data points as well as somewhat correct and complete data, in order to be reliably analysed by the algorithms, and thus to enable a reasonably valid comparison. The dataset has been collected from the kaggle.com website which provides the free dataset to the students for project purposes. The data was scrapped from the internet especially from the Internet Movie Database (IMDB).

* + 1. **Method of Collection:**

The tools used by NYC Data Science Academy for scraping all 5000+ movies is a Python library called "scrapy". Use scrapy in Python to obtain a list of 5043 movie titles of from "the-numbers" website. Save the titles into a JSON file. Search those titles from IMDB website to get the real IMDB movie links. Send HTTP request to each movie page using the links, and scrapy the page and get all data. Parse the aggregated data, clean it, and reformat it to CSV file (python code). The final CSV file can be found in kaggle.com website.

* + 1. **Input:**

The input to the program would be the dataset which consists of 5000 movie information and has attributes like director\_name, duration, director\_facebook\_likes, actors name, actors facebook likes, genres, gross collection, budget, movie\_title, num\_of\_voted\_users, cast\_facebook\_likes, title\_year and movie\_facebook\_likes. These data will be passed as an input to the algorithm and the data will be processed to obtain the result.

* + 1. **Report:**

The report would contain the results and also graphs that show the movies that would be successful after its release. We have also generated a pdf report of the results by using the R markdown package.

* 1. **System requirements:**

**3.4.1 Hardware requirements :**

1. Processor: Core 2 Duo

2. Speed: 2.3Ghz

3. Ram: 4GB

4.Storage: 10GB

5. Peripherals : Monitor, keyboard and Mouse.

* + 1. **Software Requirements :**
  1. Operating System: Windows 7
  2. IDE: R Studio
  3. Other software : Microsoft Excel, Word

**3.5 System installation and Implementation**

This system can be installed or used by the film industries all around the world. It requires very minimal specification system and also the skill set to use and understand the result. This would help the pre-production team to analyse and understand the popularity of the movie and the cast before it is released in the cinemas. The results could help the team to decide how much prints are required to release and the cost of advertising the movie could be managed. In future this system can be used to predict which type of movie or genre would be popular and brings the profit to the production team.

**3.6 Tools for Testing and Validation of Mining Models**

Analysis Services supports multiple approaches to validation of data mining solutions, supporting all phases of the data mining test methodology.

• Partitioning data into testing and training sets.

• Filtering models to train and test different combinations of the same source data.

• Measuring lift and gain. A lift chart is a method of visualizing the improvement that you get from using a data mining model, when you compare it to random guessing.

• Performing cross-validation of data sets

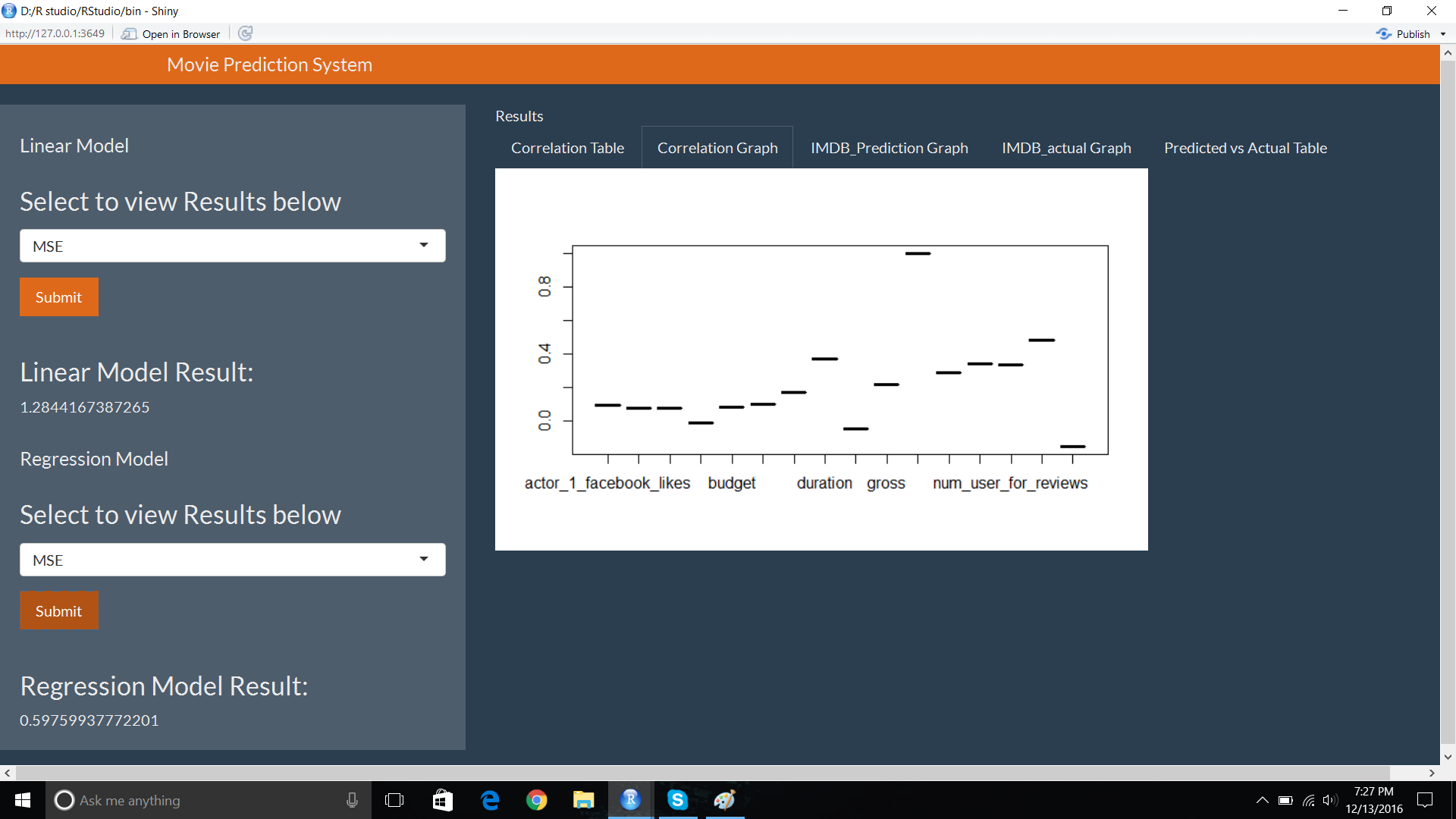
• Generating classification matrices. These charts sort good and bad guesses into a table so that you can quickly and easily gauge how accurately the model predicts the target value.

• Creating scatter plots to assess the fit of a regression formula.

• Creating profit charts that associate financial gain or costs with the use of a mining model, so that you can assess the value of the recommendations.

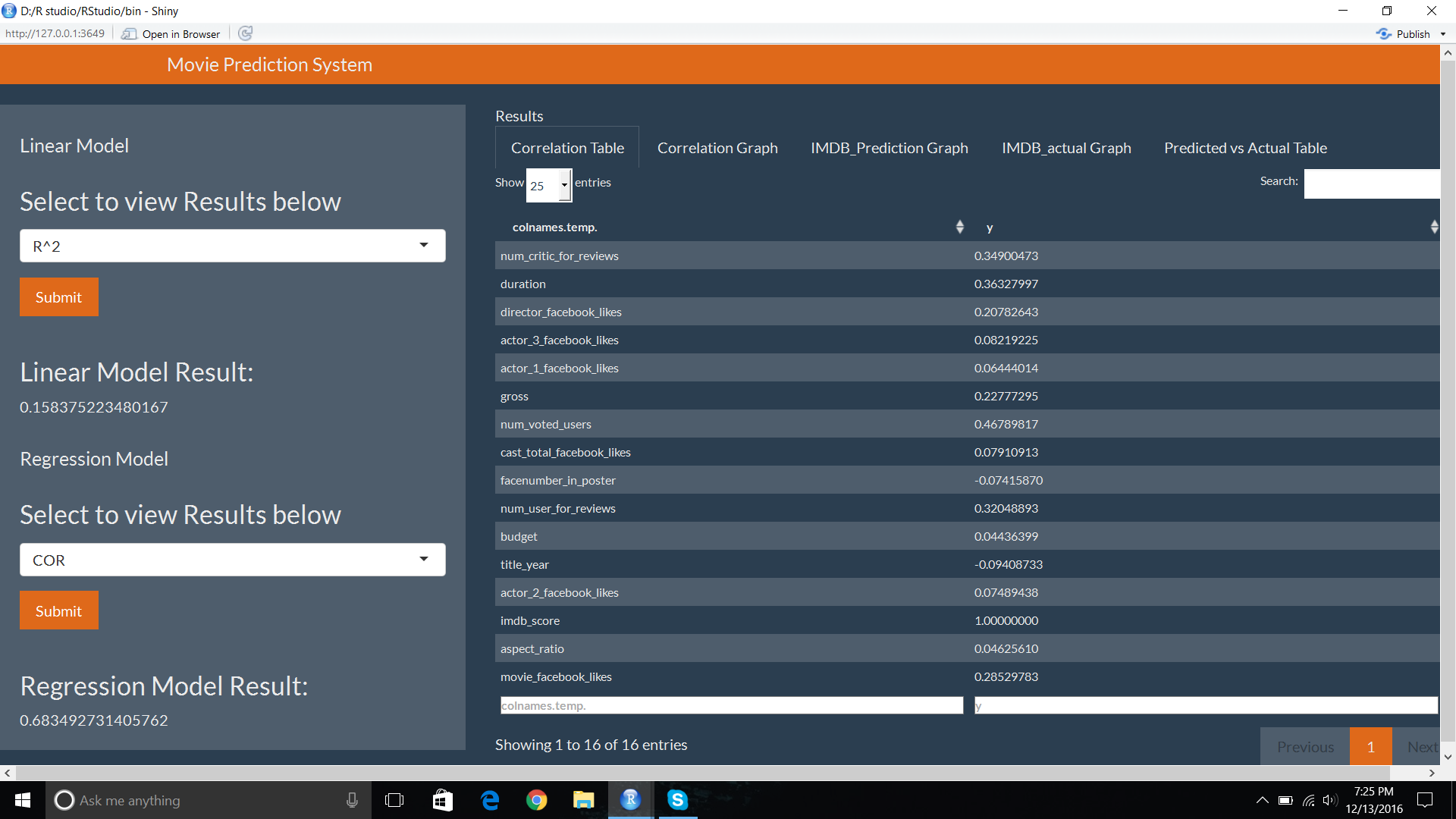
* 1. **Results and Discussions**

**1. Correlation Graph**



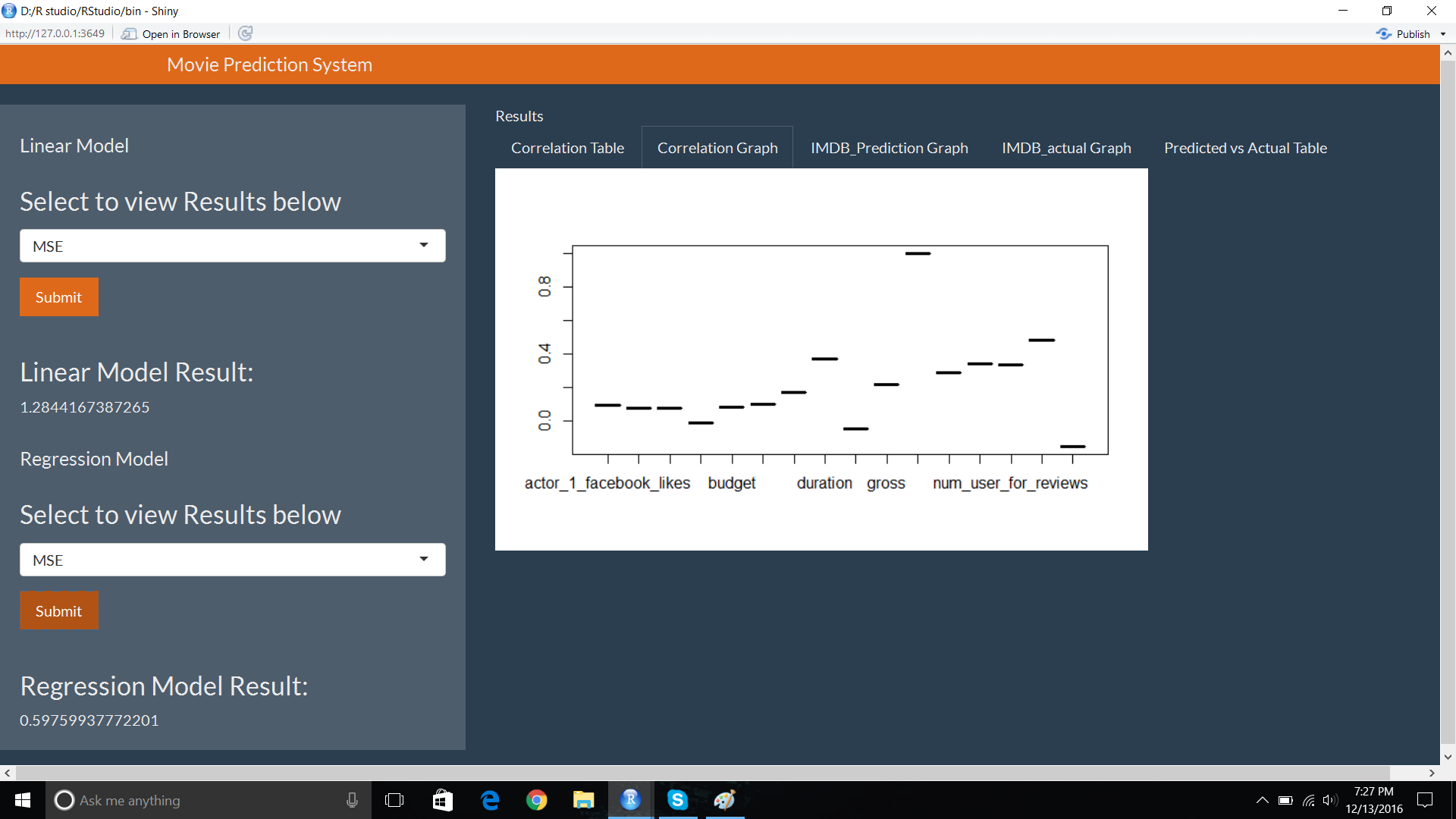
As you can see here, the graph on UI shows that attributes are mentioned on X axis and correlation values are on Y axis. The graph shows that the highest correlation value is of imdb\_score attribute and second highest correlation value is of num\_voted\_users attribute. It automatically satisfied the purpose of performing the correlation as these two attributes we should use for further prediction analysis.

**2. R\_squared for Linear Regression Model**



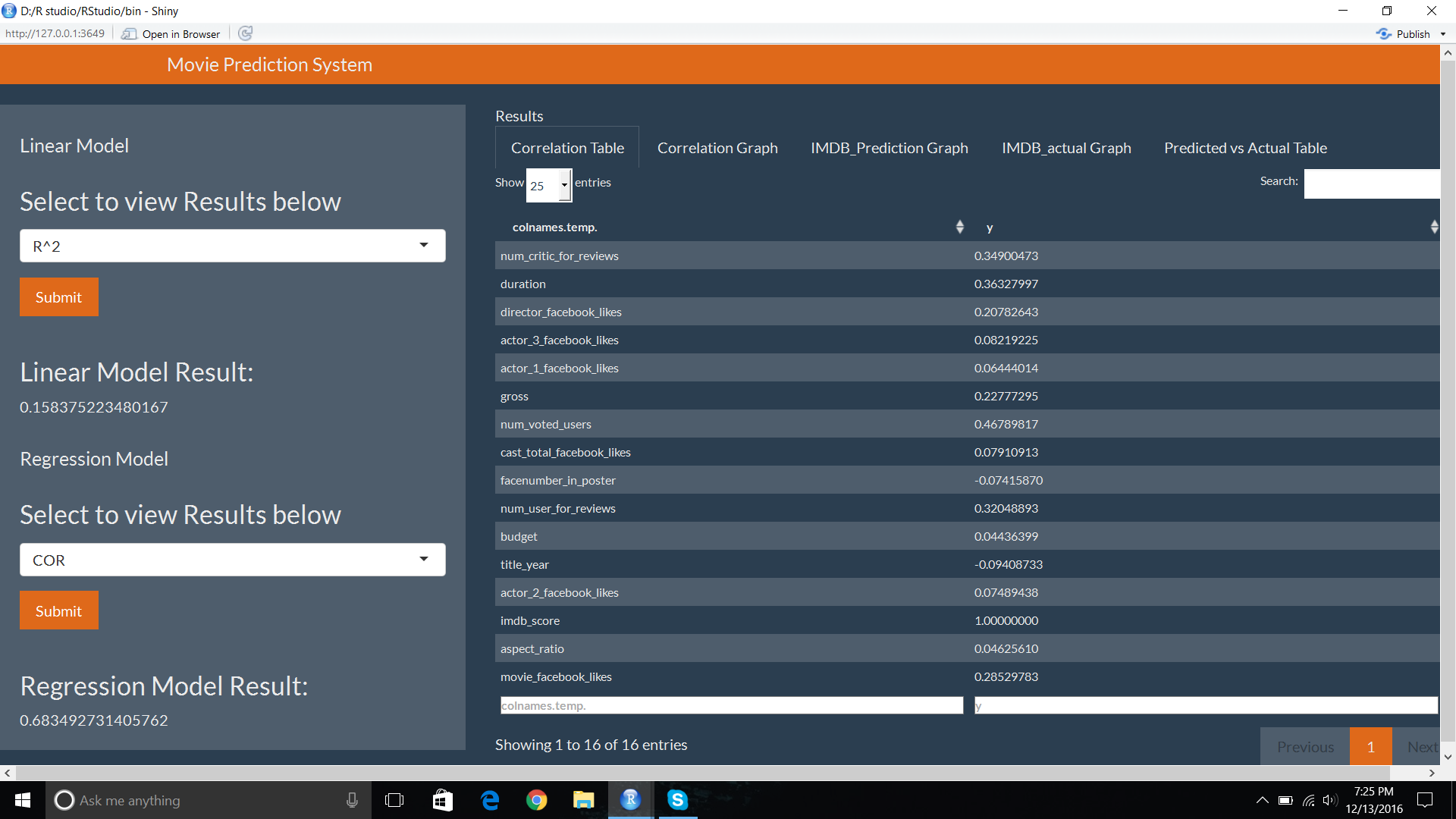
The value of R\_squared calculation for linear regression model is 0.15 for used IMDb data set. It is nearer to 0 which is exactly opposite of what is required.

**3. MSE for Linear Regression Model**



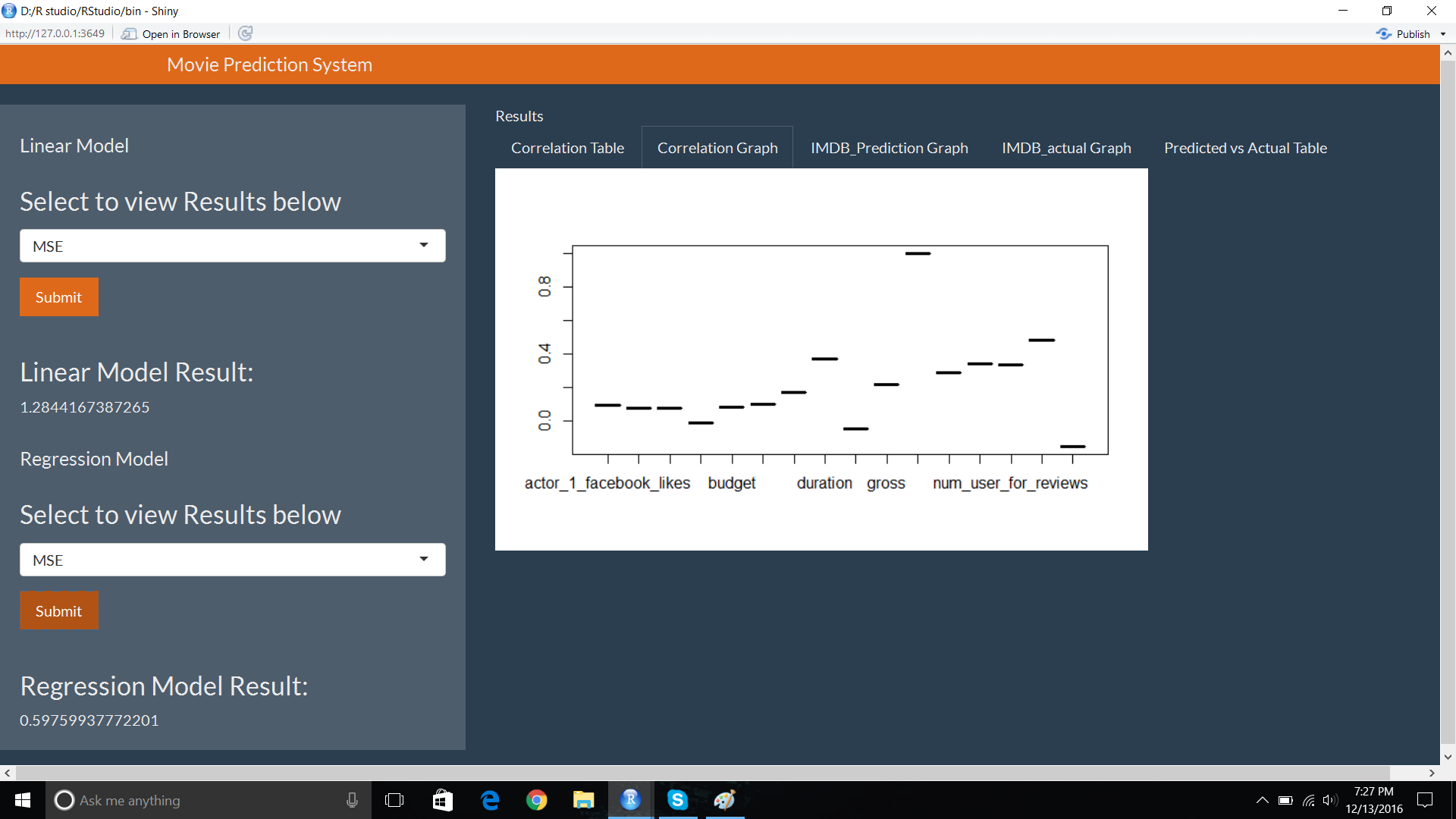
The value of MSE calculation for linear regression model is 1.28 for used IMDb data set. It is nearer to 1 i.e. opposite to what is required. So, we can conclude from 2 calculations for linear model that this linear regression model is not suitable for the IMDb data sets we are using.

**4. COR for Regression Tree Model**



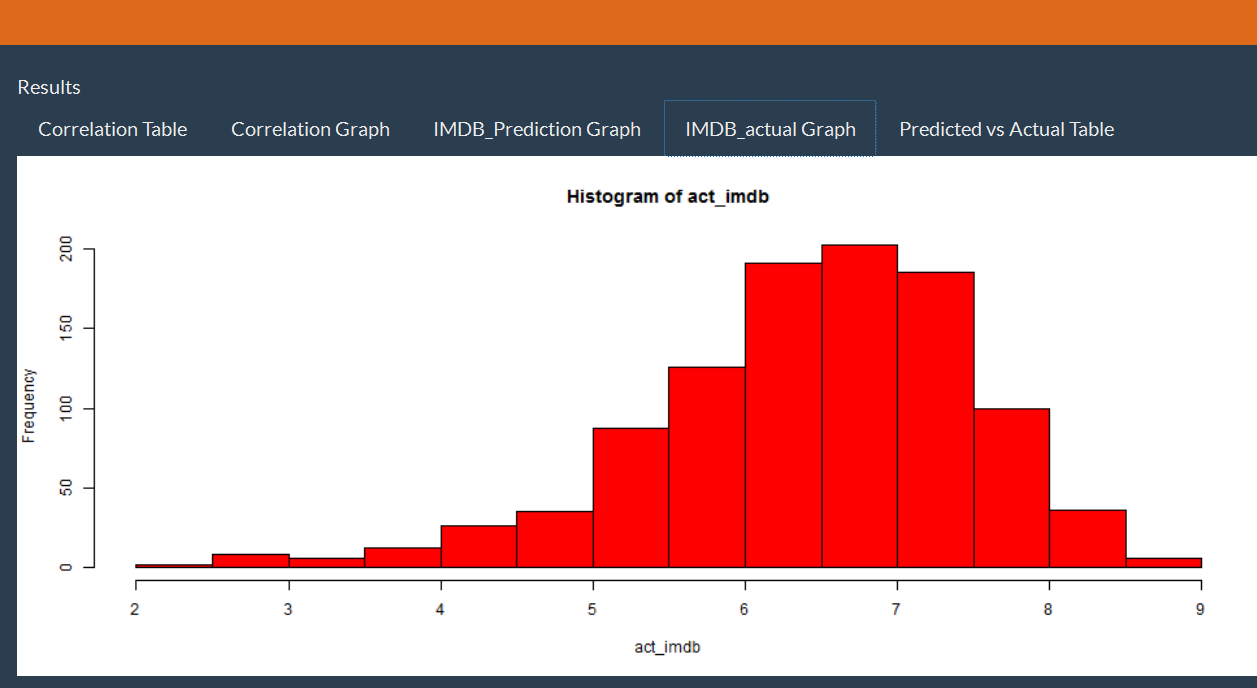
The COR value for regression tree model is 0.68 which is nearer 0 as required.

**5. MSE for Regression Tree Model**



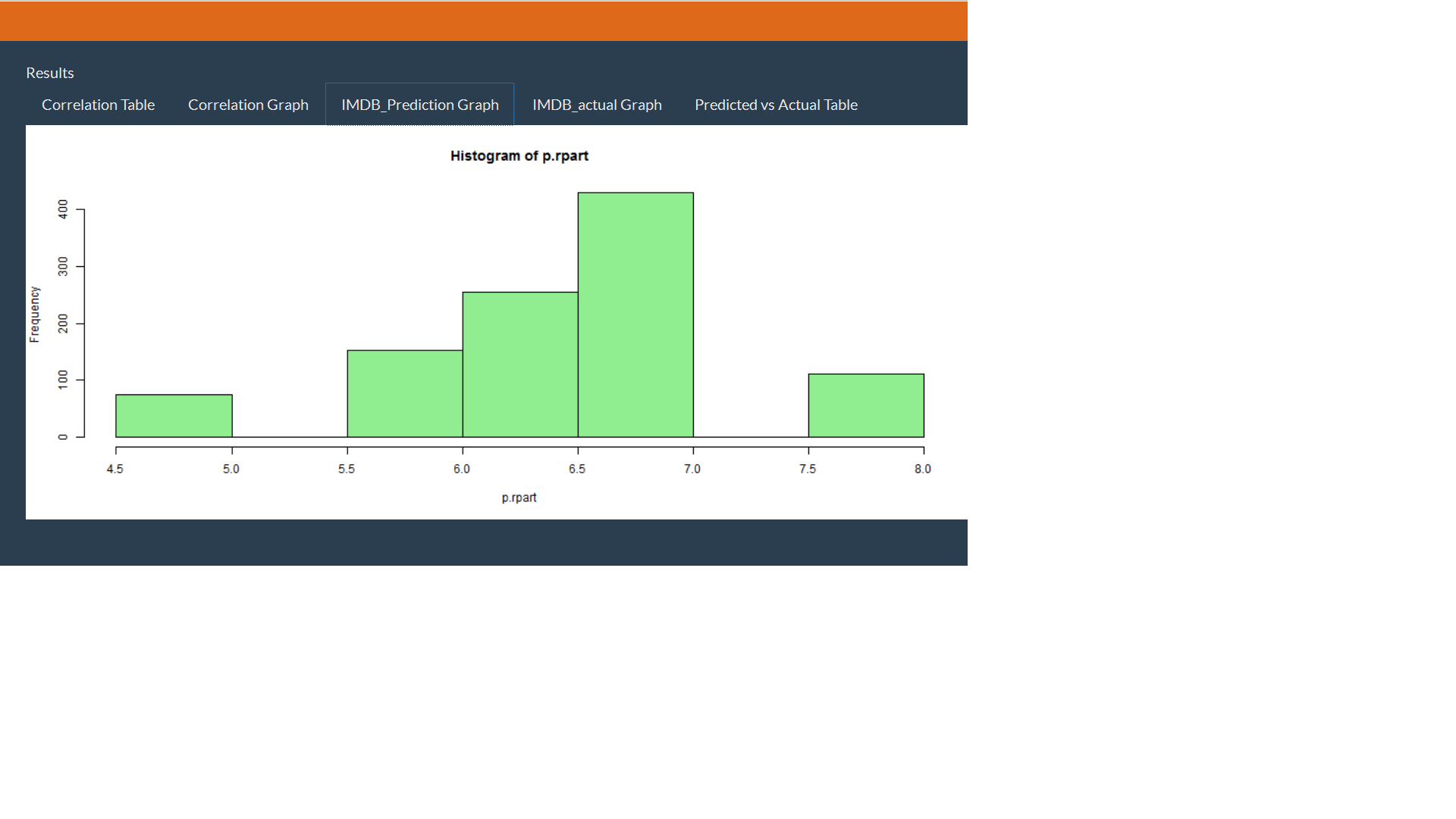
The value of MSE calculation for regression tree model is 0.59 which is nearer to 1 as required. In conclusion, regression tree model is suitable for the IMDb data set we are using for this project.

**6. Actual Imdb\_Score Graph**



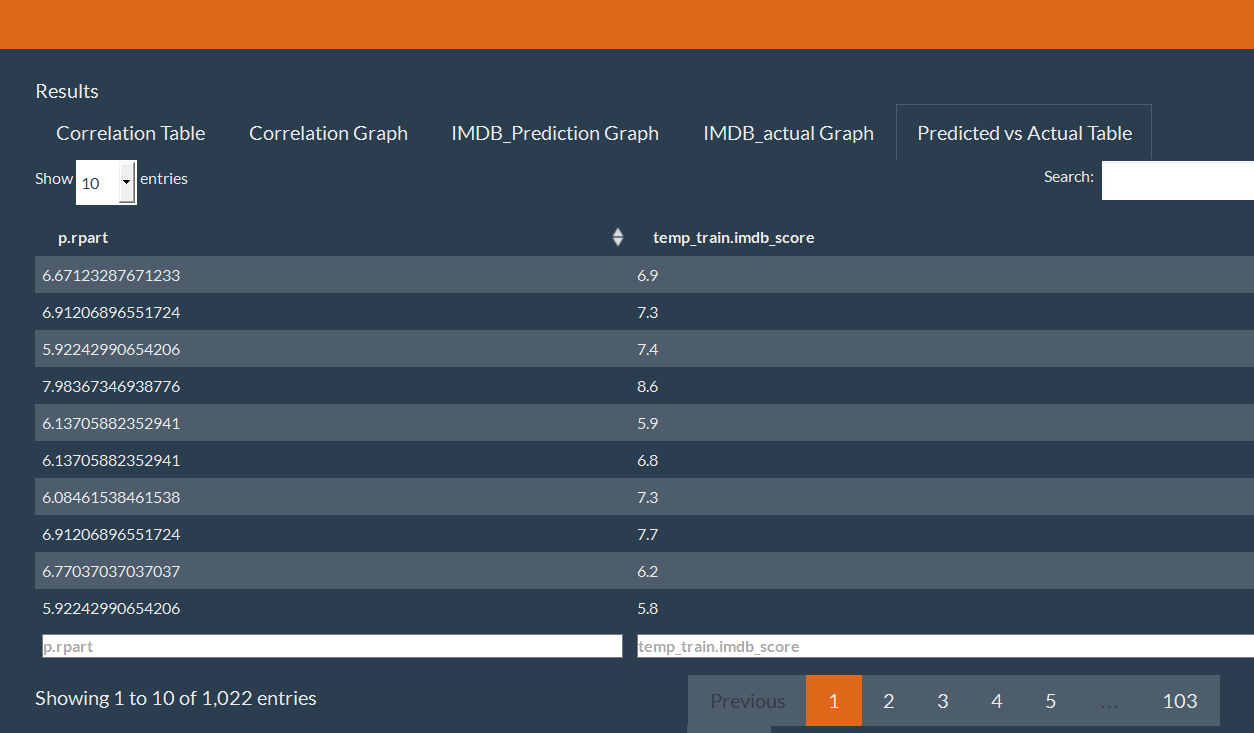
Above diagram shows actual IMDb\_scores values present into data set.

**7. Predicted Imdb\_Score Graph**



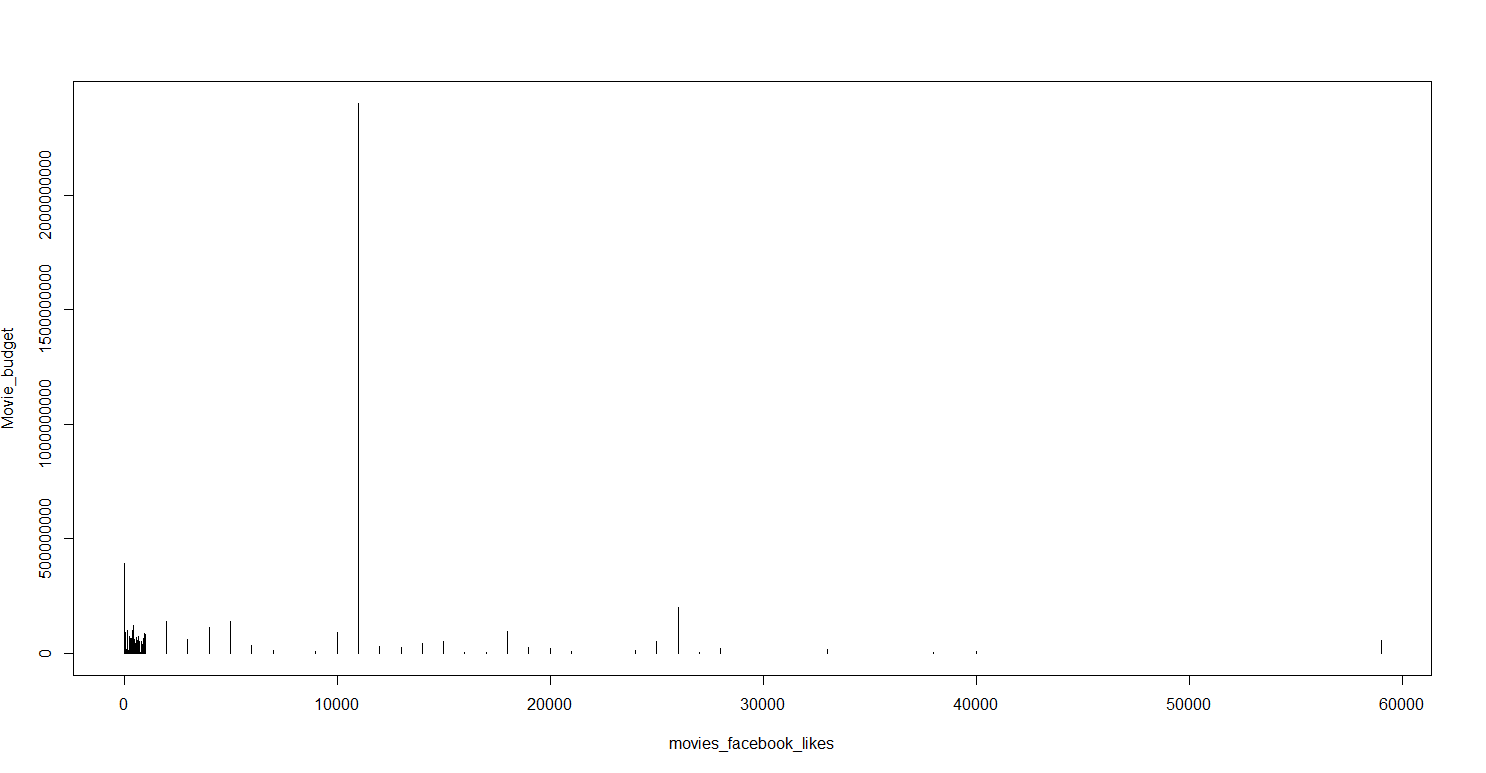
This figure shows plotted predicted IMDb\_score values.

**8. Comparison of Actual and Predicted IMDb Score**



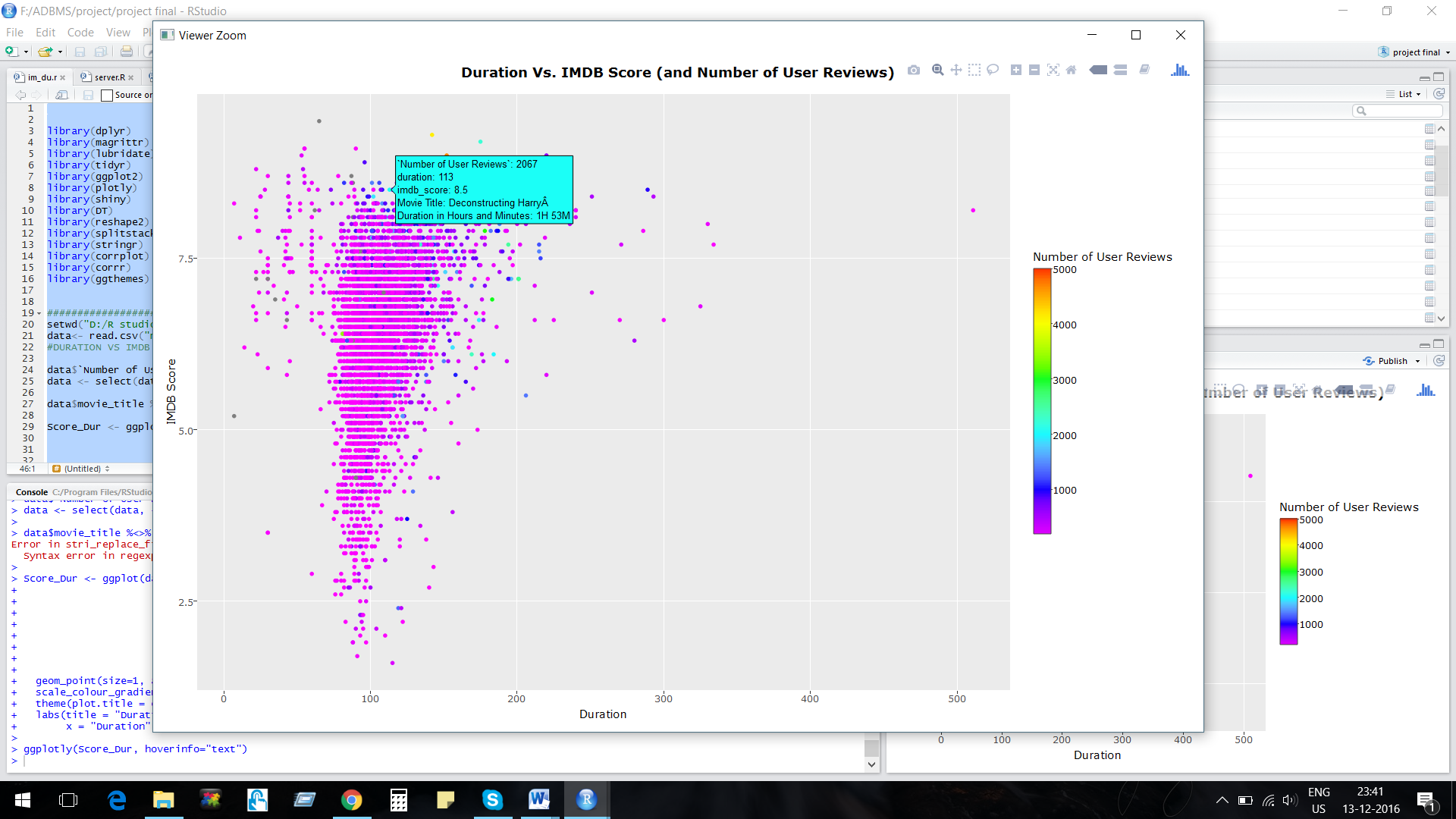
Above figure shows both values of actual and predicted IMDb\_score. By comparing these values, it can be easily understand that predicted result is almost similar to actual result. Hence, It can be proved that the prediction is correct.

**9. Budget and Movie\_Facebook\_Likes Graph**



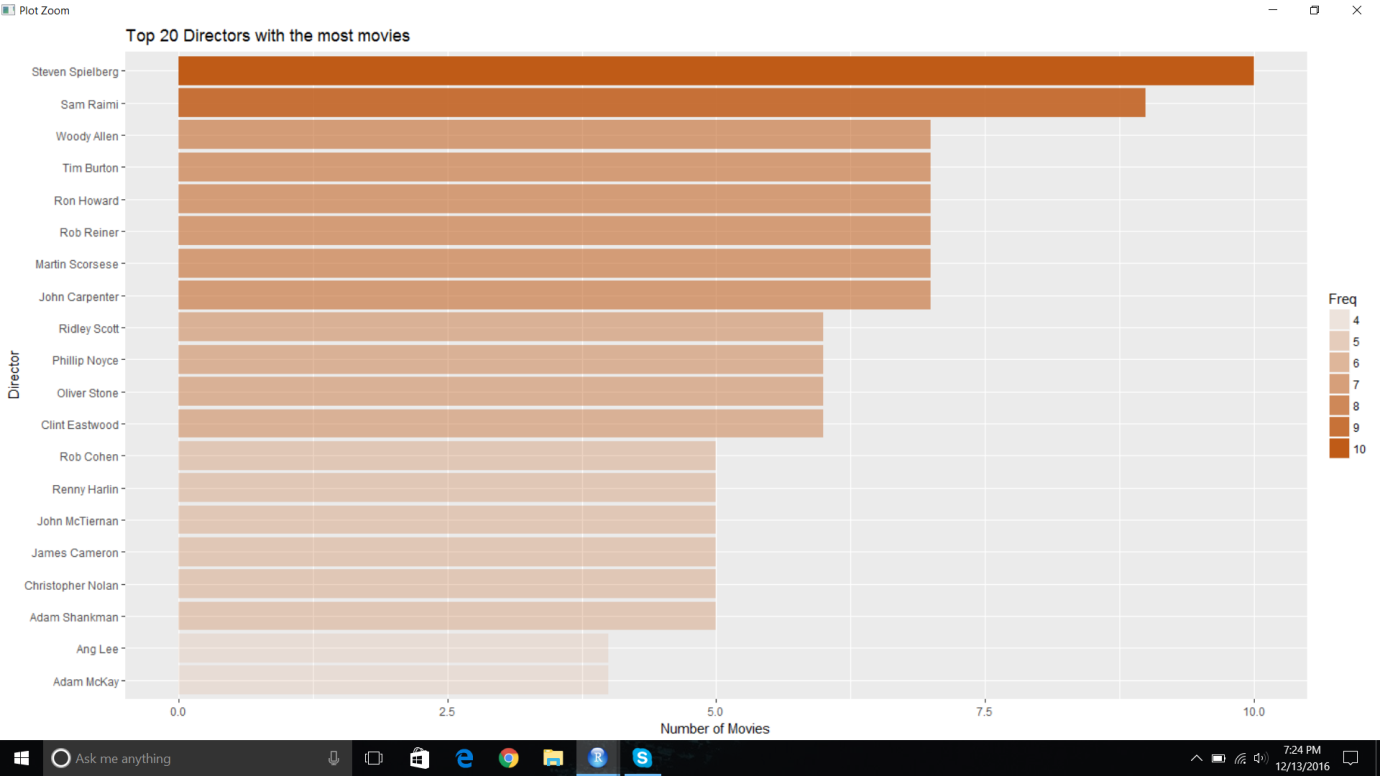
It is a statistical analysis where two variables movie\_facebook\_likes and Movie\_budget are used. Here we have to check that whether movie\_facebook\_likes are getting affected by budget or not. However, this graph states that these 2 variables are independent of each other, even though the budget of one movie is high but the movie\_facebook\_likes are low.

**10. Duration and Imdb\_Score Graph**



This figure shows that the duration and IMDb\_score values are dependent on each other. Here, IMDb\_score is high of movies whose duration is between 90 to 120 min. In conclusion, we can say that movies which have duration between 90 to 120min can be more successful than others.

**11. Top 20 Director Graph**



This figure shows name of directors on one side and number of movies on other side. This also shows top 20 directors who created lot of movies. For example, Steven Spielberg is the top director whose movie ratio is highest.

* 1. **Conclusions , Implications and Recommendation**

**5.1 Conclusion**

A project conducted a practical implementation of data mining in R language and used RStudio as environment and provided realistic implementation of linear and regression tree algorithm. We have taken the IMDb data sets from kaggle website. We divided this data set into train data set and test data set. We performed prediction for first 100 records and result of this prediction turned out to be almost accurate. After implementation of two algorithms and predicted results, we came to know its advantages as we can easily predict the success of movie before its release by using regression tree algorithm in data mining. Statistical analysis of current released movie data set also helped to make predictions for success for upcoming movies.

**5.2 Implications**

The impact of this project gives the practical experience of a website which shows the rating of current released movie according to audience reviews as well as predicts the success of upcoming movie according to its casts, production team ratings. Mainly this prediction helps production team to make analysis for their investment in movies. According to the current statistical graph and predicted result, production team can make decision about in which movie they should invest their money to make good profit. The future scope of this project could be use of machine learning for more predictions.

**5.3 Recommendations**

This project is an implementation of data mining using R language. We have used RStudio tool in which we have used shiny app for implementation of UI. The future researcher can make it more attractive and useful by using different tools like Weka etc. as well as they can make use of html, CSS and other scripting languages for making more attractive UI. Such UI requires the function to call the predicted results from used tool. These results should be shown in the form of graphs to the user. UI can also become more user friendly for audience if we give the option of selection of particular movie from the list.

# References

Ajay Siva Santosh Reddy, P. K. (2012). Box-Office Opening Prediction of Movies based on Hype. *International Journal of Computer Applications*.

Haowen Cao, D. H. (2014). Building a Predictor for Movie Ratings. *Journal of Standford Education*.

Krauss, J., Nann, S., Simon, D., & Fischbach, K. (2013). PREDICTING MOVIE SUCCESS AND ACADEMY AWARDS. *Oscar\_ECIS\_Final*.

M. Saraee, S. W. (2004). A Data Mining Approach To Analysis And Prediction Of Movie Ratings. *WIT Press*.

NevonProjects. (2015). *Movie Success Prediction Using Data Mining.* Retrieved from nevon projects: http://nevonprojects.com/movie-success-prediction-using-data-mining/

Persson, K. (1977). PREDICTING MOVIE RATINGS. *University of Skovde*.

S. Kabinsingha, S. C. (2012). A Movie Rating Approach and Application. *International Journal of Engineering and Innovative Technology* .