**BI.001.003.Final.Predicting IMDb Scores**

**Business Intelligence**

**Fall 2016**

**Predicting IMDb Scores**

**Table Of Contents:**

1. Executive Summary -------------------------------------------------------------------------------------------3
2. Project Background-------------------------------------------------------------------------------------------4
3. Data Description-----------------------------------------------------------------------------------------------4
4. Data Cleaning---------------------------------------------------------------------------------------------------4
   1. DeDuping the Dataset------------------------------------------------------------------------------5
   2. Cleaning Extraneous Data-------------------------------------------------------------------------5
   3. Identification of Surrogate Key-------------------------------------------------------------------5
   4. Outlier Treatment-----------------------------------------------------------------------------------5
5. Data Exploration------------------------------------------------------------------------------------------------5
6. Data Partitioning-----------------------------------------------------------------------------------------------6
7. Data Pre-processing-------------------------------------------------------------------------------------------7
   1. Missing Value Treatment
   2. Variable Transformation
   3. Variable Binning
   4. Variable Selection
   5. Variable Addition
8. Data Modelling-------------------------------------------------------------------------------------------------8
   1. Stepwise Linear Regression Model
   2. DMine Regression Model
   3. Partial Least Squares Model
9. Model Selection------------------------------------------------------------------------------------------------15
10. Process Selection-----------------------------------------------------------------------------------------------15
11. Business Insights-----------------------------------------------------------------------------------------------17
12. Conclusion-------------------------------------------------------------------------------------------------------19
13. References-------------------------------------------------------------------------------------------------------19

**Executive Summary:**

The report provides details of predicting IMDb score for a given movie using supervised learning techniques. The different supervised learning models used for prediction are stepwise forward selection linear regression, Dmine regression and Partial Least Square model.

Stepwise forward selection linear model starts with no variable and tests the addition of each variable based on model fit criterion. The model adds variables which statistically improves model fit and repeats this process until none of the additional variable improves the model fit. For the given data the R² value for this model was 47%.

Dmine regression model quickly identifies variables that are useful for predicting target variable. Number of voted users, movie facebook likes and movie duration are chosen as the top three factors that impact the IMDb score prediction for movies by the model. This model has a better fit than linear regression model and the R² value for this model is 51%.

Partial Least Square model combines the features from principal component analysis and multiple regression. This model is useful in predicting the output when the independent variables are highly correlated. The model selects important variable based on projecting the predicted variable and observed variable to a new space. Top variables selected by this model are number of voted users, movie duration and number of critic reviews.

On comparing the performance of the above models, it was determined that Dmine regression is the best model for predicting IMDb score. The selection criteria used for choosing the best model was minimum average least square error. Dmine regression model had the least average square error of 0.78

Along with IMDb score prediction, important factors that influence the score prediction are also identified. The final model chose number of voted users, movie duration and movie facebook likes as top three factors that influence IMDb score prediction. Some of the significant findings that impact a movie’s success are as following:

* IMDb score for a movie has a direct impact on the monetary benefits for a movie. A movie’s gross revenue and profit is high when the IMDb score is above 7.
* Movie posters which have more number of faces get a lower IMDb score.
* Movies which actors and directors with more number of oscar nominations have high IMDb score.
* Movies with genres like history, documentary or a real life story get high IMDb score.
* Movies with positive reviews on social media such as facebook get high IMDb score.

**Project Background:**

How can we tell the greatness of a movie before it is released in cinema? Many people rely on critics to gauge the quality of a film, while others use their instincts. But it takes time to obtain a reasonable amount of critics review after a movie is released and human instinct sometimes is unreliable.Given that thousands of movies are released each year, is there a better way for us to tell the greatness of movie without relying on critics or our own instincts?[2]

The Internet Movie Database (abbreviated IMDb) is an online database of information related to movies. It also provides information about television programs and video games, including cast, production crew, fictional characters, biographies, plot summaries, trivia and reviews operated by IMDb.com Inc,subsidiary of Amazon.com. It is increasingly becoming a popular go - to website for movie buffs. More number of people are deciding whether to go watch a particular movie or not based on IMDb ratings/scores. The English movie industry being a few billion US Dollars worth, we took up this dataset to provide a few business insights towards factors which contribute to the success or failure of a movie.

This project is aimed at predicting the IMDb scores for a movie that is released and identifying the significant factors that contribute to its success. Recommendations and insights are uncovered from the dataset which could in turn be used by movie makers to ensure the commercial success of a movie.

**Data Description:**

The dataset chosen for this project contains about 5000 movie titles taken from the Kaggle website.[1] This dataset contains IMDb data for various movies of different countries, languages and genres over a period of 100 years.

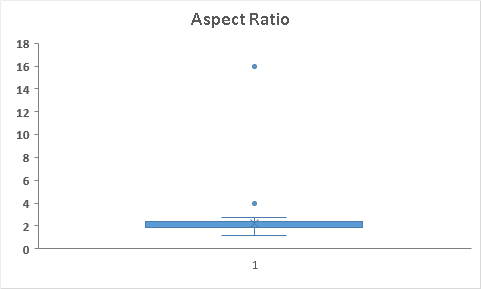
This dataset captures details about the movie cast (Main and Supporting Actors), director, movie release year, facebook likes for the actors, director, total cast and the movie itself, movie budget, gross returns of the movie and the corresponding IMDb scores. It contains 28 columns and 5043 rows. It also consists of data about number of voted critic reviews and the user reviews, duration of the movie, the content rating, aspect ratio and the number of faces in the movie poster. The data is unique at a Movie Name and Movie Year Level. The **target variable** is **IMDB Score** which is an interval variable. This variable ranges from 0.0 to 10.0.

**Data Cleaning:**

The given data has 28 columns and 5043 rows which contained lot of unnecessary and unclean data. Initial data cleaning process is essential to any predicting exercise hence below are some of the steps that were taken to achieve the same.

1. **DeDuping the Dataset :** The dataset is unique by Movie\_Title and Title\_year;some of the movie titles were repeated twice in the same year. This led to redundant rows which constituted up to 2% of the data. This duplication of records was removed and the data was converted to records having unique values by movie title and title year. This was identified manually and removed from the dataset. After this step, the records were unique at a Movie\_Title and Title\_Year level. 126 duplicates by Movie\_Title and Title\_Year were removed.
2. **Cleaning Extraneous Data:** The dataset contained special characters like Â and © in various columns such as actor\_1\_name , director\_name etc,. This was taken care by replacing these special characters with appropriate values.

1. **Identification of Surrogate Key:** The data given is unique at a Movie\_Title and Title\_Year, after the removal of duplicate rows. It would be more convenient to use a single column, which is numeric in nature, that distinguishes the records. Hence a surrogate key was introduced. This surrogate key column was named as **SNO.**
2. **Outlier Treatment:** As shown in the below graph, we observed outliers in the aspect ratio column, upon investigating, it was seen that content\_ratings column included not just movies but TV series data as well. Since predicting IMDb scores for TV- series is out of the scope for this project , we removed rows pertaining to TV-series the dataset. This constituted to about 1.5% of the data.

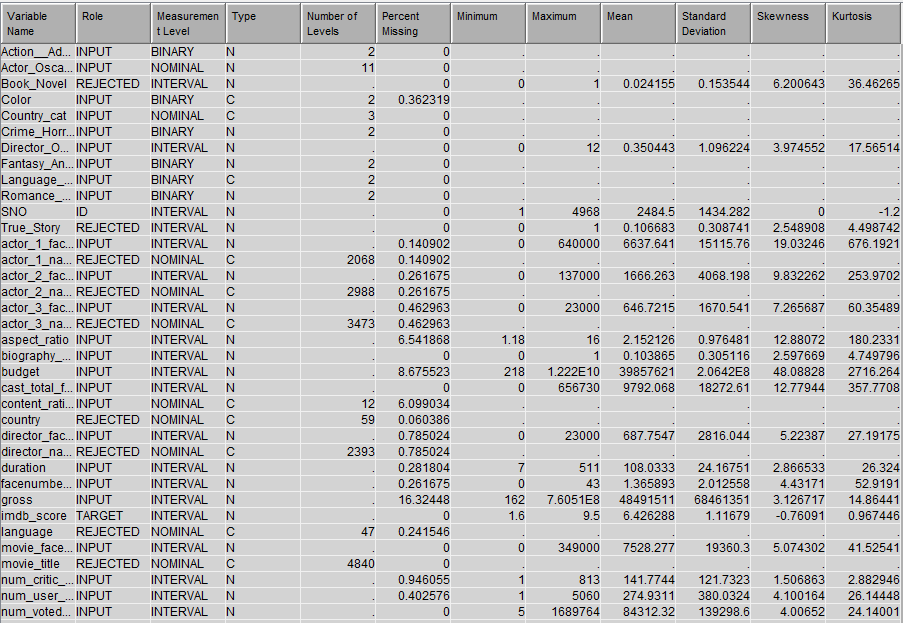


*Fig: Boxplot which detects outliers in Aspect Ratio column*

**Data Exploration:**

The dataset was originally available in the form of Comma Separated Values. This was converted to SAS file using SAS EG and imported to the SAS Enterprise Miner. Stat Explorer node was used to explore the data and to obtain an initial summary of all the variables.

The univariate distribution statistics helped us to identify the variables that are significant and discard the insignificant ones. Values of min, max ,skewness, missing values found, correlation of the variables with respect to the target variables are obtained. Below is a screenshot on the sample statistics and summary statistics of variables :

****

*Fig: Summary Statistic of Input Variables*

**Data Partitioning:**

Data Partitioning was performed to validate and test the trained model. For this purpose the data was split into train test and validation using the Data partition node under the sample tab in SAS E Miner. After the removal of rows pertaining to TV series from the dataset (4968 rows) and data cleaning, the data is split into training, test and validation data in the ratio of 40:30:30.

The train part in the data is used to train the model with the given independent variables. Since more than one model was tested , rather than splitting the data to validation and training, it was split into train validation and testing. Validation data is used to compare several models and choose the best model. Test data is used to evaluate the best model in the new data. Simple random method was chosen to randomly split the data. This method ensures that the data is split into various sets without bias. Eliminating bias is crucial to achieve the best generalised predictive function.

**Data Preprocessing:**

Data from real world mostly lack values containing errors or outliers. Hence preprocessing is an important step before data mining. Data preprocessing involves filling in missing values, removing outliers, and resolving inconsistencies.

1. **Missing Value Treatment:** Missing values in any dataset would hinder the performance of the model. Hence it is mandatory to treat the missing values before any analysis is made. 19 out of 34 variables have missing data. The variables with missing data are mostly right skewed so we are using median to impute the missing values. Budget and gross variable has major % of missing data.

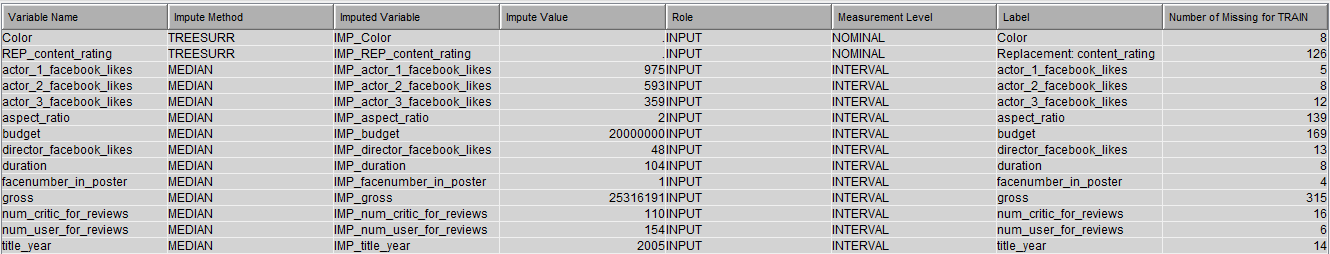


Fig: Missing values being replaced by Medians using the IMPUTE node.

1. **Variable Transformation:** Linear regression assumes that the variables are normally distributed and are not skewed. But in the given dataset some of the variables were positively skewed. In order to eliminate them, the variables were transformed to consider the log of the values present in the variables. The variables named “*Number of Voted Users*”, “*Budget*”, “free number in poster” were positively skewed. Logarithmic transformations were made to these variables. Few continuous variables were binned using optimal binning transformation methods “*Director Facebook likes*”, “*Actor 3 Facebook likes*”, “*Actor 1 Facebook likes,*“*Actor 2 Facebook likes*”, “*Movie Facebook likes*”, “*Cast total Facebook likes”, “Aspect Ratio”.* Since the primary goal of this project is to understand the variables that impact the IMDb scores, some preprocessing techniques like Principal Components is not used since this method replaces the original variables with weighted ones.
2. **Variable Selection:** Some of the variables such as Movie\_URL, Movie\_Title which showed no obvious significance for the prediction of the IMDb score were removed from the dataset. Other variables such as Director\_name , actor\_1\_name , actor\_2\_name and actor\_3\_name were also dropped for analysis as they had more than 2000 odd categories and were in turn represented by their corresponding Facebook\_Likes columns.
3. **Variable Binning**: Univariate analysis was performed on the dataset. In some variables, a single level of category was repeated for about 75% of the time throughout the column. These columns had to be bucketed into meaning buckets for further analysis.
   * *Color* - This variable had two levels; “Color” and “Black and White”. 95.8% data had “Colour” as the value. Hence it was changed into 2 buckets of Color and black and white.
   * *Language* - There were more than 40 languages under this variable but almost 94% of the data had english as the category. Hence this variable was changed into 2 buckets of English and Non English.
   * *Country:* 5 out of 250 odd countries covered over 90% of the data. In order to reduce the number of levels for this variable, new categories such as USA, UK and Others was created.
   * *Genre***:** This column describes the different genres the movie can come under. Each movie had 4-5 genres under pipe delimited values. This was cleaned in MS Excel. There were 27 distinct genres, and they were combined into the below mentioned 5 sub- categories,
     1. Action\_Adventure\_Thriller
     2. Crime\_Horror
     3. Romance\_Family\_Drama\_Music
     4. Biography\_Documentary\_History
     5. Fantasy\_Animation\_Sci-Fi
4. **Variable Addition:** Since a movie rating is dependant mainly on the cast, fields like actor name and director name cannot be neglected all together. We decided to rate the actors and directors based on their oscar nominations. The number of Oscar nominations of the actors and directors were scraped from the website and two new columns were created:
   * Director\_Oscar\_Noms
   * Actor\_Oscar\_Noms

**Data Modelling:**

Data modelling generates the predictive function that is used to forecast the dependent variable. Here the dependent variable is IMDb score as that is the variable of interest. As IMDb score is a continuous variable, methods involving linear regression have to be used. Several models were used to get the Best line fit to predict the data. All the models that were tested are mentioned below:

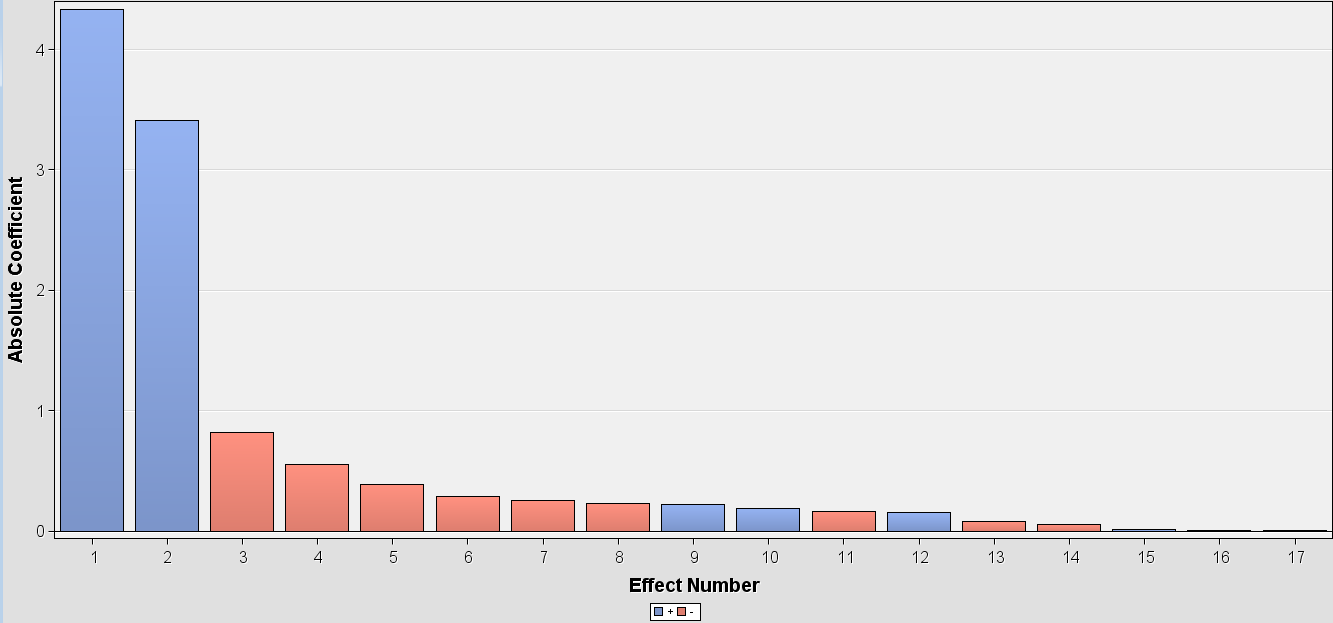
1. **Stepwise Linear Regression Model:**

Stepwise linear regression is a method of regressing multiple variables while simultaneously removing those that aren't important. Stepwise regression essentially does multiple regression a number of times, each time removing the weakest correlated variable. At the end you are left with the variables that explain the distribution best. The only requirements are that the data be normally distributed (or rather, that the residuals are), and that there is no correlation between the independent variables (known as collinearity).

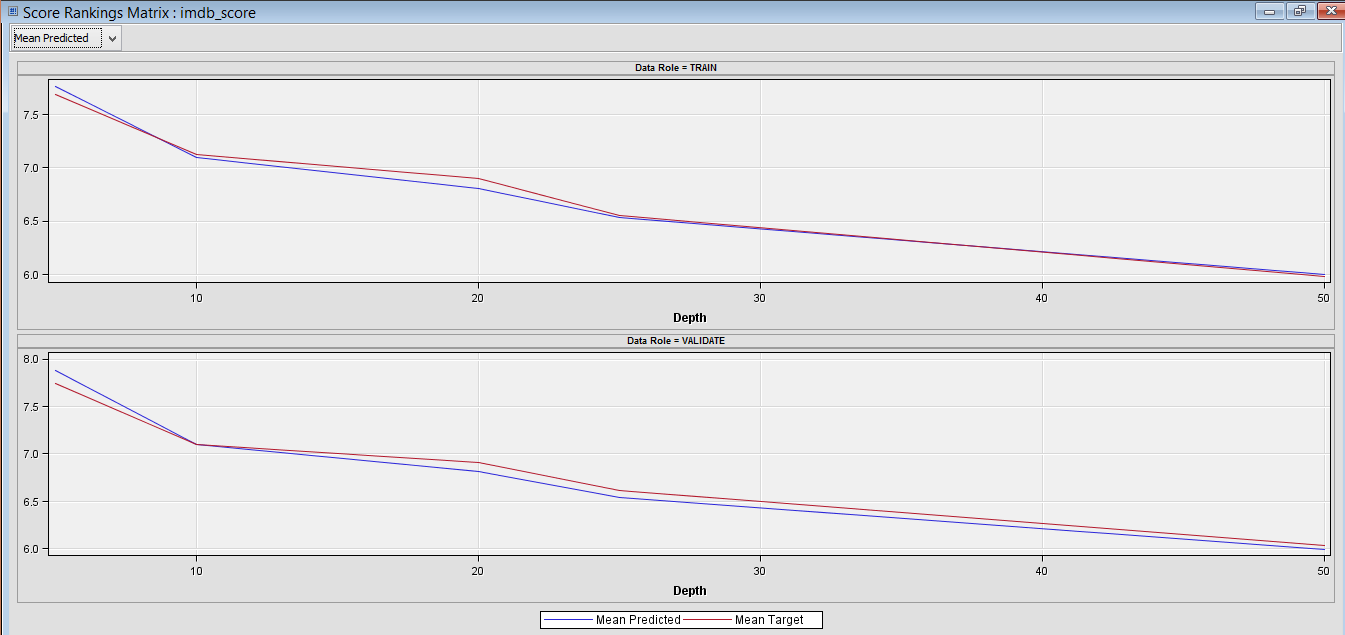
The following variables were considered for prediction,

1. Action\_Adventure\_Thriller
2. Country
3. Crime\_Horror
4. Language
5. Romance\_Family\_Drama\_Music
6. Actor\_1\_facebook\_likes
7. Actor\_3\_facebook\_likes
8. Biography\_Documentry\_History
9. Budget
10. Content\_Rating
11. Duration
12. Facenumber\_in\_poster
13. Movie\_facebook\_likes
14. Num\_critic\_for\_reviews
15. Num\_user\_for\_review
16. Num\_voted\_users
17. Title\_year

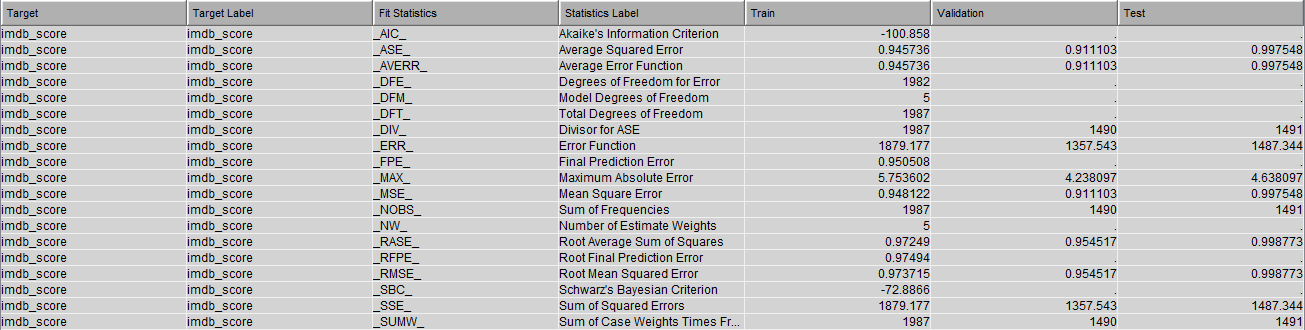
*Effects Plot*

****

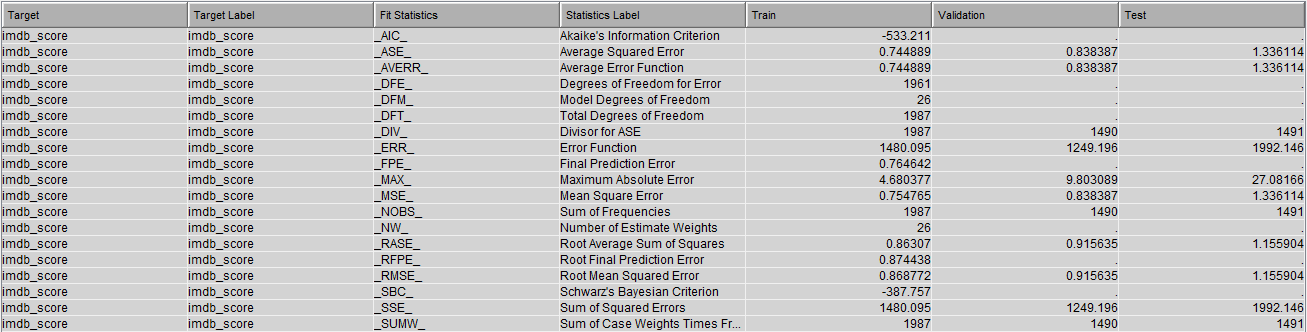
To validate the performance of a model, mean predicted value has to be compared with the mean Target value. The error rates of test train and the validation data set along with the Score rankings matrix is shown below:

****

*Fit Statistics before variable imputation and transformation*



*Fit Statistics after variable imputation and transformation*



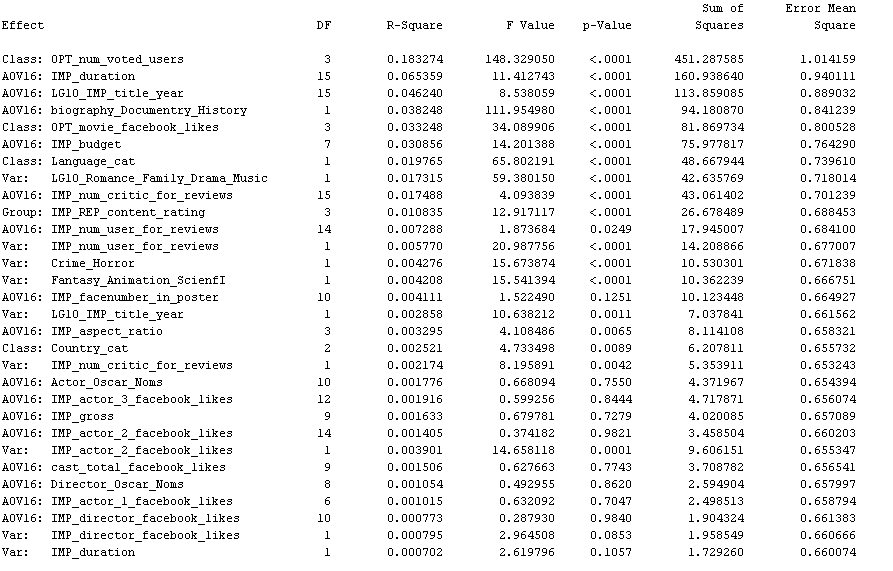
After performing variable imputation and transformation the root mean square error for the model reduced by 4%.

**2) Dmine Regression**

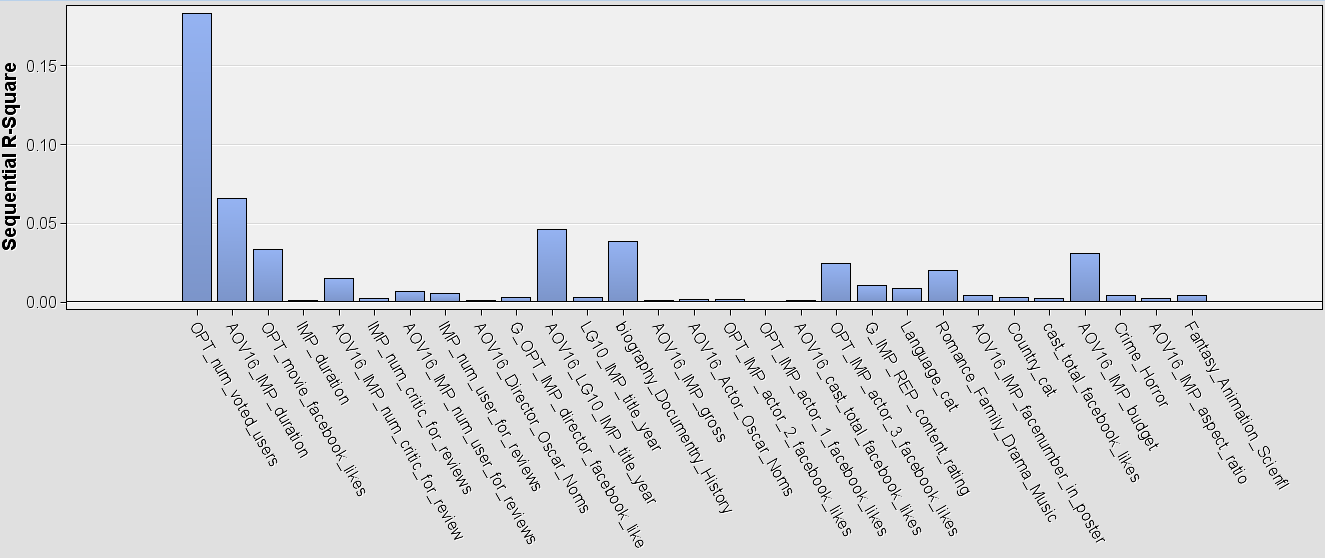
The DMINE procedure enables to quickly identify the input variables that are useful for predicting the target variable(s) based on a linear models framework. Dmine Regression node is used to compute a forward stepwise least squares regression model. In each step, an independent variable is selected that contributes maximally to the model R² value. [5]

Dmine regression helps in quickly identifying the variables in a model using forward regression techniques. Here since the outcome variable is an interval variable, least squares regression technique is used to run the model. Variables which are of primary importance are selected.

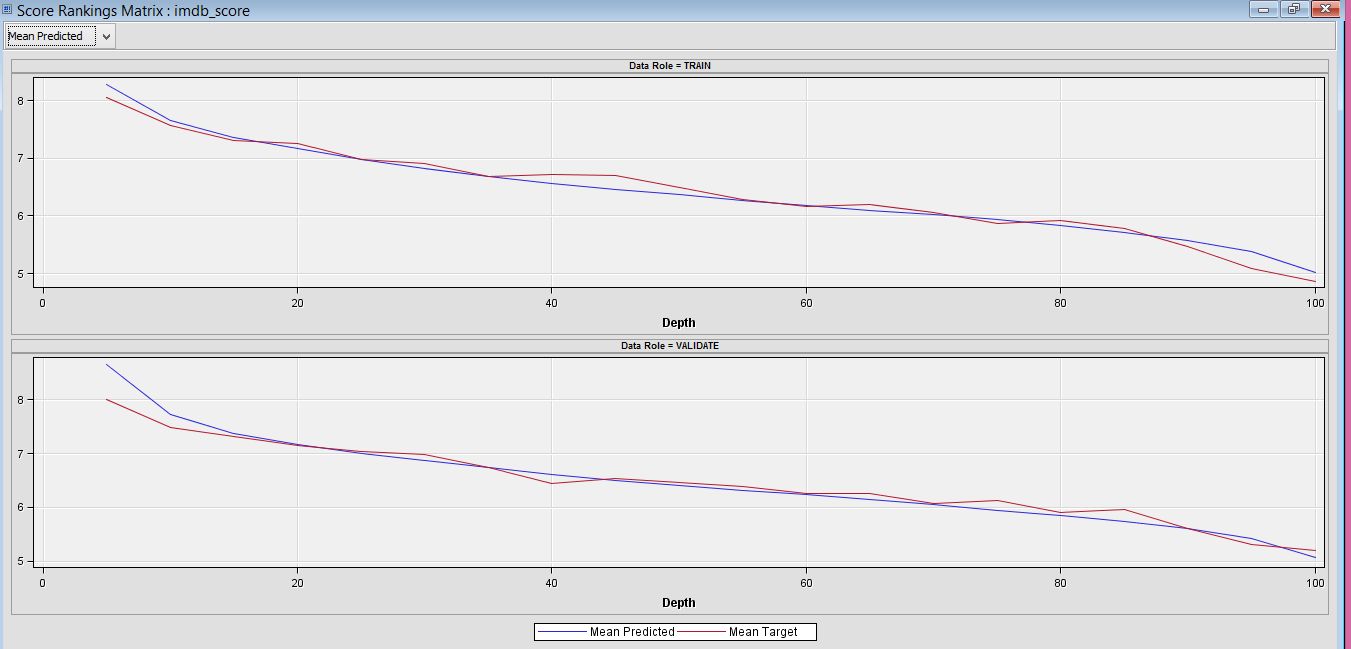
After performing the pre-processing steps, Dmine node is used to run the model. The following variables are chosen as the important variables:



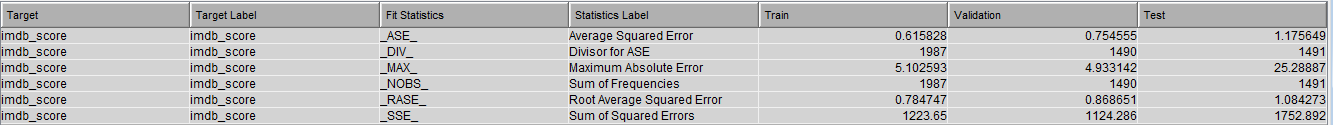
Apart from Number of voted users and facebook likes, movie duration plays a major role. Longer the movie duration, better the score. Other variables which is of importance is mentioned in the below image:



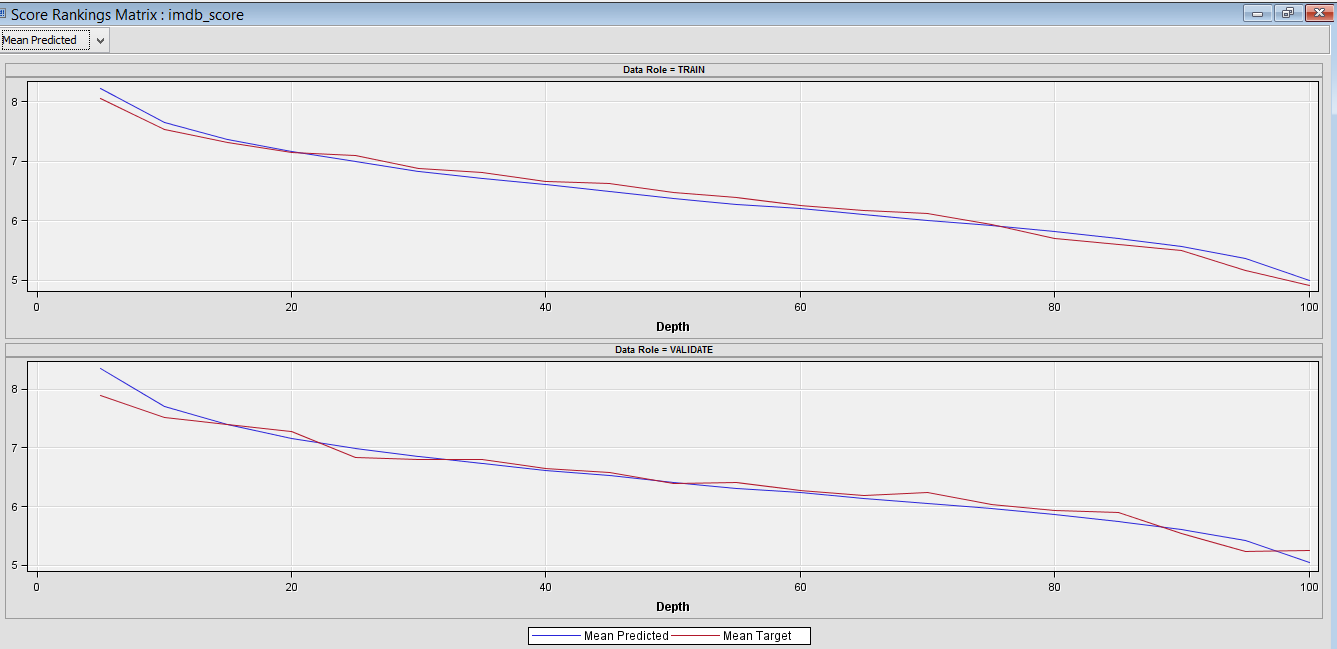
To validate the performance of a model, mean predicted value has to be compared with the mean Target value. The error rates of test train and the validation data set along with the Score rankings matrix is shown below:



*Fit Statistics before variable imputation and transformation*



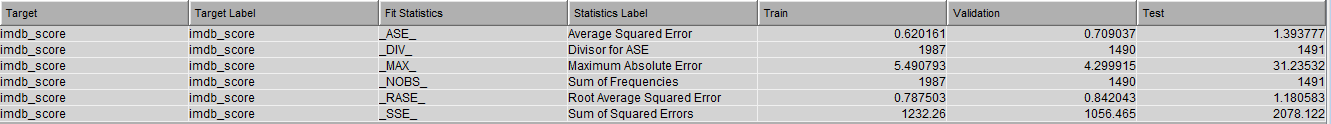
Dmine Regression is again run with imputation to arrive at better results. The missing values are imputed with Median of the entire data. The given data is slightly skewed. So median values were taken instead of mean. Once the model is run, it can be seen that the mean of predicted and target values of train and validation data are more consistent with each other when compared to the model run previously.



*Score Ranking Matrix*

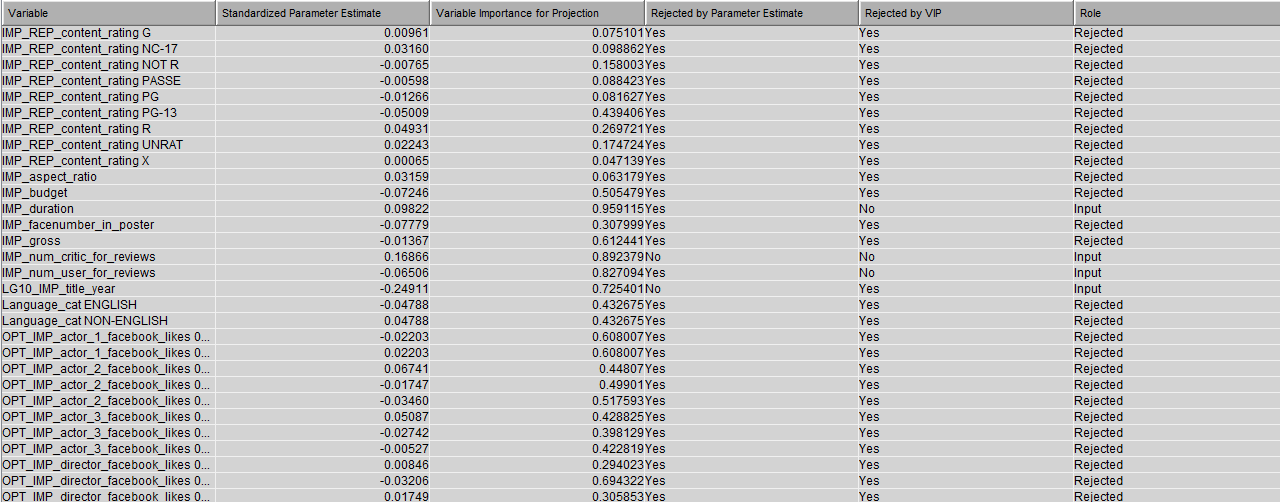
Though the error rates of train and test data are the same, the error rates for validation data has reduced after imputing and transforming the variables. The root average square error has reduced by 4% and the sum of squared errors for validation data has lowered by 6%.

*Fit Statistics after variable imputation and transformation*



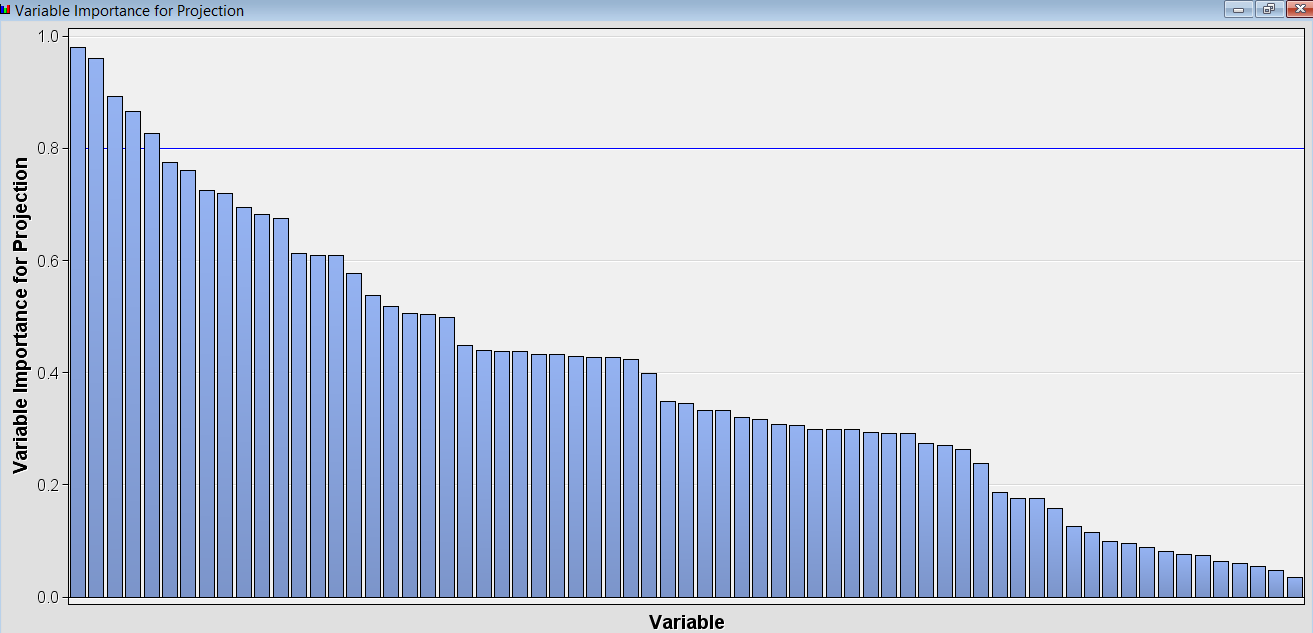
**3) Partial Least Squares**

Partial least Squares is an extension of generalised least squares method. The modeling function forms a linear relationship between the predictor variable to arrive at the optimised solution. Both binary and continuous target variables can be modelled using the least squares method.The additional benefit of partial least square model is that it accounts for variation in the predictors space that are well sampled should provide better prediction for new observations when the predictors are highly correlated.

****

The above chart shows the variables that are given as inputs. Most of the variables are rejected by the model and are considered not significant.

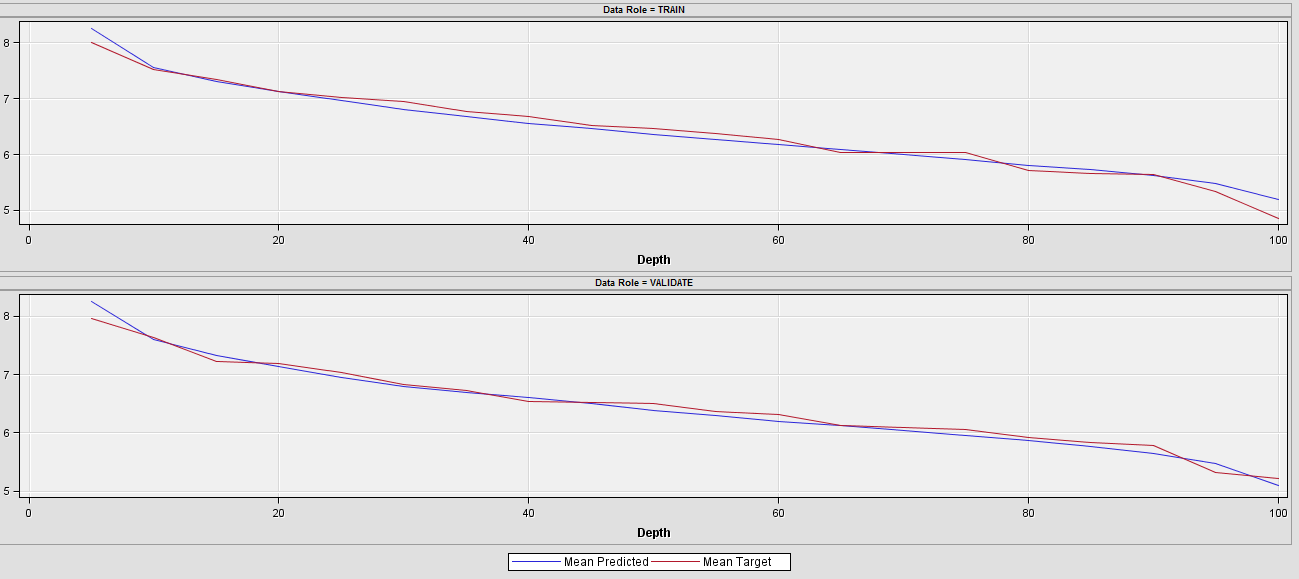
Variables selected by the model based on importance for projection are shown below:

****

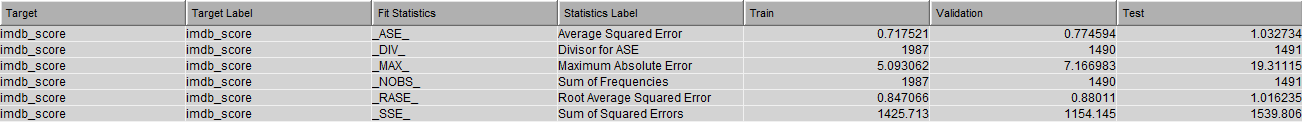
Top three variables selected by the model are below:

1. Number of voted users
2. Movie Duration
3. Number of critic reviews

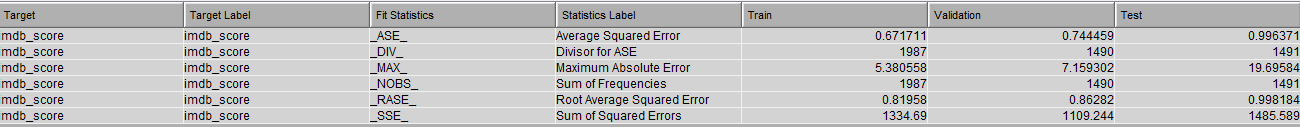
*Mean of the Predicted Value vs Target Variable*

****

*Fit Statistics before variable imputation and transformation*

****

*Fit Statistics after variable imputation and transformation*

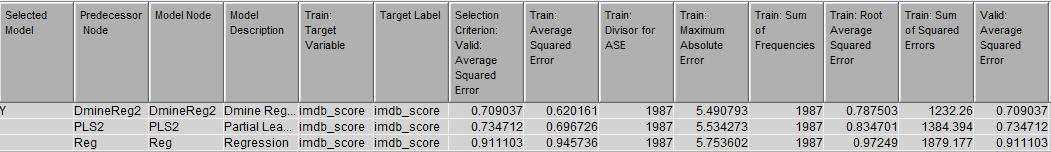
****

After performing variable imputation and transformation the root mean square error for the model reduced by 2%.

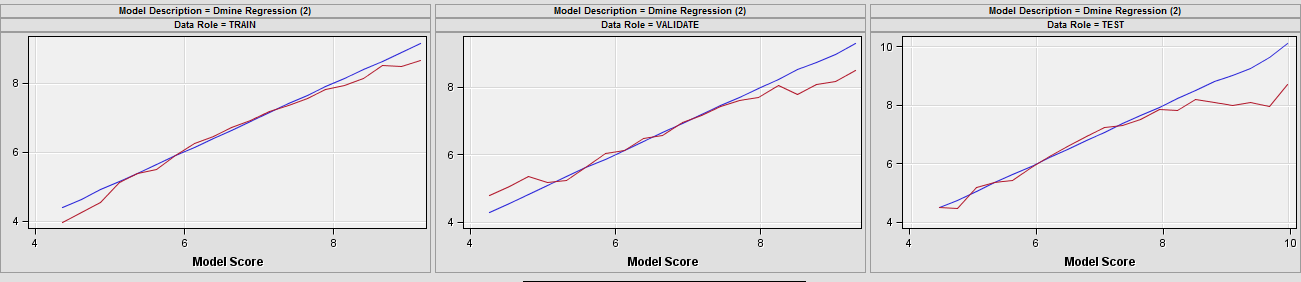
**Model Selection:**

Model selection is the crucial step in data mining tasks. Selecting a model entirely depends upon the business case at hand. For the task of predicting IMDb scores, the model with the least error rate has to be chosen. SAS Enterprise Miner has a separate node that compares the models and selects the best model which gives the least error.

After imputation, it can be seen that Dmine Regression is chosen as the best model. The average mean square error for validation data is used as the selection criteria. Dmine has the lowest average squared error for validation data.

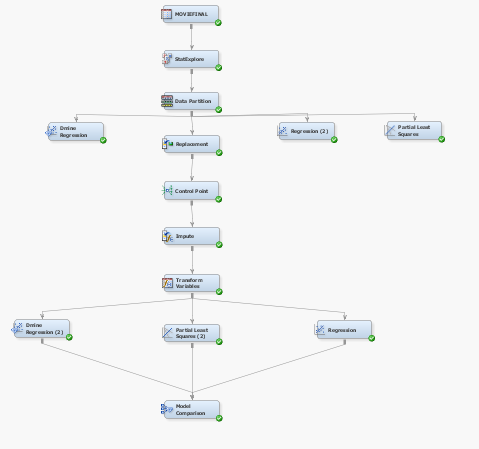


Dmine Regression is selected from SAS E-Miner. R² value for Dmine Regression model is 51% and that for Stepwise Regression is 47%. Score distribution of IMDb for Dmine Regression is shown below:

****

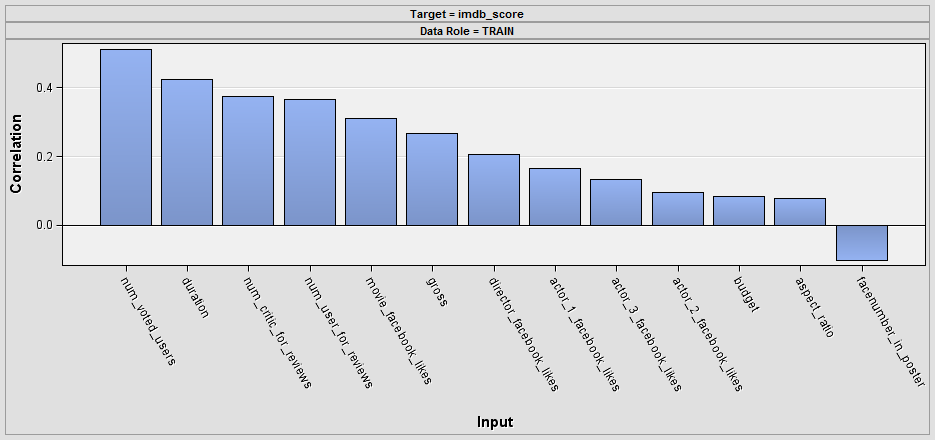
*IMDb Score Distribution: Dmine Regression*

**SAS E Miner Process Flow:**



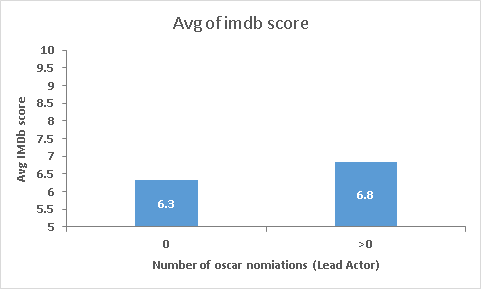
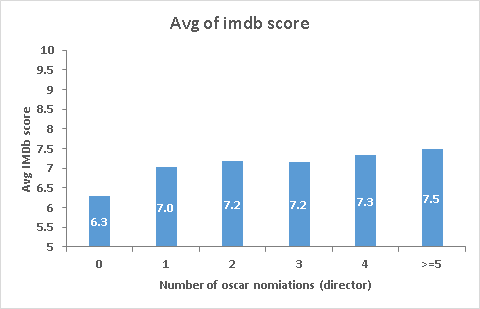
**Business Insights:**

1. ***Number of faces in posters and the IMDb scores are inversely related***. When there are too many faces in the poster, the model predicts that the score of the movies decreases. This is further shown in the below graph.

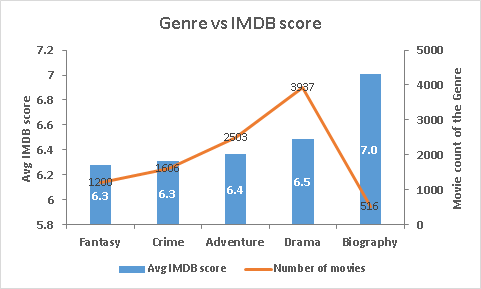


*Facenumber in Poster Vs IMDb Score*

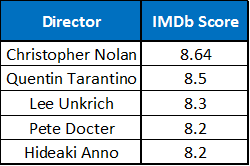
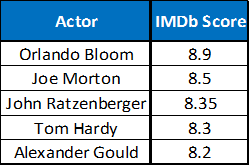
1. ***Impact of # of Oscar Nominations of Actor and Director on IMDb score.***
2. Movies with Directors that have high oscar nominations have high IMDb score.
3. The average IMDB score is high for movies in which the Lead actor is nominated for Oscar. But the difference is less significant.

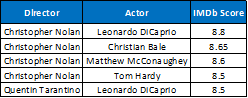


1. ***Genre Vs IMDb*** Horror movies have low IMDb scores.Biography/true stories get good IMDb ratings compared to other genres.

******

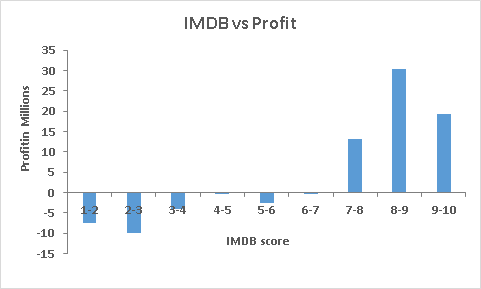
1. ***Top Actors and Directors*** Two of the main contributors to any movie are actors and directors. The top actors and directors based on their average IMDb scores are mentioned below:



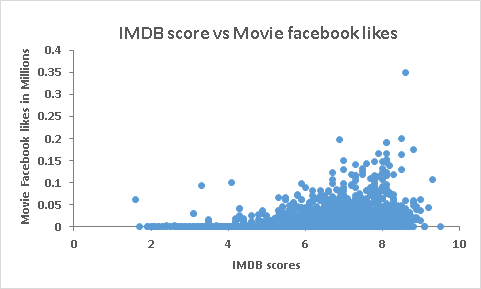


1. ***IMDB scores impact the Profit***. It is observed that movies with high IMDB scores have high

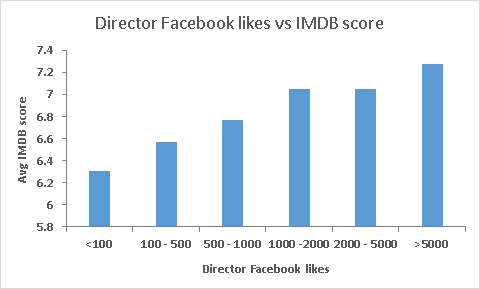
Profit.



1. ***Increased Facebook likes earns more IMDb Score.*** Movies with high IMDB scores are seen to have high popularity. Hence promoting the movies well in Social Media can earn high IMDB scores**.**



1. *High popularity of directors can lead to high IMDB scores.*



1. *Other Significant factors contributing to IMDb Score*

* Movie Budget
* Length of a movie
* Actor Facebook likes
* Movie Facebook likes
* Movie release year
* Number of voted users
* Content Rating

**Managerial implications/Conclusion:**

After incorporating the above mentioned steps, a final model was generated which will predict the IMDB scores to a certain accuracy and identify factors that influence it at the same time. The investors can leverage this information to anticipate the expected returns from the movie and plan their marketing strategy accordingly.

**References:**

The sources used for this project are mentioned below:

1. <https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset>
2. <https://blog.nycdatascience.com/student-works/machine-learning/movie-rating-prediction/>
3. <http://www.uta.edu/faculty/sawasthi/Statistics/stpls.html>
4. <https://www.casact.org/pubs/dpp/dpp08/08dpp76.pdf>
5. <http://support.sas.com/documentation/onlinedoc/miner/em43/dmine.pdf>