IMDB Predictions Report

Phanindra Varma Sagiraju

December 2016.

Predicting Future Movie Rating

Introduction:

Now-a-days data analytics is playing an important role in almost every business to identify new opportunities by harnessing their data. Now organizations are understanding the significance of how collecting the data that streams into their business and applying data analytics is helping in improving the value of their business. Data analytics is adding speed and efficiency to the business. Big data and data analytics improving the ability to analyze the data immediately and aiding in faster and better decision making. Use of analytics also making it possible to launch new products and services to the customers by analyzing the customer needs and satisfaction.[1]

Every deal, every strategy and every operation is surrounded by risks in each and every business irrespective of their size. This is true for a start-up and for a multi-billion business too. Ability to predict the future of the organization or at least predicting a small factor and its impact within a section of an organization is incredibly valuable in recent times because it will help to overcome the possible risks in the business. Predictive analytics is playing crucial role here.

Predictive analytics when applied to real world can help us in preventing some risks. For example, if we have knowledge about stock market beforehand, bears and bulls of the day etc., then we can make a better decision about buying and selling of stocks. We got interested in field of film industry. If we can predict the ratings of a movie before release, then it can help movie makers save lot of money and time. After going through different papers published on this topic, we understood that movie rating is a factor that people are depending on to watch a movie.

Movie rating is one of the important factor in making a decision of whether to watch a movie or not. These ratings are not only influencing the viewer's interest but also indirectly influencing the revenue of that particular movie i.e., for example when two movies are released at same time, then people tend to watch a movie that is having a higher rating, in this way movie ratings are affecting the revenue too. There are lot of factors that will be influencing the ratings of the movie. Sometimes a movie was expected to get good rating based on the buzz around the movie which was created by the big

[1] Thomas H. Davenport and Jill Dyché. (May 2013). Big Data in Big Companies. International Institute for Analytics Retrieved from http://www.sas.com/resources/asset/Big-Data-in-Big-Companies.pdf

names surrounded the movie like director, huge cast and budget. But the result might turn out to be opposite some times and these ambiguous situations makes it very difficult to predict the success of a movie. Jack Valenti, President and CEO of the Motion Picture Association of America (MPAA) once said, "No one can tell you how a movie is going to do in the marketplace. Not until the film opens in darkened theatre and sparks fly up between the screen and the audience.[2]

Movie making industry is one among the many industries, adopting analytics and business intelligence (BI) techniques with increased frequency. As the production costs growing in exponential manner – stakes around them also growing in direct correlation. Movie making become a high-risk investment as lot of money is involved. Hence to avoid the potential loss or failure, movie makers are adopting BI tools to analyze the matrices that will aid in the box office success.[3]

With the availability of internet and social media viewer was no more a passive receptacle, as audience are throwing tomatoes online using blogs, twitter, Facebook timelines etc., now-a-days. Movie industry is trying to recapture the lost relation between the movie maker and the viewer by converting reactions of the audience into meaningful insights with the help of emerging analytical tools. With the help of data analytics movie makers are keeping an eye on the viewers and crafting the way how movies get made, marketed and distributed in other hand.[4]

Background/Motivation:

It is very difficult to predict the performance of a movie before its release. Previously it was assumed that big names of the directors, lead cast, production houses, their crazy combinations and high budget etc., were majorly impacting the success of a movie. But sometimes the most expected success also become utter failure and made the movie makers to realize that there are so many other underlying factors were there to influence the result of a movie and the need for the factors exploration using data analytics.

[2] Davenport, Thomas H., and Jeanne G. Harris. "What People Want (and How to Predict It)."MIT Sloan Management Review 50.2 (2009)

[3] NECTO: Panorama Software. Analytics adoption in the film industry. (August 21, 2013). Retrieved from https://www.panorama.com/blog/analytics-adoption-in-the-film-industry/

[4] James Boast. Big data and Hollywood: A love Story. Retrieved from http://www.theatlantic.com/sponsored/ibm-transformation-of-business/big-data-and-hollywood-a-love-story/277/

Take an example of recent failures that results in huge losses to their production studios. One among them was sci-fi adventure film John Carter which was released in 2012. The story was based on the novel written by Edgar Rice Burroughs. Even though with the positive reviews and good overseas market the film ended up as a failure, incurring loss of $200 million to studio Disney. Poor marketing techniques in the United States was the major cause for this loss.

Another example, with a potential actor Johnny Depp as a lead actor Disney's film Lone Ranger ended up in around $190 million loss. Experts and critics say that the story's lack of relevance to the today's youth was the major factor for the movies failure[3]. Like Lone Ranger recent Ben-Hur was most disappointed film of 2016. Even though it is a replica of 1959's Oscar winning movie, this one failed miserably. As audience are expecting something fresh every time, critic reviews are influencing the result of the movie and also the fluctuations in the economies making the film makers life difficult.

User ratings and critic ratings are majorly affecting the performance of a movie at the box office. Even the big names around the movie are also unable to compensate the effects of these ratings in saving the movie from failure. These ratings are acting as "influencers" on the public in watching the movie and acting as "predictors" of the movie's box office performance too. As influencing the user's interest in watching a movie ratings are indirectly influencing the movie revenue and directly predicting the box office performance. [5]

Movie has lot of dimensions that can influence its rating. The motivation behind this project is to analyze the factors that are influencing the movie ratings. We carefully studied and evaluated the possible factors that can influence the rating of a movie with the help of Internet Movie Database (IMDb) historical data. IMDb is a user maintained resource available for free online with information related to more than 400 thousand movies, television series.[6] IMDb is one of the widely used and popular movie rating database by the users. This intern motivated us to select this database. We gathered most of the attributes that can possibly influence the ratings of the movie and bundled possible combinations to get a good rating for a movie with the help of different data visualization techniques.

[5] Suman Basuroy. (2003). How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets. *Journal of Marketing, 67(4),* 103-117. doi: http://dx.doi.org/10.1509/jmkg.67.4.103.18692

[6] M. Saraee, S. White & J. Eccleston. (2004). A data mining approach to analysis and prediction of movie ratings. WIT Press, ISBN 1-85312-729-9. Retrieved from < http://usir.salford.ac.uk/18838/1/Wessex_movie.pdf>

Literature Review:

Literature review was done to gain more knowledge about the objective of the project and the approach towards the objective. We gathered important information during this literature review which helped us a lot in choosing modelling the data. It helped us to identify the required attributes, how to clean the data, what are all the evaluation parameters and the algorithms to be used.

Method for literature review:

Relevant papers were found in Google scholar and Rutgers Library database by searching using key words like "Movie ratings", "Data analysis and movie user ratings", "IMDb database", "Influence of movie ratings on movie performance", "Precision and recall", "Data cleaning", "Neural networks in movie rating", "Predictive analysis" etc. Information from the relevant papers were gathered together.

The paper, "Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures", by Awad, Dellarocas and Zhang helped us to understand the impact of online movie ratings on the box office success of a movie. Statistical models were developed using these ratings to predict movie revenues. In addition to that they also demonstrated the correlation between critics (infomediaries) and online, offline word of mouth. This paper helped us to understand our objective in a business outcome prospective.[7]

The paper, "Predicting Movies User Ratings with IMDb Attributes", by Ping-Yu Hsu et al stated that directors, actors and writers are the important predictors for the movie ratings and neural networks was well-predicted model for predicting user ratings with less prediction absolute error. They used classification techniques (linear combination, multiple linear regression, neural networks) to develop user ratings prediction model. This paper helped us to understand how a linear or convex combination can be used for predictive analytics. A linear combination is constructed from a set of terms by multiplying each term by a constant (weight of the attribute) and adding the results. For a given finite number of predictor variables x_1, x_2,…, x_n, a linear combination of these independent variables is in the form.

$$w_1x_1+w_2x_2+\ldots\ldots..+w_nx_n$$

[7] Chrysanthos Dellarocas, Xiaoquan (Michael) Zhang, And Neveen F. Awad. (2007). Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures. *Journal of Interactive Marketing, 21(4),* 23-45

Whereas, the weights of the independent variables w is ≥ 0 and $\sum w_i = 1$ (i=1,2,……n)

This paper helped us to understand what are all the important attributes that are influencing the forecasting movie ratings, what evaluation technique should be used to determine the efficiency of the model.[8]

The paper, "Pre-release Box-Office Success Prediction for Motion Pictures", by Rohit Parimi and Doina Caragea helped us to understand how machine learning techniques/algorithms are influencing the decision making in business environments recently. This paper demonstrated how to use these algorithms in predicting the movie success before it is released. Authors used a transductive algorithm to construct features for classification which alleviated the movie independence assumption that traditional learning algorithms make. The results showed that this approach has improved the accuracy of the classification.[9]

The paper, "Predicting Movie Success Using Neural Network", by Arundeep Kaur and AP Nidhi demonstrated a methodology in which they assigned a weightage to the historical data that can influence the success or failure outcome for the movie based on the descriptive statistics. Then the data set is subjected to neural network. They checked the match between actual class and the predicted class. The accuracy obtained was high. They ran the neural network based on Levenberg-Marquardt (LM) algorithm. It is difficult to interpret the hidden layers in the neural networks. But LM algorithm which is a mixture of Gradient descent and Gauss-Newton iteration have a different approach towards hidden layers that helps to identify and constructed the combinations that can improve the true positives.[10]

The paper, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation" by Cyril Goutte and Eric Gaussier helped us in determining which parameters to be used in evaluating the dataset. It helped in adapting the confusion matrix that helps in figuring out which parameters are important when looking at the model. For this project, recall and false positive rate are the relevant parameters, as

[8] Hsu PY., Shen YH., Xie XA. (2014) Predicting Movies User Ratings with Imdb Attributes. In: Miao D., Pedrycz W., Ślęzak D., Peters G., Hu Q., Wang R. (eds) Rough Sets and Knowledge Technology. RSKT 2014. Lecture Notes in Computer Science, vol 8818. Springer, Cham

[9] Rohit Parimi, Doina Caragea. Pre-release Box-Office Success Prediction for Motion Pictures. Machine Learning and *Data Mining in Pattern Recognition. 7988,* 571-585

[10] Arundeep Kaur and AP Nidhi. (2013). Predicting Movie Success Using Neural Network. International Journal of Science and Research (IJSR), 2(9), 69-71 < https://www.ijsr.net/archive/v2i9/MTIwMTMxNTk=.pdf>

these two focused on increasing the true positives and decreasing the false positives respectively.[11]

The paper, “Competitive Dynamics and the Introduction of New Products: The Motion Picture Timing Game”, by Robert e. Krider and Charles B. Weinberg helps us to understand how critical is the timing of the new product’s introduction into the market. They have conducted an equilibrium analysis of the product introduction timing in a particular season. This intern motivated us to add additional attributes of release date, day and season to our data set to check the extent of these factors influencing the movie rating.[12]

Approach:

We considered the data set which was publicly available in Kaggle (https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset). The output data consists of 28 attributes for 5000+ movies as below,

"movie_title", "color", "num_critic_for_reviews", "movie_facebook_likes", "duration", "director_name", "director_facebook_likes", "actor_3_name", "actor_3_facebook_likes", "actor_2_name", "actor_2_facebook_likes", "actor_1_name", "actor_1_facebook_likes", "gross", "genres", "num_voted_users", "cast_total_facebook_likes", "facenumber_in_poster", "plot_keywords", "movie_imdb_link”, "num_user_for_reviews", "language", "country", "content_rating", "budget", "title_year", "imdb_score", "aspect_ratio".

Color – we have both black and white and color movies, we noticed that color movies got higher ratings compared to black and white.

movie_facebook_likes – popularity of a movie can be understood by looking its popularity in social media by having number of likes it has got

actor_1_name, actor_2_name, actor_3_name – actors in the movie have a considerable amount influence on ratings, we had information about three main characters that are present in the film

[11] Cyril Goutte and Eric Gaussier. A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation. *Advances in Information Retrieval. 3408*, 345-359

[12] Robert e. Krider and Charles B. (1998). Competitive Dynamics and the Introduction of New Products: The Motion Picture Timing Game. *Journal of Marketing Research, 35,* 1-15.

actor_1_facebook_likes, actor_2_facebook_likes, actor_3_facebook_likes – when we want to know how much an actor is influencing people to watch them, we have to know about the actor’s popularity, so these attributes about the Facebook likes these actors have got this tells about their popularity.

Gross – it is the amount of revenue they have earned after the release of the movie, by looking at the gross we can say whether this movie had made profits or not, and we know that generally a movie which has a higher rating has a high probability that it is going to get profits

Genres – It indicates the type of the movie. For example, thriller, action, comedy, love, science fiction etc., different genres attracts different sections of audience

Language – we took movies belonging to several languages, so that we can observer whether language is influencing audience to watch a movie, as regional language movies couldn’t get as higher ratings as globally spoken languages

Country – we took movies that has got release in different countries

title_year – we have movies from the year 1920 – 2016

num_critic_for_reviews – it has value of number of critics that has reviewed the movie.

num_voted_users – it has value of number of users that have rated a particular movie. Users must be registered with IMDb

facenumber_in_poster – it has value of number of faces that are present in a poster of a particular movie

plot_keywords – it has keywords that are there in a movie such as love, friendship, death, fight etc., i.e., words that are describing the plot

num_user_for_reviews – it has value of number of viewers that has given reviews about that movie

content_rating – it has different value such as PG – 13, movies which strongly caution parents to avoid children watching that content, G – all ages are admitted to that movie, etc., basically it has rating about the content in the movie, R, PG

All the data was collected using “scrapy” library in python[13]. List of 5000+ movies were obtained from the numbers website (http://www.the-numbers.com/movie/budgets/all), and saved as .json file. These titles were searched in IMDb website to obtain the actual movie links. All the data was scrapped from the individual movie page by sending HTTP request to each movie page using links. Source code for the procedure was available in https://github.com/sundeepblue/movie_rating_prediction, so we used this code and developed some parts according to our need.

To the above data we have added 4 more variables. They are “release date”, “week day”, “season”, “movielens rating”. “As Professor Christie Nelson suggested us to add ratings from other website so we have chosen movie lens ratings”. Movie lens ratings were gathered from other website not through scraping. We mapped two files i.e., IMDb dataset file and movielens ratings file and ended up with 3000+ same movie and added movie lens rating to IMDb data set. We handled the missing values for other instances who doesn’t have movie lens ratings by replacing with an average of non-missing values.

The original code included the scraping of the data from IMDB site to fetch the movies details into .json file, which consists of the HTML data. The python code loads the .json file and parse all the variables into valid format. All the 28 variables are directly from the referred tag except for the release date.

Movie release date was there as a sub part in the content rating html tag. We extracted the date in python. But the output date was in character form. We have changed the release date from character to date, fetched the week day from the date and calculated the season of the period using conditions on number of days of a season.

Below is the code to pull the date from content rating tag

```
if movie['content_rating'] is None or len(movie['content_rating']) == 0:
    parsed_movie['week_day'] = "No_Date"
  else:
    if len(movie['content_rating']) == 18:
      x = movie['content_rating'][15].strip()
      if len(x) > 0:
        try:
          k = x.split(" ")
```

[13] Chuan Sun. Predict Movie Rating. August 22, 2016 https://blog.nycdatascience.com/student-works/machine-learning/movie-rating-prediction/

```
            if len(k) == 4:
                k = " ".join(k[0:3])
                d = datetime.strptime(k, '%d %B %Y')
                parsed_movie['week_day'] = d.weekday()
                parsed_movie['date'] = d
                parsed_movie['season'] = get_season(d.timetuple().tm_yday)
        except:
            parsed_movie['week_day'] = x
            parsed_movie['date'] = "Exception_18"

elif len(movie['content_rating']) == 16:
    y = movie['content_rating'][13].strip()
    if len(y) > 0:
        try:
            k1 = y.split(" ")
            if len(k1) == 4:
                k1 = " ".join(k1[0:3])
                d1 = datetime.strptime(k1, '%d %B %Y')
                parsed_movie['week_day'] = d1.weekday()
                parsed_movie['date'] = d1
                parsed_movie['season'] = get_season(d1.timetuple().tm_yday)
        except:
            parsed_movie['week_day'] = y
            parsed_movie['date'] = "Exception_19"

elif len(movie['content_rating']) == 14:
    y = movie['content_rating'][11].strip()
    if len(y) > 0:
        try:
            k1 = y.split(" ")
            if len(k1) == 4:
                k1 = " ".join(k1[0:3])
                d1 = datetime.strptime(k1, '%d %B %Y')
                parsed_movie['week_day'] = d1.weekday()
                parsed_movie['date'] = d1
                parsed_movie['season'] = get_season(d1.timetuple().tm_yday)
        except:
            parsed_movie['week_day'] = y
            parsed_movie['date'] = "Exception_19"
else:
    parsed_movie['week_day'] = "Series_No_Date"#movie['week_day']
```

To fetch the seasons, the logic used is

```
spring = range(80, 172)
summer = range(172, 264)
fall = range(264, 355)

def get_season(doy):
    # "day of year" ranges for the northern hemisphere
    # winter = everything else
    if doy in spring:
      season = 'spring'
    elif doy in summer:
      season = 'summer'
    elif doy in fall:
      season = 'fall'
    else:
      season = 'winter'
    return season
```

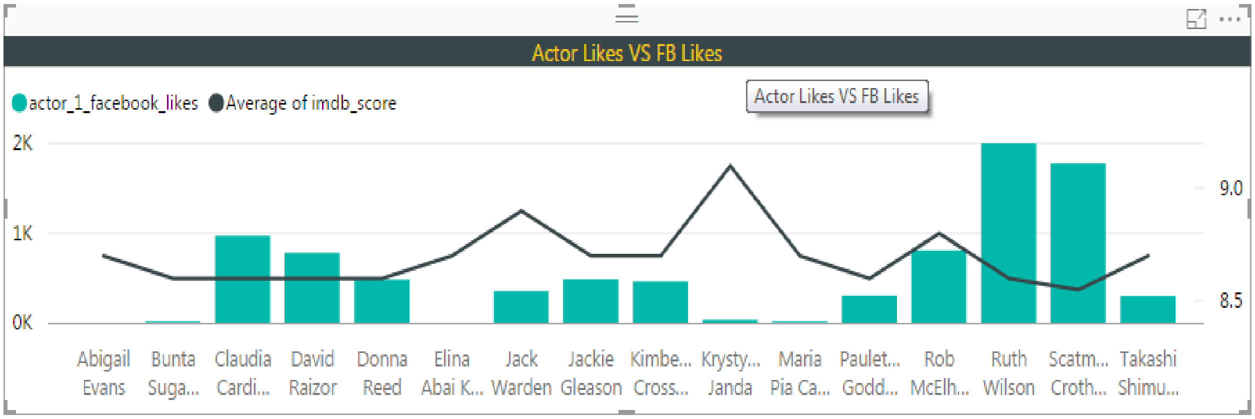
Numeric values were assigned to the week day as "0 – Monday", "1 – Tuesday", "2 – Wednesday", "3 – Thursday", "4 – Friday", "5 – Saturday" and "6 – Sunday". Assigned seasons were "winter", "spring", "summer" and "fall". Found missing values for around 400 columns, which are TV series like friends, Star Wars etc. We don't have a release date for these and we considered them as missing values.

The reason behind adding day, season attributes is that we thought that movie release in holiday season for example, summer might affect ratings and movie released on weekends might also have a high influence on ratings. When we analyzed the pattern of days on which movie was released, we observed that majority of the movie were release on Fridays.

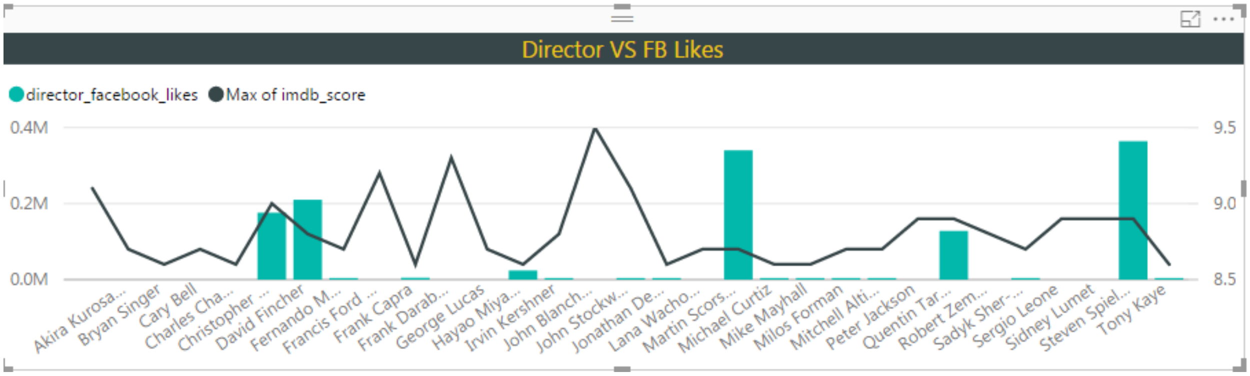
Data Cleaning:

After having all the attributes in hand, we started cleaning the data. Before going to cleaning the data, as a part of analytics, we want to find what has happened in the past to find out the future inferences. So, we have analyzed different attributes to our target variable and how they have effected in the past. We have more than 30 attributes, and of them some of the attributes are futile. So we have decided to reject those attributes in order to make our analyses more effective. We have used "Microsoft PowerBI" to do this task. We have checked different attributes against target variable to check whether the respective variable is effective or not. Consider for example, popularity of actors on social media. The variables in our data, which corresponds to this are actors Facebook likes and directors Facebook likes.

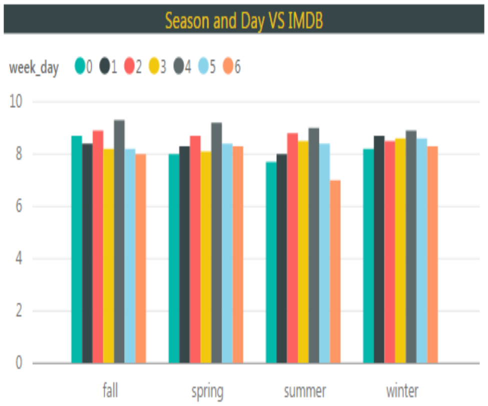
By observing the below graph, even though actors like Abigail Evans, Bunta Sugavara has relatively very low popularity in terms of social media, the average IMDB score is very similar to the actors who has high popularity.

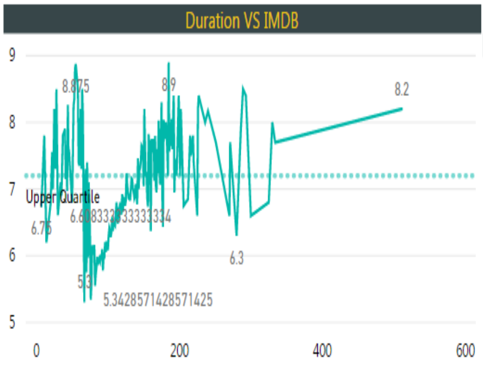


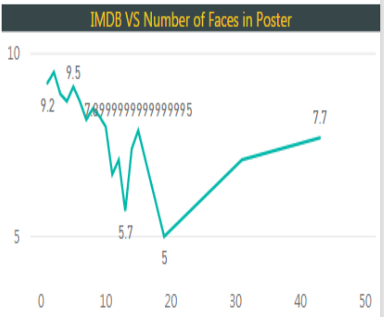
Similar to the analyses made in terms of actors, same is applied for director's popularity.



Consider the following analyses.





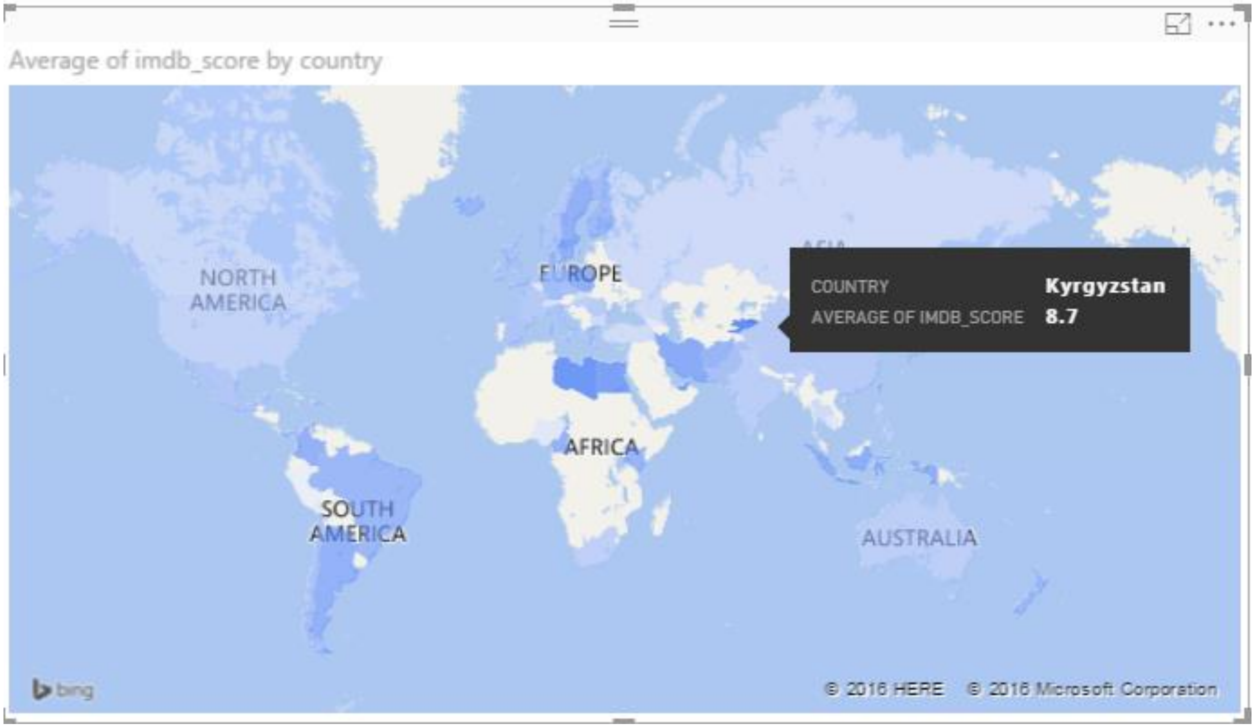


From graph, Season and Day vs IMDB, we can say that the movies which are released on day 4 i.e. Friday, have higher IMDB ratings than the movies, which are released on other days of the week. We can also say that the movies, which were released on Sunday of summer got lesser ratings than the movies, which released on same day but different season. (Ex: Sunday in winter). So, Season and day are an important attribute in finding the future IMDB ratings.

Consider the graph, Duration vs IMDB. We can observe that in a particular range of duration, the average IMDB ratings are higher than the upper quartile (median of the upper half of a data set). Even though there are some outliers in the IMDB ratings for a particular duration range, we can generalize that if duration was in between two points, the IMDB rating was greater than upper quartile or lower.

Similarly for number of Faces in poster Vs IMDB Ratings. We can see that from 0 to 5, if the number increases, the IMDB and from 5 to 13, if the number of faces increases, IMDB rating decreases. So, Number of faces on poster is an important attribute.

Likewise, we have analyzed for every interval attribute.



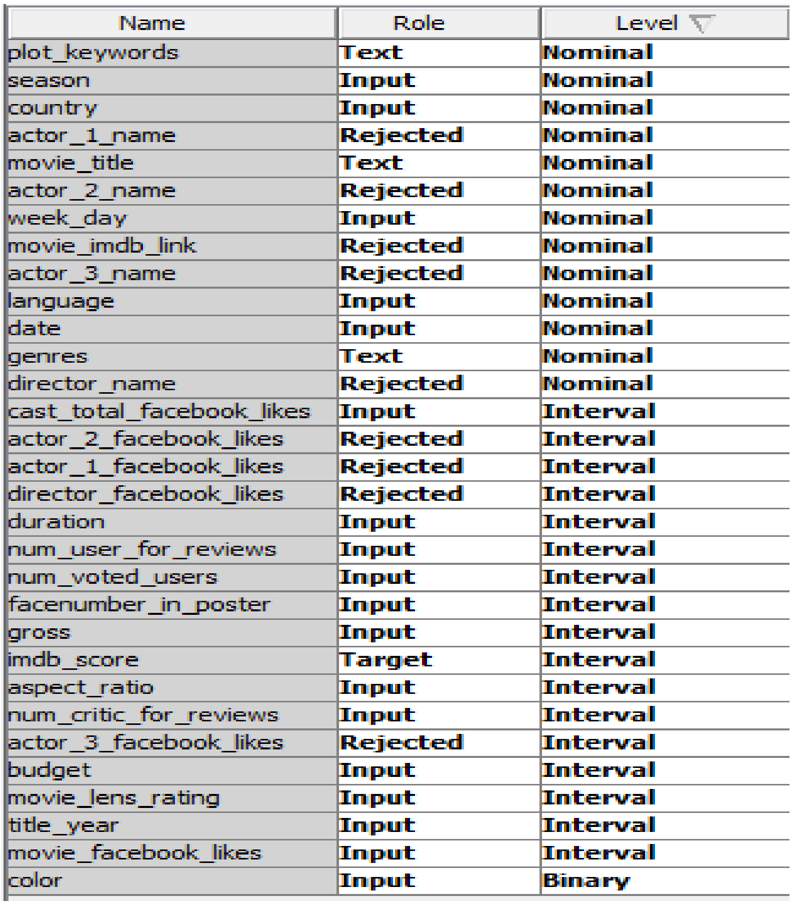
Consider the nominal attribute country. We can see that the data colors are varied according to country. I.e. In the map, we can observe that the average of IMDB scores varies according to country. In the graph highlighted portion, we can observe that the average IMDB score of Kyrgyzstan is 8.7. This might not mean that the movies released in Kyrgyzstan are highly rated. We have not excluded a single attribute, which has slightest possibility of effecting IMDB ratings.

Final Data:

We can see that, even though weekday is interval (0-7), the editor has detected as nominal. This is because of the missing and the wrongly scraped data. In order for our model to perform well, we have to get rid of these. We have replaced all the missing values of the Interval variables with the mean of the non-missing values. Similarly for the class variables, using replacement editor[14], for nominal attributes, which are interval attributes without missing values, we have replaced with most occurring (Mode). For all other missing values, we have observed the pattern and replaced with the following.

For Language: English

[14] SAS E-Miner 14.1 Replacement Node.



Name	Role	Level
plot_keywords	Text	Nominal
season	Input	Nominal
country	Input	Nominal
actor_1_name	Rejected	Nominal
movie_title	Text	Nominal
actor_2_name	Rejected	Nominal
week_day	Input	Nominal
movie_imdb_link	Rejected	Nominal
actor_3_name	Rejected	Nominal
language	Input	Nominal
date	Input	Nominal
genres	Text	Nominal
director_name	Rejected	Nominal
cast_total_facebook_likes	Input	Interval
actor_2_facebook_likes	Rejected	Interval
actor_1_facebook_likes	Rejected	Interval
director_facebook_likes	Rejected	Interval
duration	Input	Interval
num_user_for_reviews	Input	Interval
num_voted_users	Input	Interval
facenumber_in_poster	Input	Interval
gross	Input	Interval
imdb_score	Target	Interval
aspect_ratio	Input	Interval
num_critic_for_reviews	Input	Interval
actor_3_facebook_likes	Rejected	Interval
budget	Input	Interval
movie_lens_rating	Input	Interval
title_year	Input	Interval
movie_facebook_likes	Input	Interval
color	Input	Binary

For Week_Day: 4

For Country: USA

For Date: 10/16/2016 0:00.

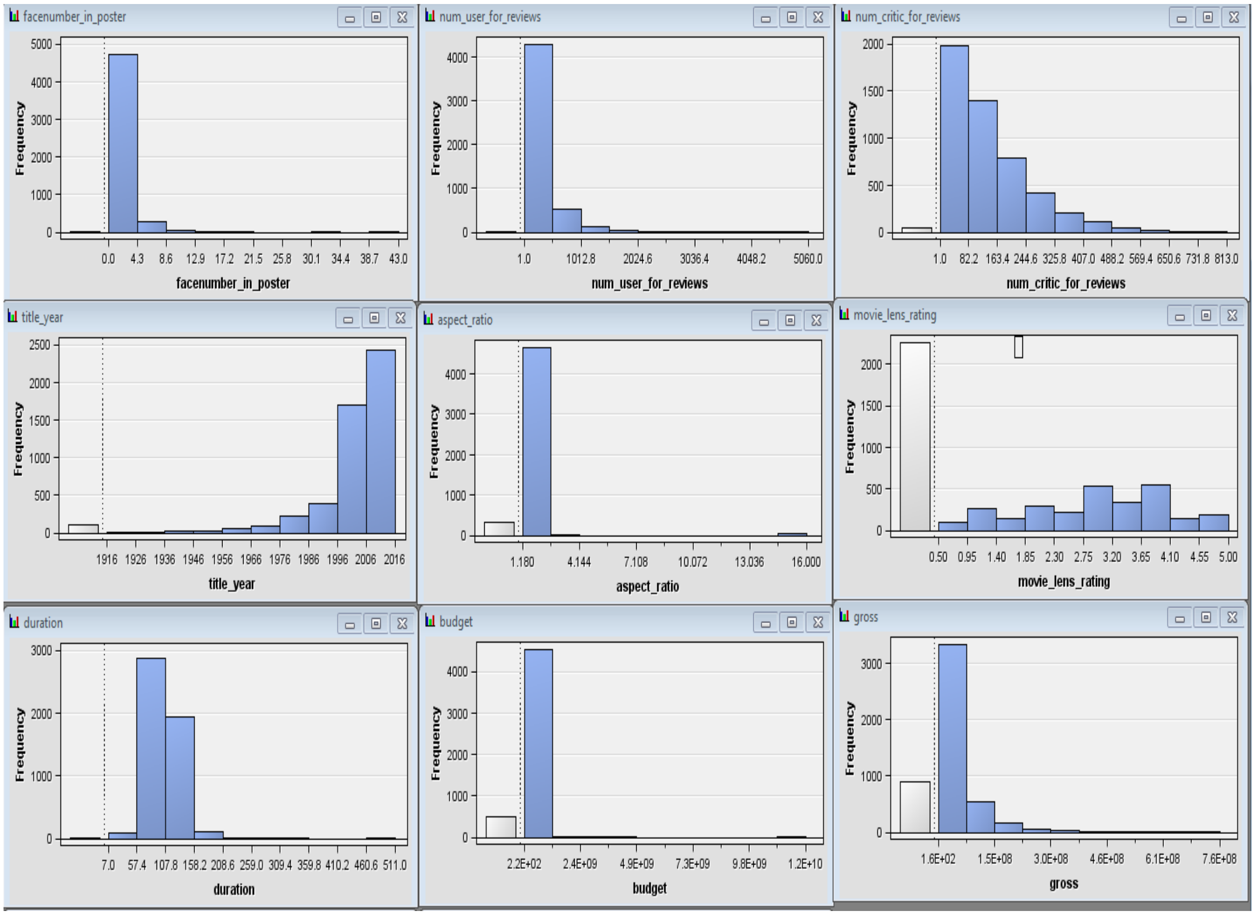
For Color: Color

For Season: Fall

After defining the data, we have to divide the data into training, validation and test data sets. Training data for preliminary model fitting, validation to tune and also used for model assessment. Test data set, also used for prediction using model assessment. This is mainly to avoid "Over fitting of test data set".

Pareto Principle: In programmer sense, we usually sit on developing an application for 80% of the time (includes correcting bugs) and 20% of the time in testing the developing application. So, we have divided our data into 60:20:20, considering Pareto principle.

Before directly jumping to modeling of the given data, we have to see all the variables and remove the skewed distributions if any. Because, we have partitioned our data into three sets, we want to check different types of algorithms to model our training data in order to perform well on validation data set. So, taking the logarithm of a skewed variable and fit by altering the scale and making the variable more "normally" distributed.[15]



We can observe that the interval variables like budget, gross, and duration are skewed. So, we have applied logarithmic transformation in order to get rid of model selecting inputs with highly skewed or highly kurtotic distributions over inputs that yield better overall predictions.

Modeling:

Using this final dataset, we have decided to predict IMDB ratings. For this prediction, we have used some of the supervised learning algorithms such as Decision trees, Logistic

[15] Exploratory Data Analysis, John W Tukey.

Regression and Neural networks. But, while doing the analyses and from literature review analyses, we have realized that the text based inputs were not being analyzed properly. So, we have decided to do text mining on text based inputs such as plot-keywords, genres and Movie title.

Evaluation Metrics:

Different model performances have been compared according to Average Squared Error (focused more on estimate predictions), which is defined as follows.

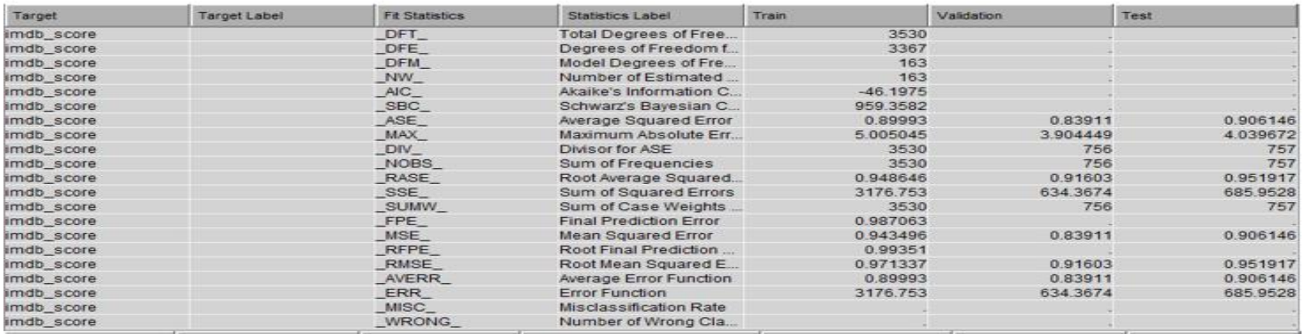
ASE = Sum of Squared Errors (SSE) / Number of Cases (N).

Our main aim is to decrease Average squared error and to decrease the difference of this statistic on different portioned data sets. The best model will minimize the average square error and the difference between train average square error, validation squared error is as minimal as possible.

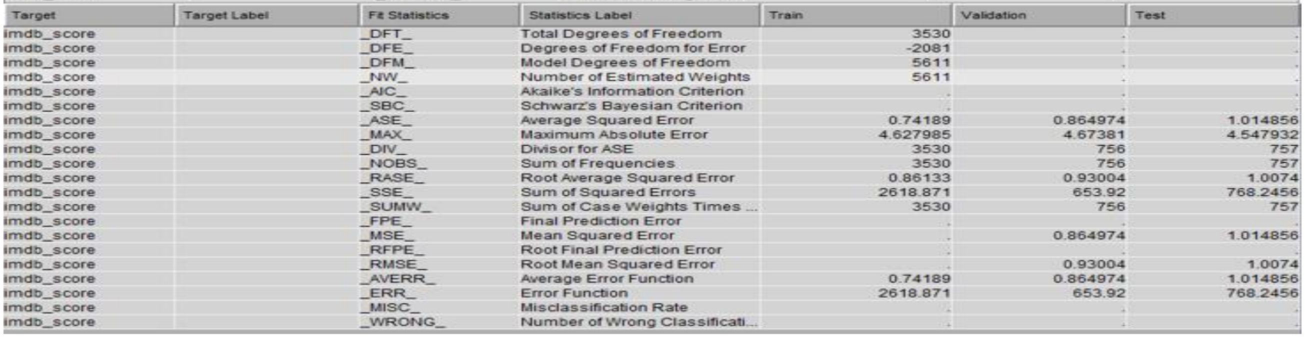
We have modeled two decision trees. One - the maximal tree and the other Interactive decision tree. The first one to analyze the importance of different inputs on the target variable, which have highest –log (p) value and the second one to analyze all the remaining inputs on the training data, which were not used for analyses in the first step.

Choosing inputs for a classification algorithm is very tricky.

For neural networks, we have used inputs from dataset directly and output from regression model to neural network in order to minimize the difference between the validation and training average errors. Consider the following two cases:



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
imdb_score		_DFT_	Total Degrees of Free...	3530	.	.
imdb_score		_DFE_	Degrees of Freedom f...	3367	.	.
imdb_score		_DFM_	Model Degrees of Fre...	163	.	.
imdb_score		_NW_	Number of Estimated ...	163	.	.
imdb_score		_AIC_	Akaike's Information C...	-46.1975	.	.
imdb_score		_SBC_	Schwarz's Bayesian C...	959.3582	.	.
imdb_score		_ASE_	Average Squared Error	0.89993	0.83911	0.906146
imdb_score		_MAX_	Maximum Absolute Err...	5.005045	3.904449	4.039672
imdb_score		_DIV_	Divisor for ASE	3530	756	757
imdb_score		_NOBS_	Sum of Frequencies	3530	756	757
imdb_score		_RASE_	Root Average Squared...	0.948646	0.91603	0.951917
imdb_score		_SSE_	Sum of Squared Errors	3176.753	634.3674	685.9528
imdb_score		_SUMW_	Sum of Case Weights ...	3530	756	757
imdb_score		_FPE_	Final Prediction Error	0.987063	.	.
imdb_score		_MSE_	Mean Squared Error	0.943496	0.83911	0.906146
imdb_score		_RFPE_	Root Final Prediction ...	0.99351	.	.
imdb_score		_RMSE_	Root Mean Squared E...	0.971337	0.91603	0.951917
imdb_score		_AVERR_	Average Error Function	0.89993	0.83911	0.906146
imdb_score		_ERR_	Error Function	3176.753	634.3674	685.9528
imdb_score		_MISC_	Misclassification Rate	.	.	.
imdb_score		_WRONG_	Number of Wrong Cla...	.	.	.

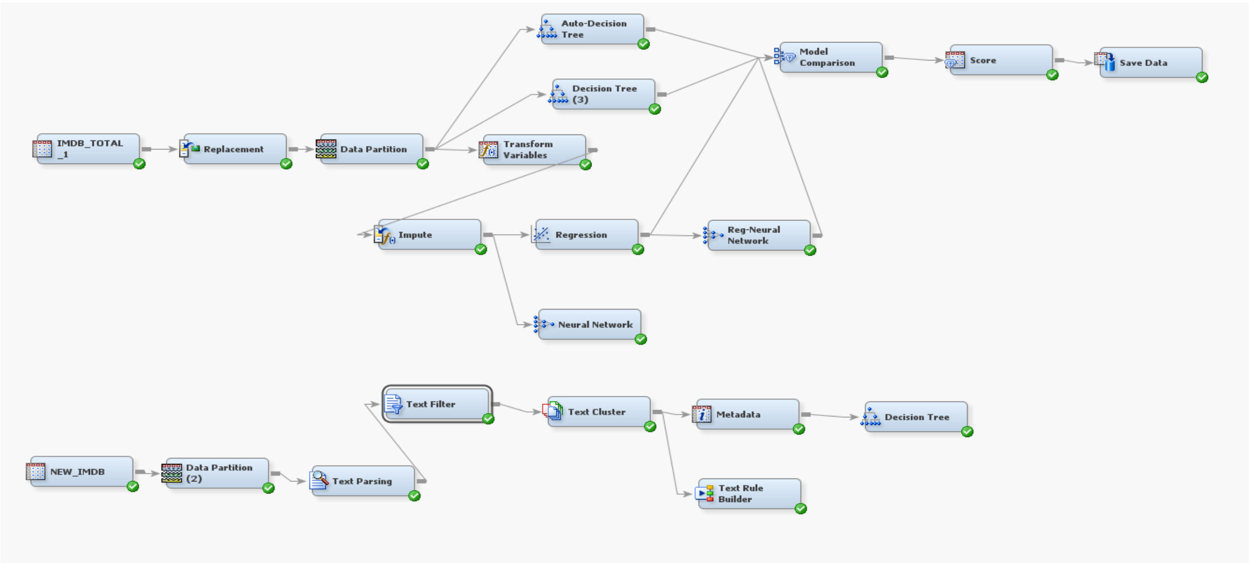


Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
imdb_score		_DFT_	Total Degrees of Freedom	3530	.	.
imdb_score		_DFE_	Degrees of Freedom for Error	-2081	.	.
imdb_score		_DFM_	Model Degrees of Freedom	5611	.	.
imdb_score		_NW_	Number of Estimated Weights	5611	.	.
imdb_score		_AIC_	Akaike's Information Criterion	.	.	.
imdb_score		_SBC_	Schwarz's Bayesian Criterion	.	.	.
imdb_score		_ASE_	Average Squared Error	0.74189	0.864974	1.014856
imdb_score		_MAX_	Maximum Absolute Error	4.627985	4.67381	4.547932
imdb_score		_DIV_	Divisor for ASE	3530	756	757
imdb_score		_NOBS_	Sum of Frequencies	3530	756	757
imdb_score		_RASE_	Root Average Squared Error	0.86133	0.93004	1.0074
imdb_score		_SSE_	Sum of Squared Errors	2618.871	653.92	768.2456
imdb_score		_SUMW_	Sum of Case Weights Times ...	3530	756	757
imdb_score		_FPE_	Final Prediction Error	.	.	.
imdb_score		_MSE_	Mean Squared Error	.	0.864974	1.014856
imdb_score		_RFPE_	Root Final Prediction Error	.	.	.
imdb_score		_RMSE_	Root Mean Squared Error	.	0.93004	1.0074
imdb_score		_AVERR_	Average Error Function	0.74189	0.864974	1.014856
imdb_score		_ERR_	Error Function	2618.871	653.92	768.2456
imdb_score		_MISC_	Misclassification Rate	.	.	.
imdb_score		_WRONG_	Number of Wrong Classificati...	.	.	.

Although, the error rate increased from case 1(Regression input) to case 2(No Regression Input), the difference between validating models has been reduced to a very significant amount.

For analyzing the impact of genres, plot-keywords such as computer, lab, data etc in the description of movie, and as the data is in the form of text, we used text-mining. Making use of different algorithms, we came to an understanding what should be in a genre and in plot-keywords to make sure of higher ratings.

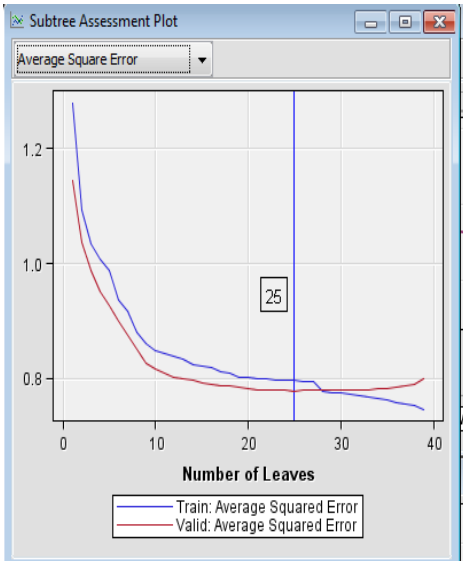
Final Model:

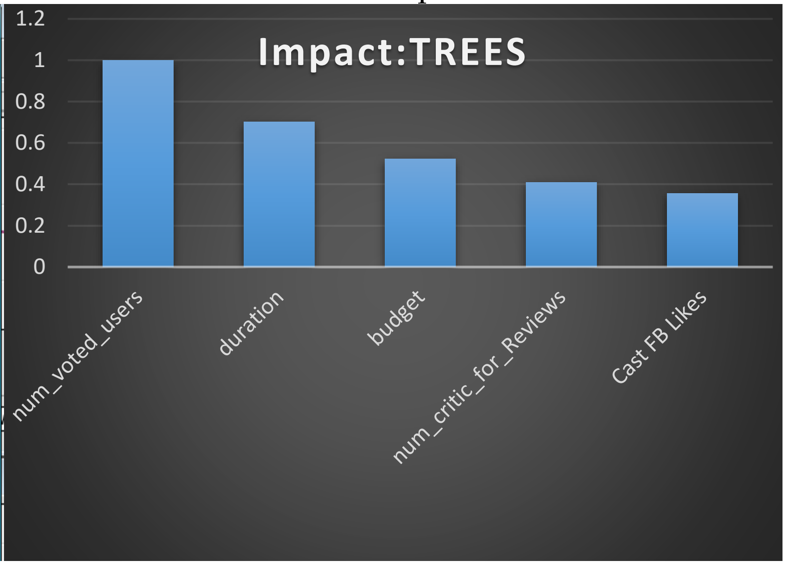


Results:

Decision Tree:

As the depth of the decision tree increases, we can observe that the difference between the mean predicted and mean target is being reduced. This model has been tested on inputs: num_voted_users, duration (modified value for duration), budget, num_critic_for_reviews, and cast_total_facebook_likes. When we observe the results of the decision tree, we have observed that the model has been optimized at 25th leaf.

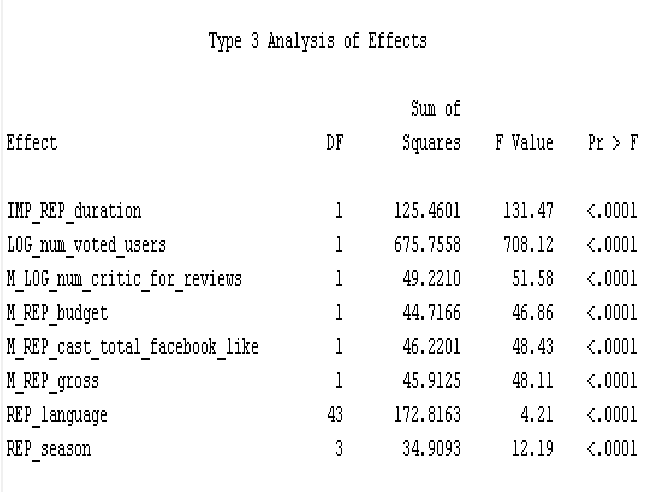




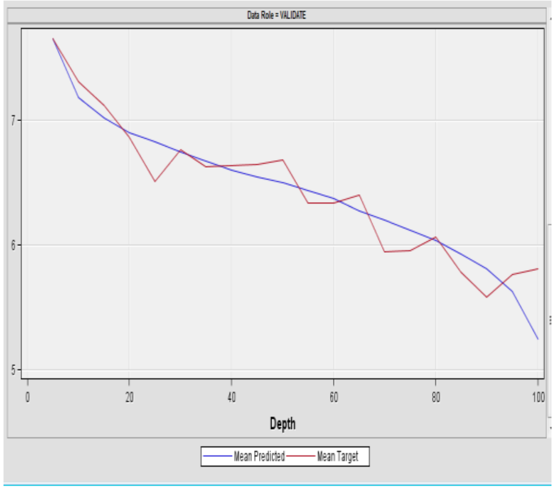
Based on the analyses on the training data, the most effective attributes on predicting IMDB are: number of voted users, duration of the movie, budget, number of critics reviewed, and cast Facebook likes.

Neural Networks:

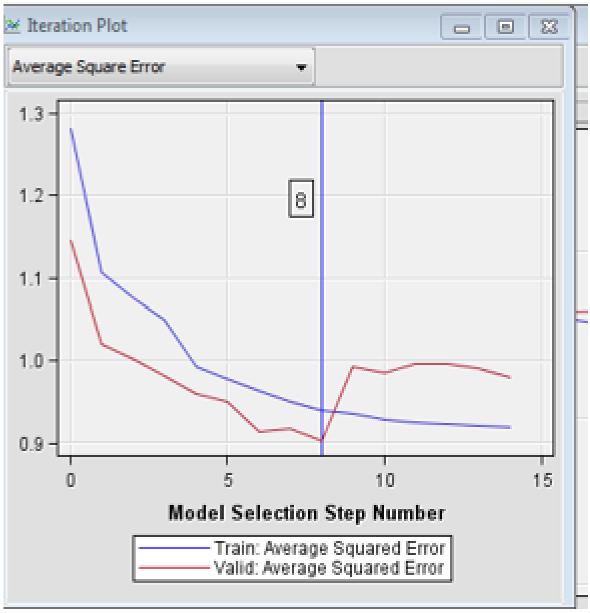
Before analyzing the results of neural network, let us first analyze inputs of neural network i.e. logistic regression analyses.

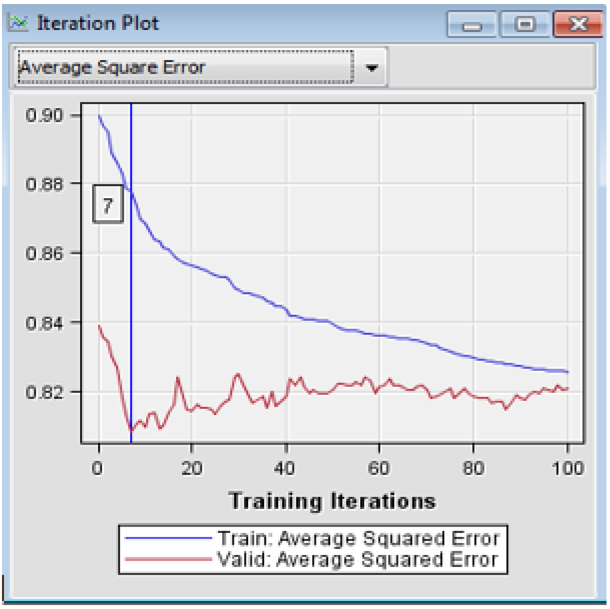


```
                     Type 3 Analysis of Effects

                                          Sum of
Effect                          DF       Squares    F Value    Pr > F

IMP_REP_duration                 1      125.4601     131.47    <.0001
LOG_num_voted_users              1      675.7558     708.12    <.0001
M_LOG_num_critic_for_reviews     1       49.2210      51.58    <.0001
M_REP_budget                     1       44.7166      46.86    <.0001
M_REP_cast_total_facebook_like   1       46.2201      48.43    <.0001
M_REP_gross                      1       45.9125      48.11    <.0001
REP_language                    43      172.8163       4.21    <.0001
REP_season                       3       34.9093      12.19    <.0001
```

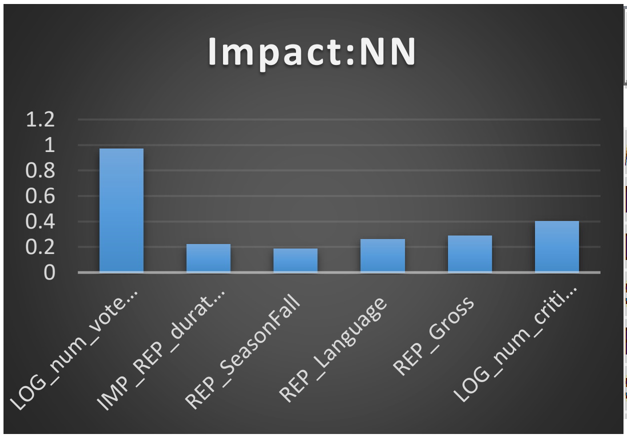


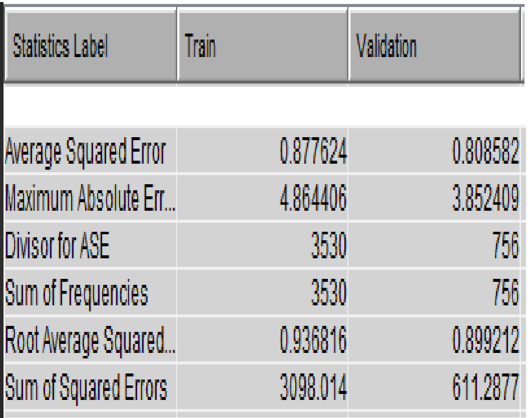
So, we can say that from above, although the inputs are classified properly, when we calculate ratio of mean predicted vs mean target, the error rate is high. So, in order to reduce the error, we have chosen to give this output as input to neural network.


Iteration Plot
Average Square Error
1.3
1.2
1.1
1.0
0.9
8
0
5
10
15
Model Selection Step Number
Train: Average Squared Error
Valid: Average Squared Error



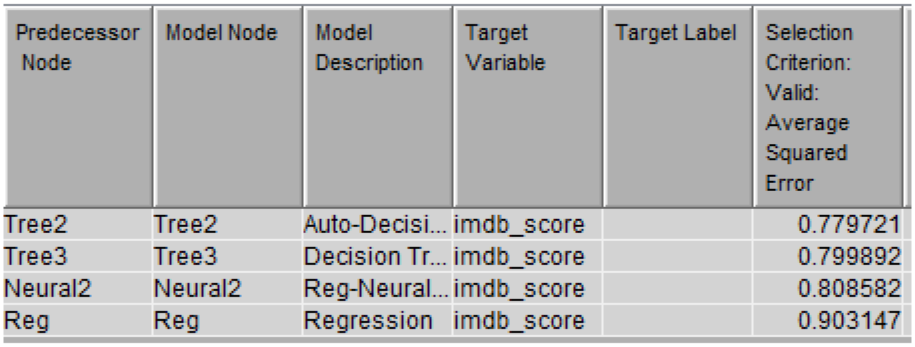
Comparing the Iteration plots of the two models, we can say that neural network has been optimized at an earlier iteration and the final weights used to analyze neural network are as follows. And final statistics of neural network model are also listed.





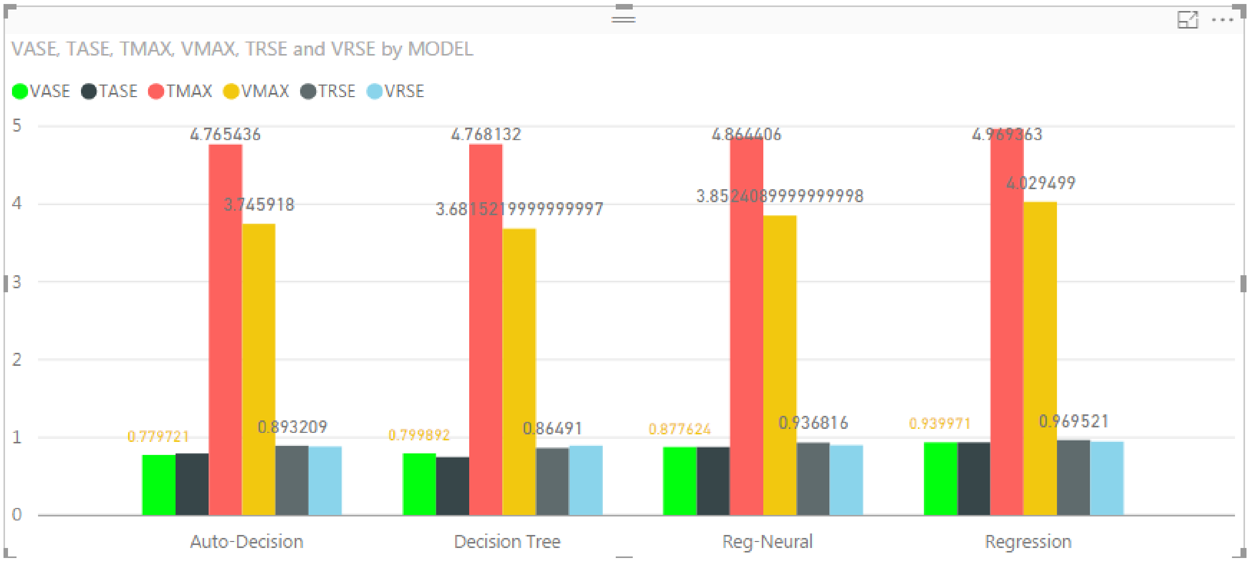
Statistics Label	Train	Validation
Average Squared Error	0.877624	0.808582
Maximum Absolute Err...	4.864406	3.852409
Divisor for ASE	3530	756
Sum of Frequencies	3530	756
Root Average Squared...	0.936816	0.899212
Sum of Squared Errors	3098.014	611.2877

Model Comparison: From the analyses made, we have removed some of the models, which are not performing well. The different model's final statistics are analyzed based on errors.



Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error
Tree2	Tree2	Auto-Decisi...	imdb_score		0.779721
Tree3	Tree3	Decision Tr...	imdb_score		0.799892
Neural2	Neural2	Reg-Neural...	imdb_score		0.808582
Reg	Reg	Regression	imdb_score		0.903147

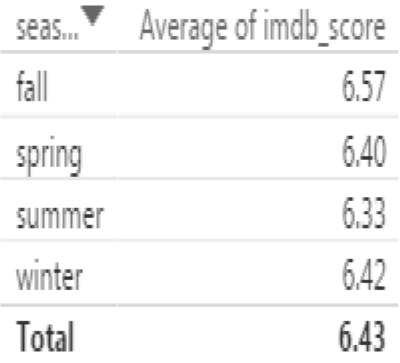
We can see that, Auto-Decision tree based on valid: Average Squared error is selected.



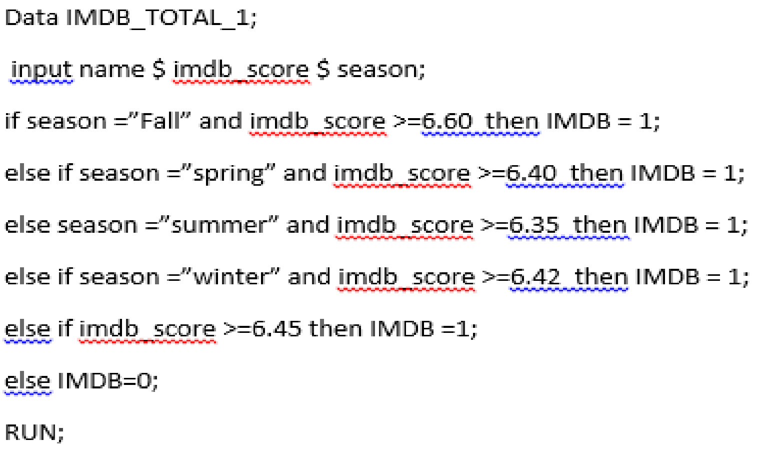
Comparison of different models based on different metrics (Green-VASE)

Text-Mining:

For optimized results, we have changed target variable to binary. The formula, which we have used for converting the values to binary is calculated average IMDB for season and if the value is exceeded the average we have assigned as 1(Positive) and if not 0(Negative).

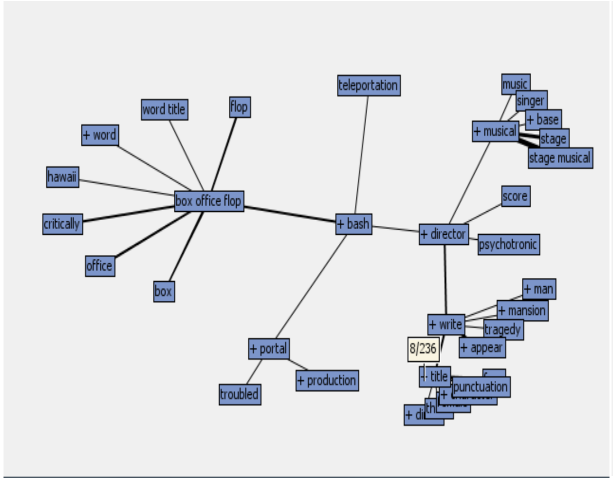


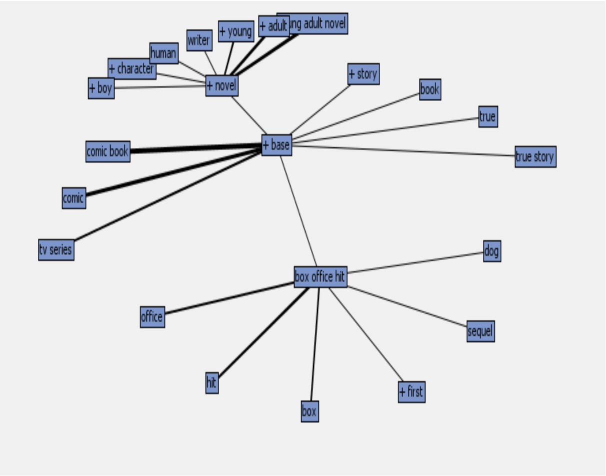
seas...	Average of imdb_score
fall	6.57
spring	6.40
summer	6.33
winter	6.42
Total	6.43



```
Data IMDB_TOTAL_1;
 input name $ imdb_score $ season;
if season ="Fall" and imdb_score >=6.60  then IMDB = 1;
else if season ="spring" and imdb_score >=6.40  then IMDB = 1;
else season ="summer" and imdb_score >=6.35  then IMDB = 1;
else if season ="winter" and imdb_score >=6.42  then IMDB = 1;
else if imdb_score >=6.45 then IMDB =1;
else IMDB=0;
RUN;
```

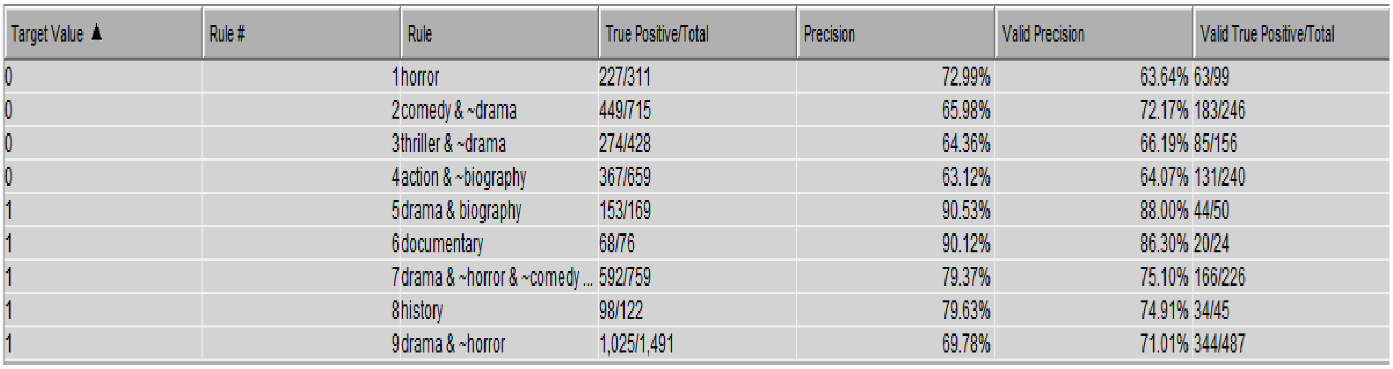
Concept-Linking:





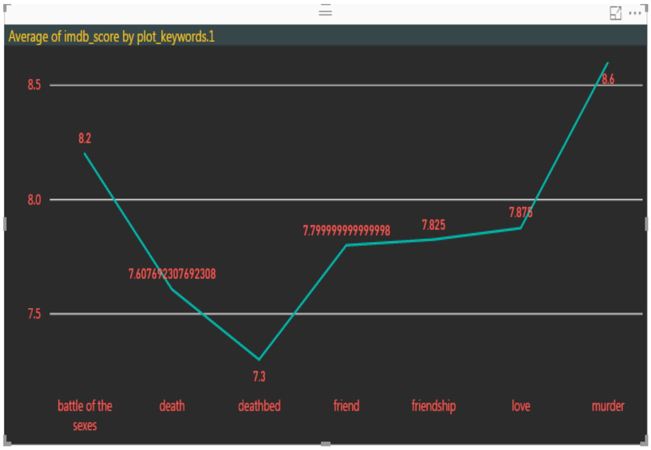
From the interactive filter[16], we have analyzed some of the attributes in the lot keywords, which results in a box office hit (1) or box office flop (0).

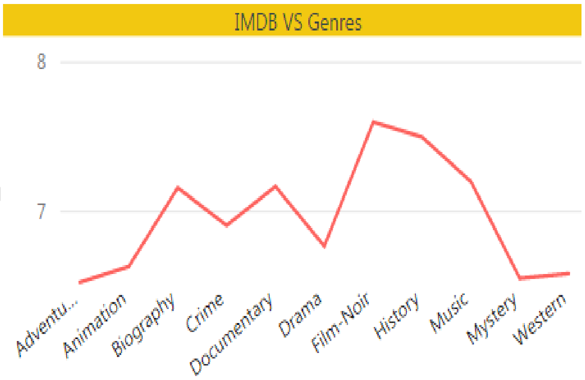
From two concept links analyses: flop is associated with title, director, production and music. And Hit is more associated with base of the movie, which in turn related more to comic book. (Width of the line between the centered term and a concept link represents how closely the terms are associated). From the analyses, we have plotted the weights for different keywords. We have done the same analyses for the genres, we ended up getting the following results.



Target Value ▲	Rule #	Rule	True Positive/Total	Precision	Valid Precision	Valid True Positive/Total
0	1	horror	227/311	72.99%	63.64%	63/99
0	2	comedy & ~drama	449/715	65.98%	72.17%	183/246
0	3	thriller & ~drama	274/428	64.36%	66.19%	85/156
0	4	action & ~biography	367/659	63.12%	64.07%	131/240
1	5	drama & biography	153/169	90.53%	88.00%	44/50
1	6	documentary	68/76	90.12%	86.30%	20/24
1	7	drama & ~horror & ~comedy ...	592/759	79.37%	75.10%	166/226
1	8	history	98/122	79.63%	74.91%	34/45
1	9	drama & ~horror	1,025/1,491	69.78%	71.01%	344/487

From this, we can analyze that drama coupled with biography is the best genre (based on precision) followed by documentary. Horror genre is the least I getting high ratings, which followed by comedy that does not involve drama. So, final conclusions include:

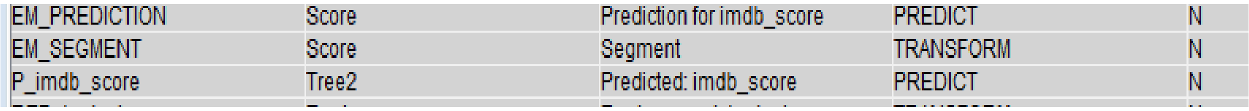




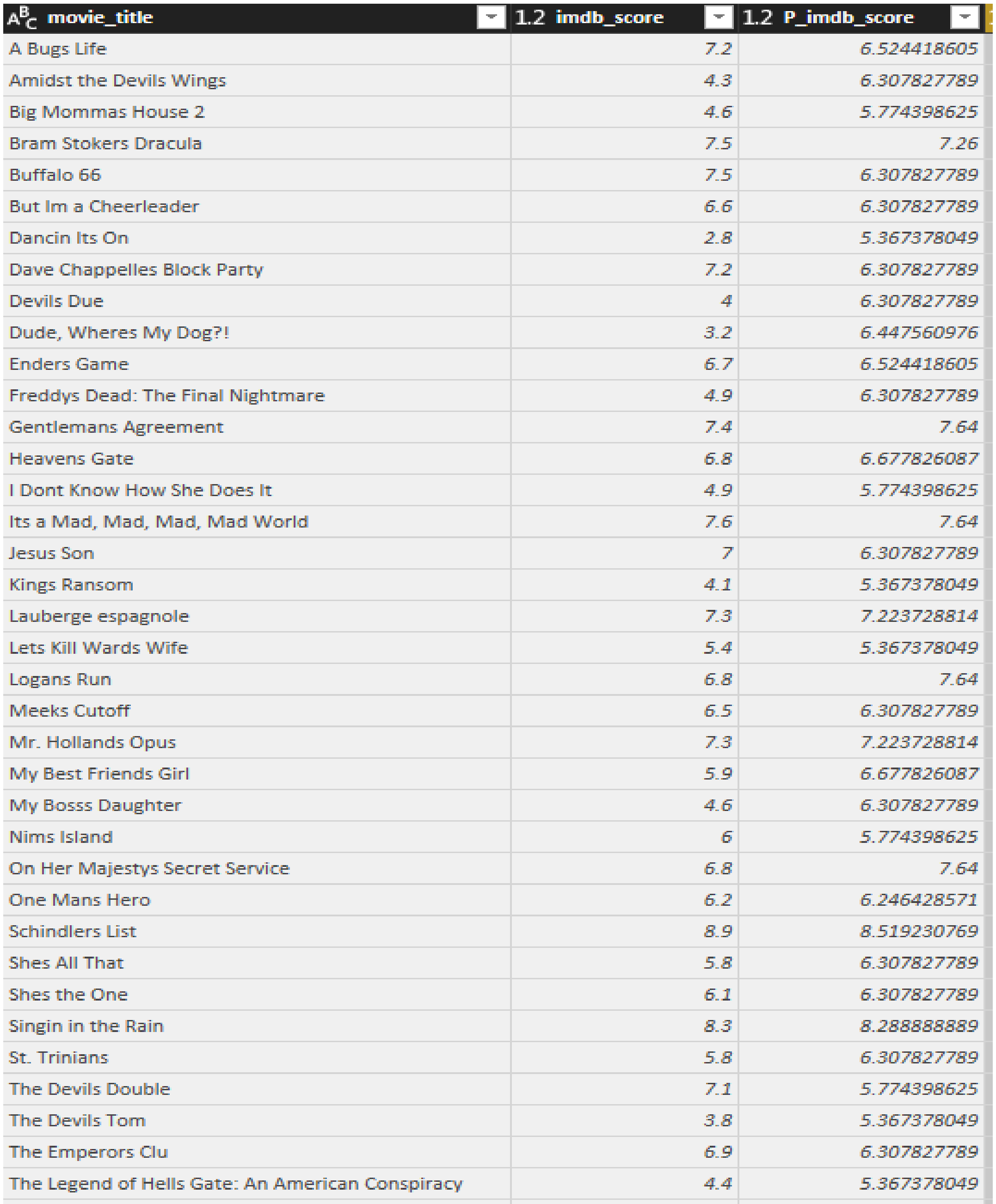
[16] SAS E-Miner Text Filter node

Final Test-data Output:

Average Square Error on the test data has been calculated considering same statistic as before and Decision tree has been selected to predict the variables as ASE is low. We can see that in predicting, model has chosen decision tree 2.



EM_PREDICTION	Score	Prediction for imdb_score	PREDICT	N
EM_SEGMENT	Score	Segment	TRANSFORM	N
P_imdb_score	Tree2	Predicted: imdb_score	PREDICT	N



movie_title	1.2 imdb_score	1.2 P_imdb_score
A Bugs Life	7.2	6.524418605
Amidst the Devils Wings	4.3	6.307827789
Big Mommas House 2	4.6	5.774398625
Bram Stokers Dracula	7.5	7.26
Buffalo 66	7.5	6.307827789
But Im a Cheerleader	6.6	6.307827789
Dancin Its On	2.8	5.367378049
Dave Chappelles Block Party	7.2	6.307827789
Devils Due	4	6.307827789
Dude, Wheres My Dog?!	3.2	6.447560976
Enders Game	6.7	6.524418605
Freddys Dead: The Final Nightmare	4.9	6.307827789
Gentlemans Agreement	7.4	7.64
Heavens Gate	6.8	6.677826087
I Dont Know How She Does It	4.9	5.774398625
Its a Mad, Mad, Mad, Mad World	7.6	7.64
Jesus Son	7	6.307827789
Kings Ransom	4.1	5.367378049
Lauberge espagnole	7.3	7.223728814
Lets Kill Wards Wife	5.4	5.367378049
Logans Run	6.8	7.64
Meeks Cutoff	6.5	6.307827789
Mr. Hollands Opus	7.3	7.223728814
My Best Friends Girl	5.9	6.677826087
My Bosss Daughter	4.6	6.307827789
Nims Island	6	5.774398625
On Her Majestys Secret Service	6.8	7.64
One Mans Hero	6.2	6.246428571
Schindlers List	8.9	8.519230769
Shes All That	5.8	6.307827789
Shes the One	6.1	6.307827789
Singin in the Rain	8.3	8.288888889
St. Trinians	5.8	6.307827789
The Devils Double	7.1	5.774398625
The Devils Tom	3.8	5.367378049
The Emperors Clu	6.9	6.307827789
The Legend of Hells Gate: An American Conspiracy	4.4	5.367378049

Where P_IMDB_score is predicted IMDB score.

Discussion:

From the analyses, we have observed that data mining is a step by step process and carefully selecting inputs for different models is a very critical step. Even though the main goal of our analyses is to find the future ratings, the backbone of this prediction is analyzing the inputs. Considering the attributes we had at the starting, and while adding the new attributes from our knowledge on movies, we thought some of the attributes from the original data and the attributes we have added will have higher impact on movie ratings rather than the other attributes. But to our surprise, from our classification algorithms analyses, the impact of these attributes is very less and even negligible some times. With more than 32 attributes, it is not easy to attain less error rates with such variables in hand. From the initial analyses, we cut shorted to 20 attributes. In decision tree analyses, if we see the final weights analyses infers that number of users rated the movie made very high impact rather than season or day of release. we have observed that the movie ratings indicates popularity of the movie by number of users but the popularity of the movie, which we think as social media has very less impact on the movie ratings.

In our analyses of different algorithms, the impact of selecting the inputs for the algorithms made more impact than the algorithm itself. By splitting the decision trees according to high log worth value of different attributes and previous analyses of the attributes, we have tried to use the interactive splitter[17], but every time puzzled to see the higher error rates than the original tree (Based on higher log worth). Although the thumb rule while predicting an interval variable is to go with regression, the main reason for the poor analyses of logistic regression is the number of attributes. In order to with the transformed inputs and imputed variables on the other hand, regression model performs poorly when the inputs are extreme or outlying values in space. So, in order to not have this repeated again, we have chosen the regression output instead of selecting inputs on our own to neural network model. The final weights analyses of neural network model made the prediction even easier.

The better model among Decision tree, Interactive Decision tree, Regression and neural network is decision tree. Number of users rated the movie, duration of the movie, budget of the movie and total Facebook likes of the cast made more impact on prediction of IMDB ratings. Even with the analyses of these, we have not seen any analyses on text based inputs, as it is very difficult to find nominal statistics on such variables.

[17] SAS Decision tree modeler

With text-mining, what makes a movie with a particular genre to be having higher rating than the others has been analyzed. As we have number of users as our main predictor (From Decision tree), what makes a user to rate a movie? What keywords makes the user more attracted towards rating the movie? Precision has been chosen as our statistic to differentiate the impact.

Conclusion:

From the above analyses, we can say that the future ratings of IMDB ratings are a result of different attributes mix. I.e. If IMDB ratings has to be higher, just making sure of one or two functions of the movie going well is not sufficient. Some of the rules for higher/Lower IMDB ratings include:

- Number of Users ➔ Comic Book reference in genre ➔ Sex in Plot_Keyword ➔ Duration ➔ Budget. (High)
- Thriller =>Drama(Coupled) ➔Number Of Users ➔ DeathBed in Plot_Keyword ➔ Cast Facebook likes. (Low)
- Autobiography => Drama (High)
- Language ➔ Duration > 110.5 ➔ budget > 31500000 (High)

So, explaining one of the above, if thriller coupled with drama, which is then coupled with death bed (instead of death), results in lesser IMDB rating.

Future Objective:

First effective improvement might be improving the data attributes. As we know actors/directors have significant influence on ratings, we can make them more effective by not just looking at their popularity with Facebook likes but also adding the hit rate of a particular actor or directors hit rate, which might increase our model performance of predicting ratings. Some of the interval attributes such as length of the title for a particular director or a particular hero can be added for more analyses. In extra, for predicting gross, number of public appearances before release of the movie can be added.

We have used some post release attributes to model, but if we an exclusively use all the prerelease attributes for modelling then the performance of model to predict future movie rating might increase.

If we possible, we can use word of mouth (WOM) and blogposts data to model the algorithm. WOM has lot of influence on ratings of a movie. Further interest is to evaluate whether this prediction performance can be applied to multiple

datasets as well as large and small datasets so that we can draw more conclusions, as currently we used a well-established data.
Major future improvement can be whether this can be applicable to box office predictions such that scope is widened. Further improvement can be done by using some other algorithms like support vector machines (SVM), M5P etc., to get more accurate results.