

**School of Business**

**OPIM 5604 – Predictive Modeling Project**

**MOVIE RATING MODEL**

17th April, 2018

**Team Members:**

Amit Gangrade

Akshata Nayak

Gargi Yadav

Gaurav Sharma

Juhi Maheshwari

Shobana Ravichandran

**Table of Contents**

Executive Summary

1 Introduction

1.1 Problem Statement

1.2 Data Description

2 Methodology

2.1 Data Pre-processing

2.1.1 Resolve Inconsistencies

2.1.2 Missing Value

2.1.3 Outlier Analysis

2.1.4 Variable Transformations

2.1.5 Descriptive Statistics & Visualizations

2.2 Data Modeling Process for Objective 1

2.2.1 Logistic Regression

2.2.2 Neural Networks

2.2.3 Discriminant Analysis

2.3 Data Modeling Process for Objective 2

2.2.1 Linear Regression

2.2.2 Neural Networks

2.2.3 Boosted Trees

2.3.4 K Nearest Neighbor

3 Conclusion & Recommendations

4 References

5 Appendix

**Executive Summary**

This paper describes detailed analysis on Internet Movie Database (IMDb) dataset, a free, user-maintained, online resource of production details for over 5043 movies. The movies range from the year 1916 to 2016, from 59 countries across the globe. The various data cleaning and reduction techniques include handling missing values, outlier analysis, data transformation and creating derived variables. Some interesting facts and correlations have been gathered using various predictive

modeling techniques in JMP. In particular, attributes such as budget of a movie, actor’s popularity on social media, movie reviews and genres play a vital role in the success of a movie. The analysis

helps to answer key questions such as, if the budget of a movie, famous actor face on a movie poster, number of reviews received from critics, duration of a movie or director’s popularity on Facebook impacts the movie success. Also, a predictive model has been built to predict the IMDB score equivalent for rating the movies, ranging on a scale of 1 to 10.

The objective has been achieved by building prediction models by using techniques such as Linear regression, Bootstrap Tree, Logistic Regression, Neural Network and KNN to predict a movie as a hit or a flop and the IMDB score of a movie based on various significant parameters. Among all the models that were tested on the movie dataset, Logistic regression model predicted most accurate results for categorical target variable hit/flop on the basis of confusion matrix -model accuracy of 75.45%. For continuous target variable, IMDB rating, Neural model predicts most accurate results based on the most significant RSquare value of 49.5% and RMSE of 76.28%.

1. **Introduction**

The Internet Movie Database (IMDb) is the world's most popular and authoritative source for any movie. It provides valuable information about general trends in films. Through this project the SEMMA approach to Data Mining techniques have been applied to extract interesting patterns or knowledge from the IMDB movie data. It will reveal information, also confirm or disprove the assumptions about the movies, and allow us to predict the success of a future film given selected information about the film before its release.

**1.1 Problem Statement**

The average number of movies released per year are rapidly increasing. Predicting the success of a movie is a huge concern for directors and producers who are looking for financial returns and credibility. For many years people relied on critics to gauge the quality of the movie which is a time taking process and also there are chances of bias involved. So, devising a method which can predict not only the success of the movie but also help understand what factors contribute to the success of the movie could really benefit the film industry. This will enable them to focus on their advertising campaigns, also help them to find the most appropriate time to release a movie by looking at multiple factors. Through this project, we want to leverage predictive analytics to enhance the box office success rate for the movie.

**1.2 Data Description**

The dataset is taken from “data.world” website. It contains 28 variables for 5043 movies between years 1916 to 2016. IMDB\_Score is the target variable while rest of the 27 variables are possible predictors. The link for the dataset from Data world is mentioned below:

<https://data.world/popculture/imdb-5000-movie-dataset>

Following is the description of the variables of the dataset:

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| movie\_title | Title of the Movie |
| Duration | Duration in minutes |
| director\_name | Name of the Director of the Movie |
| director\_facebook\_likes | Number of likes of the Director on his Facebook Page |
| actor\_1\_name | Primary actor starring in the movie |
| actor\_1\_facebook\_likes | Number of likes of the Actor\_1 on his/her Facebook Page |
| actor\_2\_name | Other actor starring in the movie |
| actor\_2\_facebook\_likes | Number of likes of the Actor\_2 on his/her Facebook Page |
| actor\_3\_name | Other actor starring in the movie |
| actor\_3\_facebook\_likes | Number of likes of the Actor\_3 on his/her Facebook Page |
| num\_user\_for\_reviews | Number of users who gave a review |
| num\_critic\_for\_reviews | Number of critical reviews on imdb |
| num\_voted\_users | Number of people who voted for the movie |
| cast\_total\_facebook\_likes | Total number of facebook likes of the entire cast of the movie |
| movie\_facebook\_likes | Number of Facebook likes in the movie page |
| plot\_keywords | Keywords describing the movie plot |
| facenumber\_in\_poster | Number of the actor who featured in the movie poster |
| Color | Film colorization. ‘Black and White’ or ‘Color’ |
| Genres | Film categorization like ‘Animation’, ‘Comedy’, ‘Romance’, ‘Horror’, ‘Sci-Fi’, ‘Action’, ‘Family’ |
| title\_year | The year in which the movie is released (1916:2016) |
| Language | English, Arabic, Chinese, French, German, Danish, Italian, Japanese etc |
| Country | Country where the movie is produced |
| content\_rating | Content rating of the movie |
| aspect\_ratio | Aspect ratio the movie was made in |
| movie\_imdb\_link | IMDB link of the movie |
| Gross | Gross earnings of the movie in Dollars |
| Budget | Budget of the movie in Dollars |
| imdb\_score | IMDB Score of the movie on IMDB |

**2. Methodology**

There are two types of prediction variables:

Categorical Prediction: Yes/No for identifying if the movie was a Hit or Flop

Numeric Prediction: Identifying the rating of the movie.

**2.1 Data Preprocessing**

The major tasks covered in Data Preprocessing were:

**Data Cleaning:** Following steps are involved in the data cleaning process.

**2.1.1 Resolve inconsistencies:**

It is observed in the data that there are 16 rows of data having greater than 8 missing columns. These rows are deleted and not used for analysis. The reason to eliminate these rows was to minimize any adverse impact on the analysis due to large number of missing values in the same row.

**2.1.2 Missing Value:**

|  |  |
| --- | --- |
| **Column name** | **Missing value action** |
| Color | 19 row had missing value which got imputed to color if yr >1940 |
| duration | 20 rows imputed manually |
| director\_name | 96 rows had missing value. Deleted the rows |
| actor\_3\_name | 9 rows had missing value. Deleted the rows |
| content\_rating | 186 rows had missing value which got imputed with rating as Unrated |
| Budget | Budget is imputed with Median 20000000. Since existing outlier increases the mean value, so it should not be imputed with the mean value. |
| Gross | Gross is imputed with the mean value 40718315, since the data  is widespread |

**2.1.3 Outlier Analysis:** Based on the distribution of all variables following Outliers were identified and excluded.

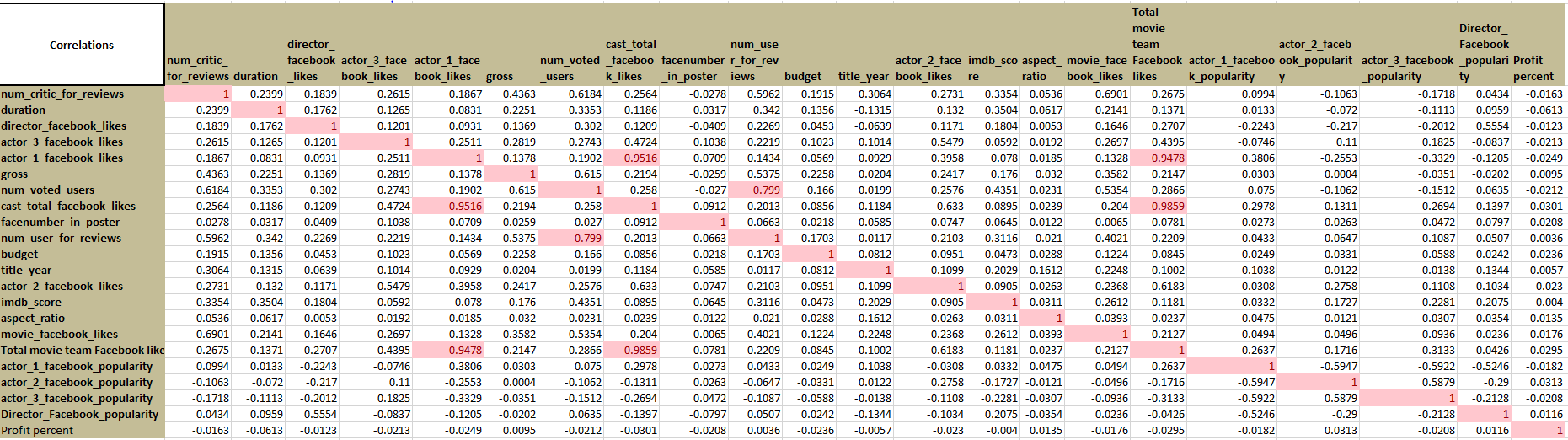
|  |  |
| --- | --- |
| **Column name** | **Handling Outliers** |
| actor\_1\_facebook\_like | Excluded two values of outliers(650000 , 260000). These date seems erroneous based on actual facebook likes |
| Gross | The Outlier here is Avatar movie and its Gross income is correct. Exclude this row |
| Budget | The outlier here is 4200000000 for the movie Lady Vengeance. Excluded this row |

**2.1.4 Variable Transformations:**

|  |  |
| --- | --- |
| **Column name** | **Column value description** |
| Content\_rating 2 | This is a new Column variable created whose value is segreagated into four categories. The column value is created based on the combination of original content\_rating and countries. |
| Total movie team Facebook likes | This column is created by adding the facebook likes of the columns - actor\_1\_facebook\_likes,actor\_2\_facebook\_likes,actor\_3\_facebook\_likes, director\_facebook\_likes |
| Popularity % of actors and directors - actor\_1\_facebook\_popularity, actor\_2\_facebook\_popularity, actor\_3\_facebook\_popularity, Director\_Facebook\_popularity | The popularity percent is calculated by dividing the No of facebook likes of respective actor and director by the Total Facebook likes.  For eg :  actor\_1\_facebook\_popularity = actor\_1\_facebook\_likes /Total movie team Facebook likes |
| Profit percent | This Column is created based on budget and gross. Profit percent = (Gross-Budget)/Budget |
| Profit | If the Profit percent is positive the Profit value is set a '1'. If the profit percent is not positive , the Profit is set as '0'. |
| Hit/Flop | The Column has a value of '1' if Profit = 1 and imdb\_score > 6, else the column value is set as '0' |
| New\_genre\_catergory | The column contains various movie genres that have been grouped based on broader movie themes. The excel for transformation is embedded below. |

**2.1.5 Descriptive Statistics & Visualizations**

Correlation Analysis



From the correlation analysis on the predictor variables, it is observed that there is a high correlation between the following pairs of variables:

1. actor1\_facebook\_likes and cast\_total\_facebook\_likes

2. num\_voted\_users and num\_users\_for\_reviews

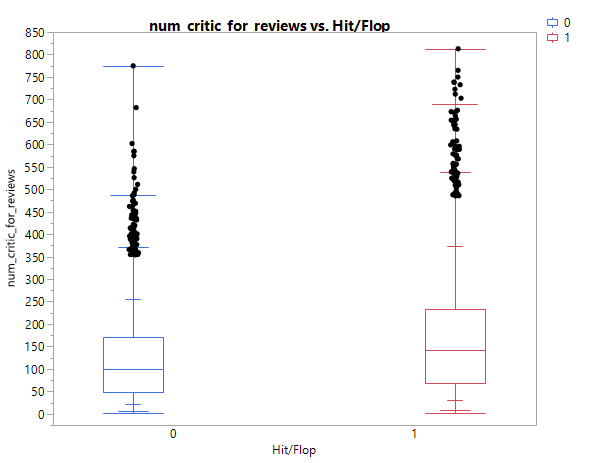
3. cast\_total\_facebook\_likes and Total\_movie\_team\_facebook\_likes

4. actor1\_facebook\_likes and Total\_movie\_team\_facebook\_likes

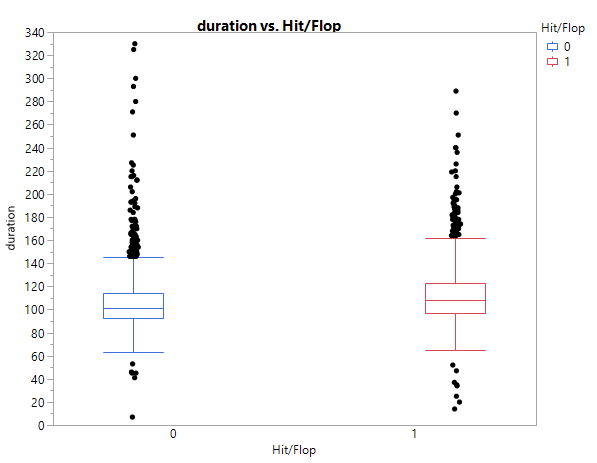
In order to proceed with further analysis involving these predictor variables, from each pair listed above, the variable with a higher correlation with the outcome variable (imdb\_score) is retained.

As a result, Total\_movie\_team\_facebook\_likes and num\_voted\_users are retained for further analysis from among all the variables listed above.

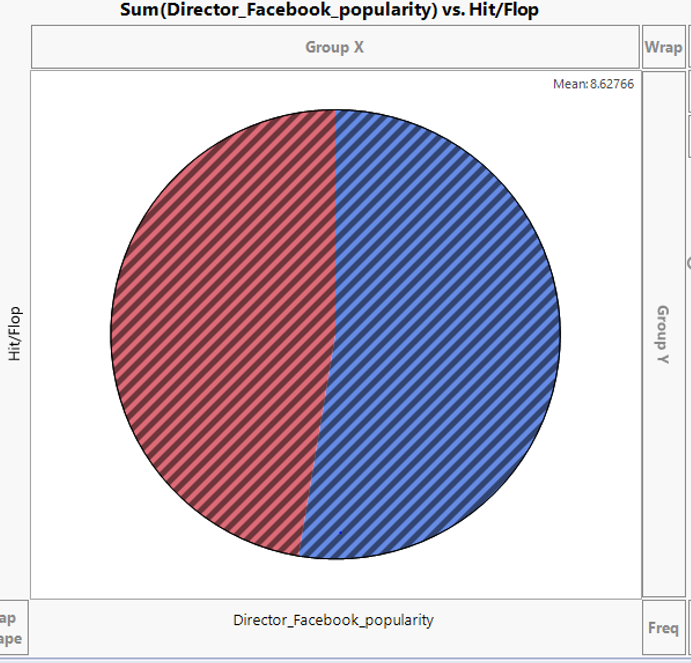
**Visualizations:** Below are the visualizations to understand the correlation and descriptive statistics.

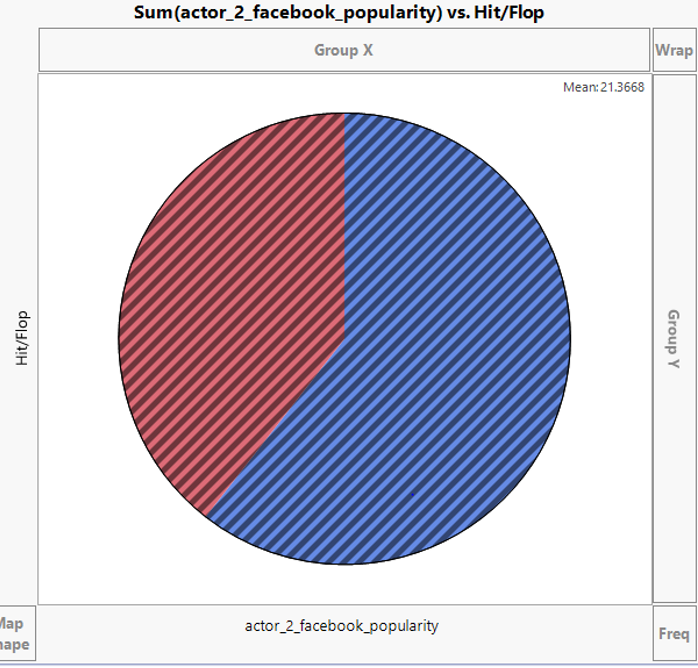
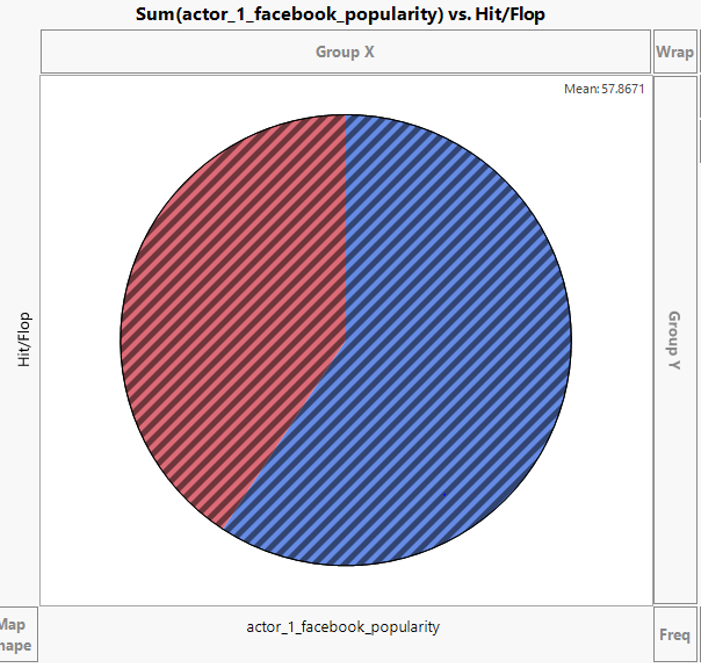


It is observed that the movie success is driven by higher number of critic reviews on an average.

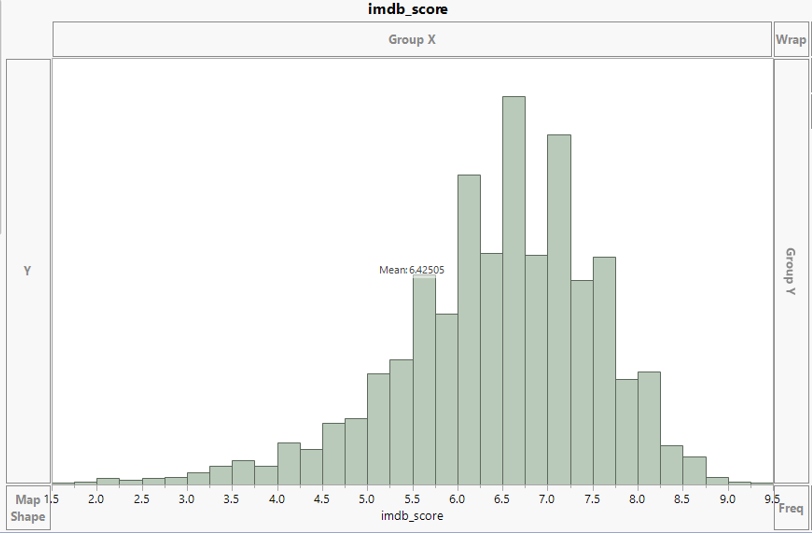


It can be observed that duration of the movie does not have a significant impact on the movie success.

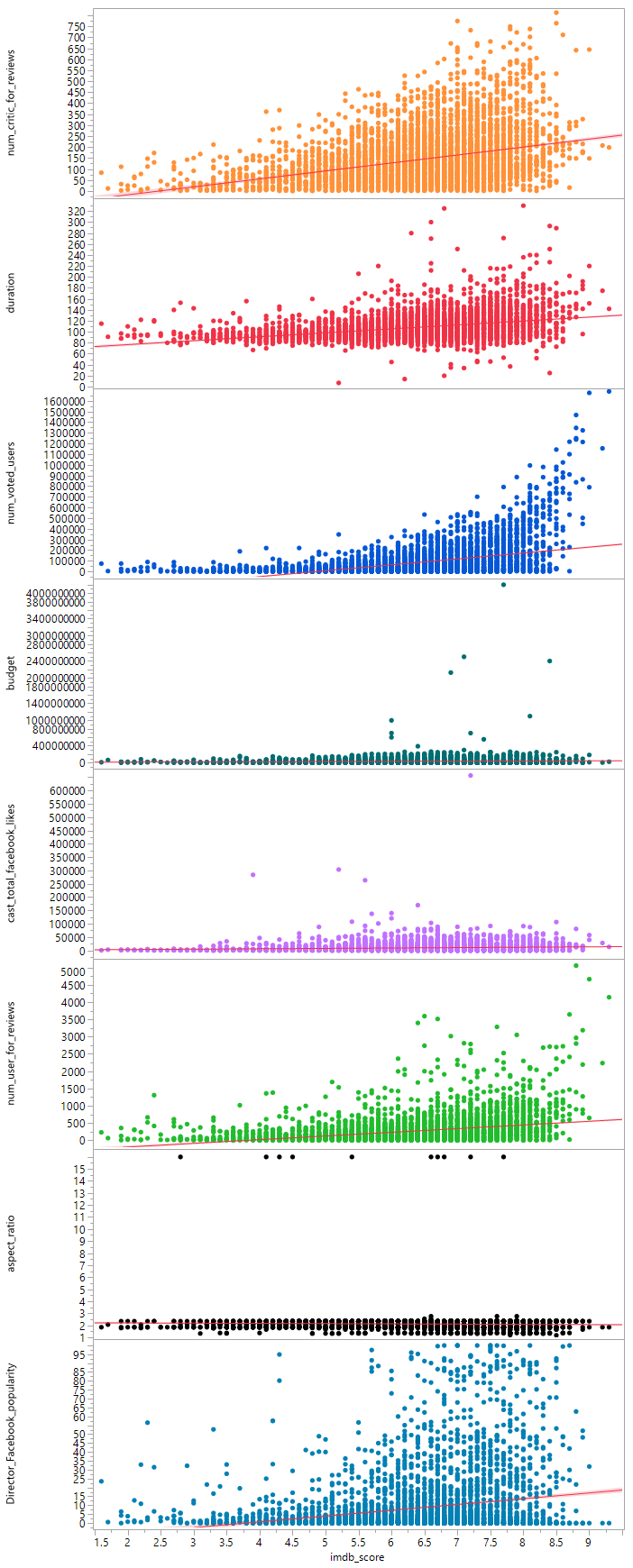




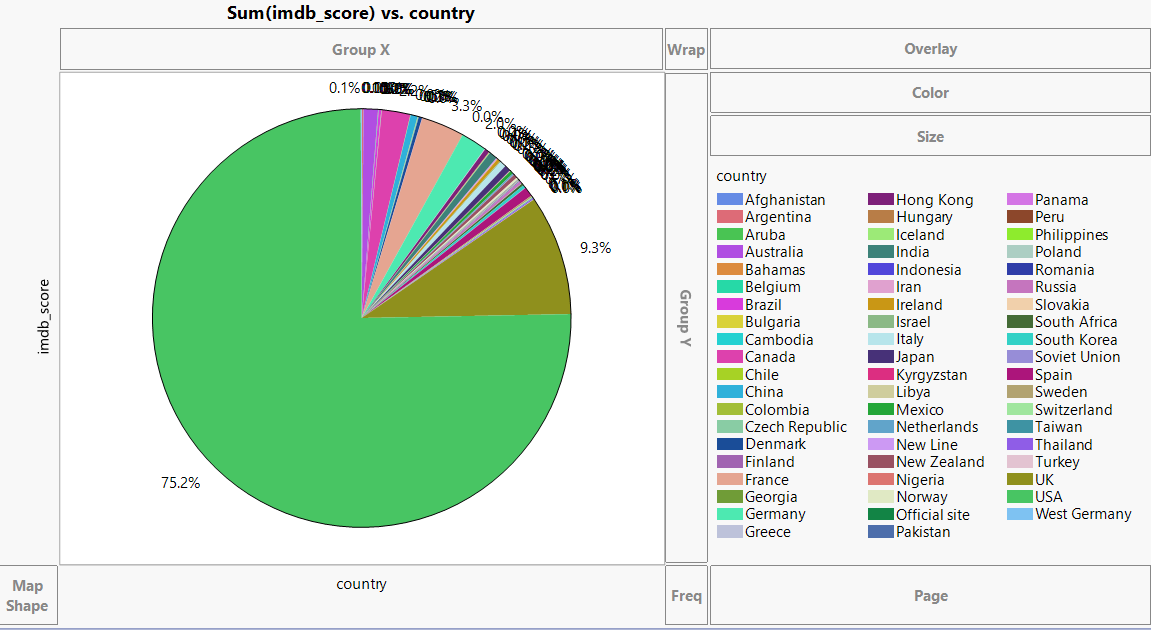
It can be observed that the actor 1 and actor 2 Facebook popularity drive movie success at a higher rate as compared to director Facebook popularity.



Above is the distribution of the IMBD score which states that most of the movie ratings lie between 5.5 to 8.



Above scatterplots display correlation analysis of IMDB score with various predictor variables.



Top 2 countries contributing to IMDB ratings are US and UK with 75.2% and 9.3% respectively.

Two objectives have been designed for exploration:

**2.2 Data Modeling Process for Objective 1:**

**Objective 1.** To be able to understand the factors influencing movie success (Hit or Flop).

In order to assess the objective defined, a new variable Hit/Flop has been created in the dataset.

Hit/Flop has been defined as Profit = 1 and imdb\_score > 6.

Below is the Year -wise Hit Rate of movies.

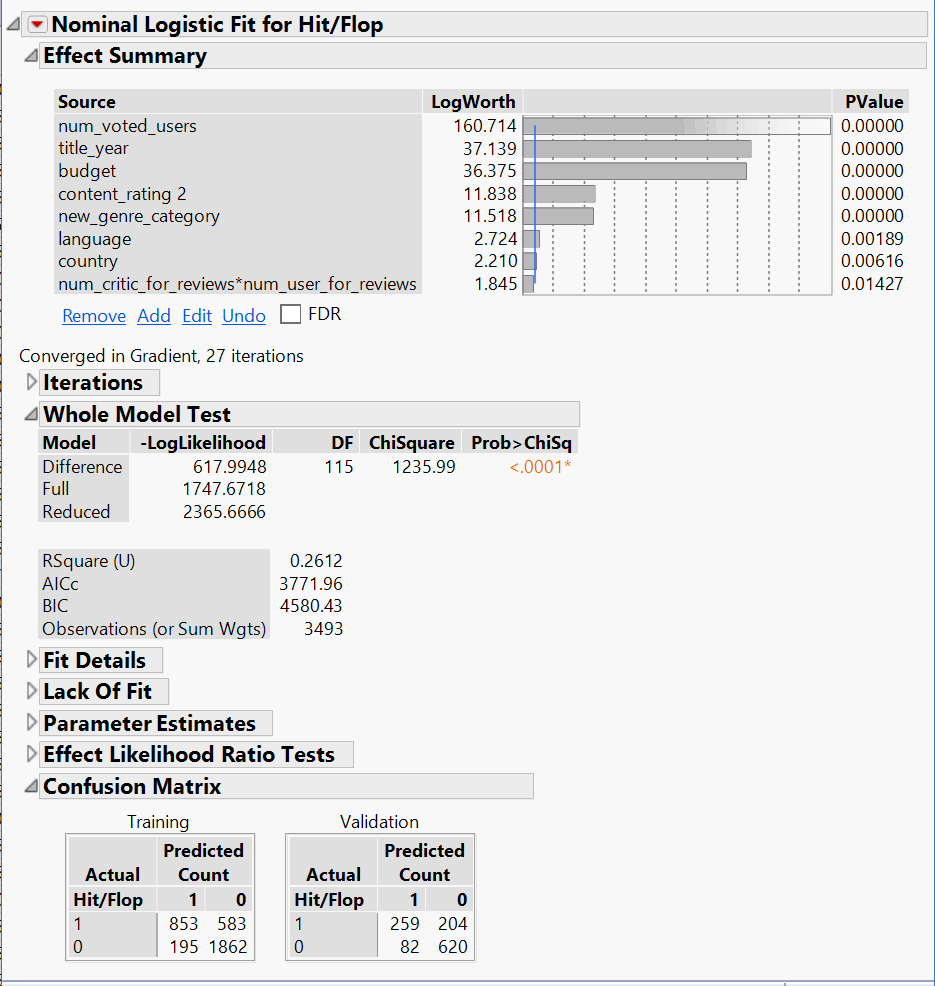
|  |  |  |  |
| --- | --- | --- | --- |
| **Year\_of\_Release** | **Sum of Flop** | **Sum of Hit** | **Hit\_%** |
| 1916-1939 | 4 | 17 | 80.95% |
| 1940-1950 | 2 | 23 | 92.00% |
| 1951-1960 | 1 | 28 | 96.55% |
| 1961-1970 | 13 | 67 | 83.75% |
| 1971-1980 | 23 | 96 | 80.67% |
| 1981-1990 | 102 | 180 | 63.83% |
| 1991-2000 | 537 | 383 | 41.63% |
| 2001-2010 | 1405 | 676 | 32.48% |
| 2011-2016 | 673 | 429 | 38.93% |
| **Grand Total** | **2760** | **1899** | **40.76%** |

In order to achieve the objective of defining movie success, 3 types of models have been built and consecutively the model with the highest accuracy was selected.

**2.2.1 Logistic Regression**

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

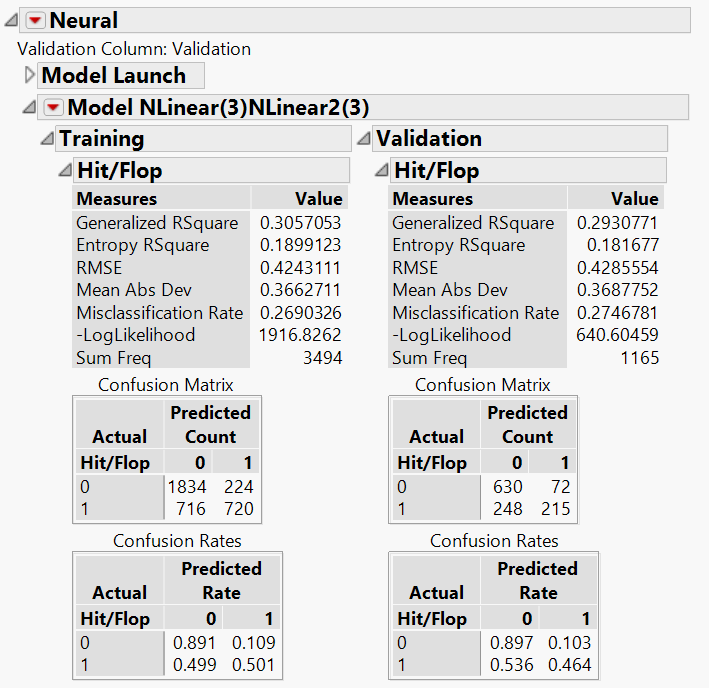
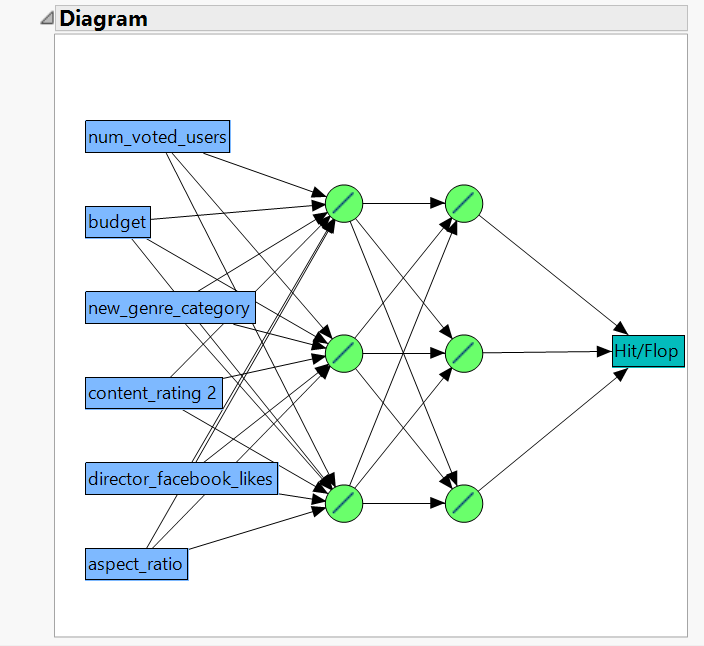
Based on several iterations we arrive at a final model in Logistic regression with below listed variables as highly significant.



Result: RSquare for this Model is 26% and Model Accuracy is 75.45%

Here, Model accuracy for Training data is 77.72 %., hence accuracies are consistent between Training and Validation datasets.

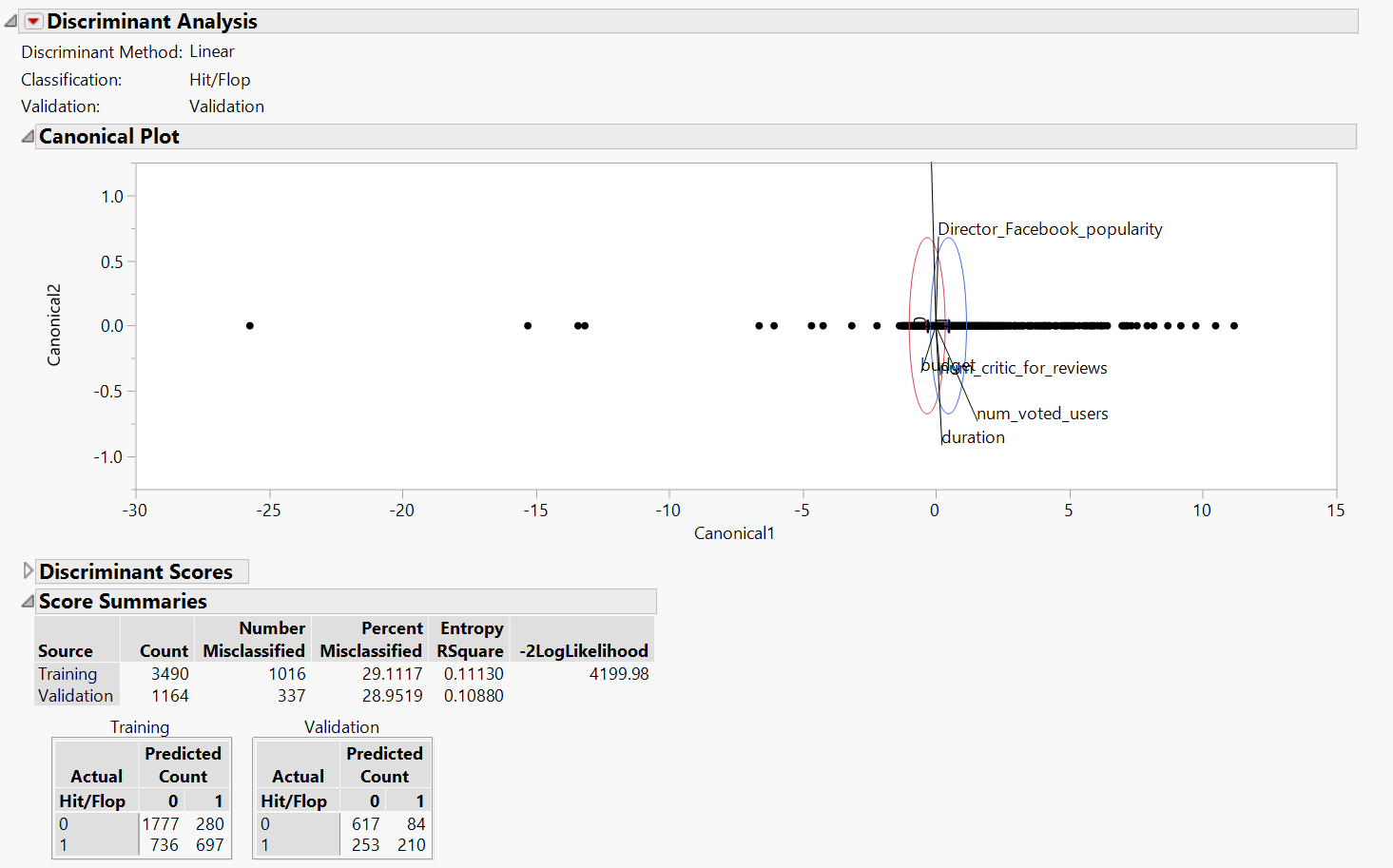
**2.2.2** **Neural Networks:** Neural networks are typically organized in layers. Layers are made up of a number of interconnected 'nodes' which contain an 'activation function'. Patterns are presented to the network via the 'input layer', which communicates to one or more 'hidden layers' where the actual processing is done via a system of weighted 'connections'. The hidden layers then link to an 'output layer' where the answer is output.



Result: RSquare of the Training and Validation Models are 30.57% and 29.30% respectively.

Model accuracy of Validation dataset is: 72.53% whereas Model accuracy of Training Dataset is 73.09%.

**2.2.3 Discriminant Analysis:** Discriminant function analysis is a statistical analysis to predict a categorical dependent variable (called a grouping variable) by one or more continuous or binary independent variables (called predictor variables).



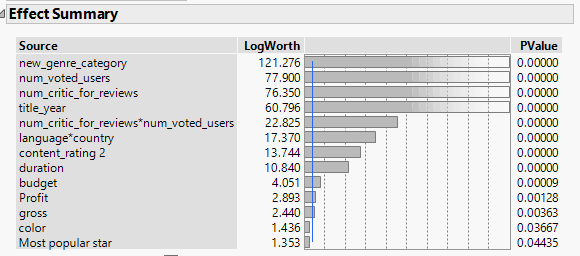
Result: Model accuracy of Validation dataset is : 71.04% whereas Model accuracy of Training Dataset is 70.89%.

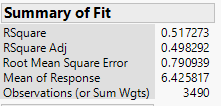
**Conclusion on Categorical Model:** Based on the above analysis it can be observed that Logistic Regression Model gives the highest accuracy in Training and Validation datasets . Hence the Logistic Regression Model with the Predictor variables : number of voted users, title year, budget, content rating, genre category, language, country, number of critic for reviews \* num\_user\_for\_reviews has been selected to predict Movie success – Hit/Flop.

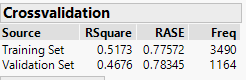
**2.3 Data Modeling Process for Objective 2:**

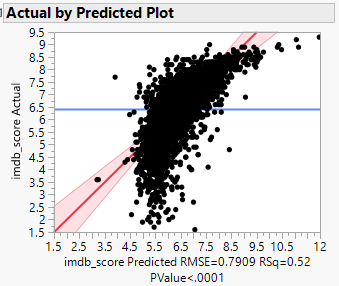
**Objective 2**. To understand the factors influencing the prediction of movie ratings (0 to 10).

**2.3.1 Linear Regression Model:** Linear regression is a linear approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.





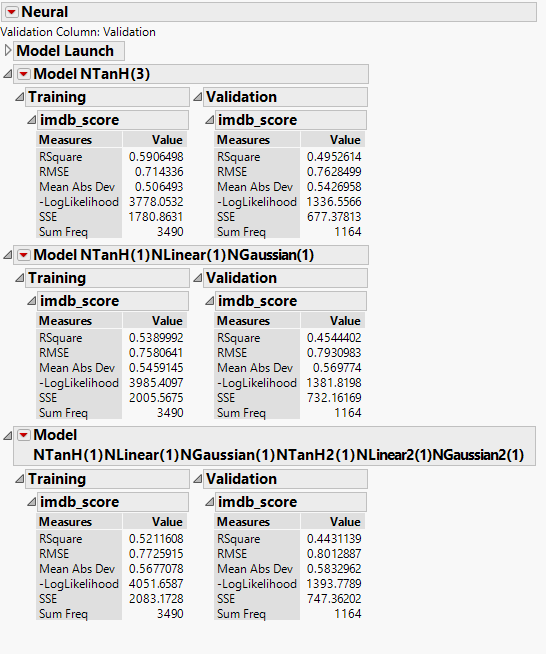


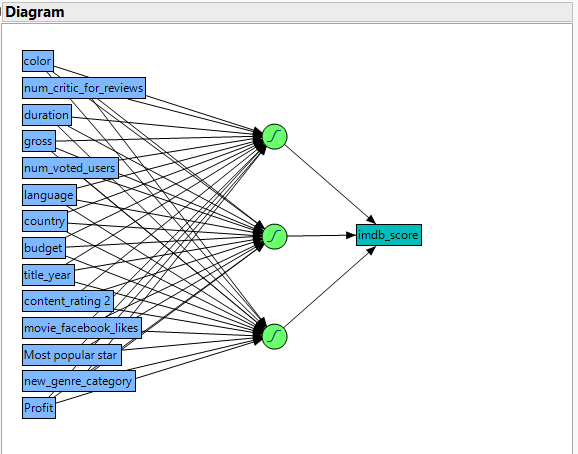


Results: RSquare of the Training Model is 51% and Validation Model is 47%.

RASE for the Model is 78.34 %

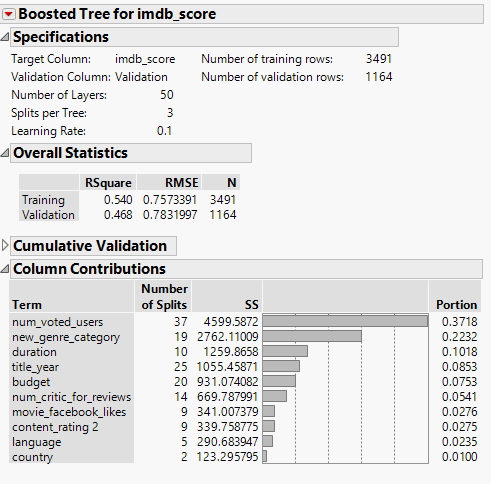
**2.3.2 Neural Networks:** Neural networks are typically organized in layers. Layers are made up of a number of interconnected 'nodes' which contain an 'activation function'. Patterns are presented to the network via the 'input layer', which communicates to one or more 'hidden layers' where the actual processing is done via a system of weighted 'connections'. The hidden layers then link to an 'output layer' where the answer is output.





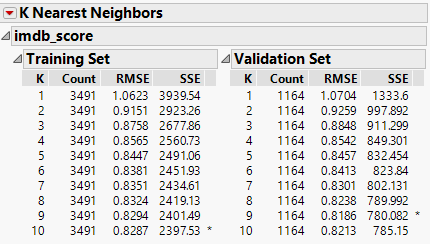
Results: RSquare of the best of the above mentioned three Training Models is 59.04% and Validation Model is 49.5%. RMSE for the Model is 76.28 %.

**2.3.3** **Boosted Trees :** The algorithm for Boosting Trees evolved from the application of boosting methods to regression trees. The general idea is to compute a sequence of (very) simple trees, where each successive tree is built for the prediction residuals of the preceding tree.



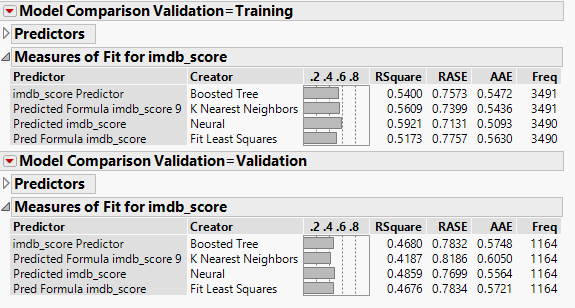
Results: RSquare of the Training Model is 54% and Validation Model is 46.8%. RMSE for the Model is 78.3%.

**2.3.4 K Nearest Neighbor (KNN)**: KNN classification is one of the most fundamental and simple classification methods and should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution of the data. K-nearest-neighbor classification was developed from the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine.



Results: RSquare of the Training Models is 56.09% and Validation Model is 41.87%. RASE for the Model is 81.86%.

Conclusion on Numeric data: Model Comparison table that has been created with the four models – Linear regression, Neural networks, Boosted Trees and K Nearest Neighbor. After reviewing the results, the Neural Network method has been selected to predict the Movie ratings based on the predictor variables.



1. **Conclusion & Recommendations:**

Since the best fitted model for categorical data is Logistic regression and for numeric data is Neural Networks, the below insights are based on these models.

* The highest Hit movie rate was in the years between 1951 to 1960. There has been a decline in the percentage of Hit movies since then.
* The top two countries contributing to highest number of IMBD score are US and UK with 75.2 and 9.3% respectively.
* Actor 1 and actor 2 Facebook popularity drive the movie success at a higher rate as compared to director Facebook popularity.
* The number of voted users is an important factor impacting movie success.
* Budget is important and there is a strong correlation between budget and movie rating.
* The Logistic Regression model has been selected as the best model to predict categorical outcome Hit / Flop, based on the accuracy of the model in Validation dataset at 75.45%
* In order to predict IMDB equivalent movie ratings, the Neural Network method was selected as the best model with a R – square of 49% and RASE of 77%.

**4 References**

Galit Shmueli, Peter C. Bruce, Mia L.Stephens, Nitin R.Patel, Data Mining for Business Analytics Book

<https://nycdatascience.com/blog/student-works/web-scraping/movie-rating-prediction/>

<https://help.imdb.com/imdb?ref_=cons_nb_hlp>

<http://usir.salford.ac.uk/18838/1/Wessex_movie.pdf>

[https://nycdatascience.com/blog/student-works/web-scraping/movie-rating-prediction/](https://www.google.com/url?q=https://nycdatascience.com/blog/student-works/web-scraping/movie-rating-prediction/&sa=D&source=hangouts&ust=1524002431353000&usg=AFQjCNGU-TB4gPyuOeUVjBaetpLJCDXMLA)

[https://drive.google.com/drive/u/1/folders/17VICkHIIg-5tCxAdHGfDZcby5kEFd9-Q](https://www.google.com/url?q=https://drive.google.com/drive/u/1/folders/17VICkHIIg-5tCxAdHGfDZcby5kEFd9-Q&sa=D&source=hangouts&ust=1524002431353000&usg=AFQjCNFPZZ88lRrg3B5rldPgdQSJjHZcdA)

[https://en.wikipedia.org/wiki/Linear\_discriminant\_analysis](https://www.google.com/url?q=https://en.wikipedia.org/wiki/Linear_discriminant_analysis&sa=D&source=hangouts&ust=1524015716256000&usg=AFQjCNEbrCfpOztdhylXxk5gqA31eMsUYw)

<https://en.wikipedia.org/wiki/Linear_regression>

Appendix

Original Dataset referenced for above analysis - <https://data.world/popculture/imdb-5000-movie-dataset>

Processed Dataset

**ANOVA**

ANOVA Analysis conducted to select the significant variables are documented below.

