

**MIS 6324**

**Business Analytics with SAS**

***Prediction of success of a movie***

***Group 11 – Project Report***

**Submitted By:**

*Anurag Kanchibhotla Subhramanya*

*Mohit Manoj Shah*

*Raghavan Krishnan Veera*

*Ragini Raj*

*Sambit Sahoo*

**Table of contents**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **TITLE** | **PAGE NO.** |
| **1** | **Executive Summary** | **3** |
| **2** | **Business Objective** | **3** |
| **3** | **Dataset Description** | **4** |
| **4** | **Motivation** | **5** |
| **5** | **Data Processing** | **6** |
| **6** | **Imputation** | **7** |
| **7** | **Clustering** | **8** |
| **8** | **Data Partition** | **11** |
| **9** | **Decision Tree** | **11** |
| **10** | **Logistic Regression** | **14** |
| **11** | **Neural Network** | **19** |
| **12** | **Model Comparison** | **21** |
| **13** | **SAS E-Miner Diagram** | **23** |
| **14** | **Conclusion** | **23** |
| **15** | **References** | **24** |
| **16** | **Acknowledgement** | **24** |

**Executive Summary:**

We are now in the information era where analytics is ruling the industry. Predicting the likelihood of something happening with the data that we have will always be interesting and money making. Analytics and business intelligence plays a vital role in running every business, because it enables the use of data to predict all sorts of things that will be useful in terms of improving the business and the value that it creates for the end user through their products.

Analytics is widely spread in all the fields, right from analyzing the power consumption to artificial intelligence. In the former, one can analyze the power consumption by a family or household over a period to build a specific plan that will best suit that house resulting in the nominal payable amount, eventually gaining their support towards one electricity provider. Whereas in the latter, analytics can be used to in the field of artificial intelligence to train the artificial system to learn about what could possibly happen, in turn making them decide by themselves about what can possibly happen in the future? Likewise, there are many applications where analytics can be used. One field where analytics can be widely used is movie industry, which produces a lot of data that can be used in analyzing a database. Making a movie isn’t a simple, where only one or two person’s works will bring in a huge revolution. It involves many people and many small, yet, powerful dimensions that make it huge success.

Predicting the success of a movie will help the movie makers save a huge amount of time and money. For example, assuming the prediction is correct over 95% then the movie makers can reduce spending the time in making promotions and with the feedback that they receive, they can give what people expect, thereby making the movie a big success the next time.

How do we determine the success of a movie before it is released? What parameters do you think are responsible in determining the success of a movie? Do you think the directors and the actors are solely responsible for success of a movie? Movie has a lot of dimensions that could possibly be used to determine the success. Selecting the right parameters is a difficult task and they must be logical as well. Our project is all about finding the right parameters that contribute to the success of a movie and building a model that will used in predicting the success of a movie with the valuable inputs that are logical and have good inclination towards predicting the success of a movie. Can we assure that the success of the movie can be predicted accurately with a mathematical model, because most of the time but not with what people think about it and the everlasting influence that the movie has got in people’s mind? Not all the predictions can be accurate but there can be some margin of error, but this doesn’t make the model inefficient. This report contains the details, explanations and conclusions of all the predicting models.

**Business Objective:**

* To determine the success of a movie before it gets released.
* To determine the cutoff point for a movie to be good.
* To determine the parameters responsible for the success prediction.
* To determine the best model that has maximum predicting accuracy.

**Dataset:** IMDB Movie ratings Data–This data is about the rating for different types of movies of various actors, directors and genres.

**Dataset Source:** The data-set is a second-hand data downloaded from Kaggle,

https://www.kaggle.com/zcbmxvnyico/notebook6635037081

The dataset has attributes related to number of voted users, number of critic for reviews, gross, budget, etc.

**Description of the dataset**:

The IMDB movie data set consists of 28 columns and 5040 records of data. It includes columns like color, top 3 actors, name of the director, number of critics for review, duration, facebook likes of the director, likes of the top three actors, gross, budget, genre, name of the actors, movie title, number of the voted users, total facebook likes of the cast, total faces in the poster, plot keywords, number of user for review, language, country, title year, IMDB score, movie facebook likes.

**Color**: In our data we have both black and white and color movies.

**num\_critic\_for\_reviews:** This attribute specifies the total number of critics who have reviewed the movie.

**Duration:** This attribute specifies the duration of each of the movies.

**actor\_1\_facebook\_likes**

**actor\_2\_facebook\_likes**

**actor\_3\_facebook\_likes**

**director\_facebook\_likes**

These fields indicate the number of facebook like that each of the actors and directors have got on their facebook page.

**movie\_facebook\_likes:** Total number of likes the movie has got in its facebook page.

**actor\_1\_name, actor\_2\_name, actor\_3\_name, director\_name:** Actors and directors in the movie have considerable influence on IMDB ratings.

**Gross:** It is the total monetary value that the movie has earned after the release. This includes revenue from the national, international market and satellite rights.

**Genres:** It indicates the type of movies. For example: thriller, action, comedy, love story, fictional etc.

**Language:** This column contains the languages that the movie has been released in.

**title\_year:** The year in which the movie was released. Our dataset contains movie from 1900’s until 2016.

**num\_voted\_users:** The number of users who have rated the movie, or the total number of voted that the movie has scored.

**facenumber\_in\_poster:** It is the number of faces that are present in the poster of the movie.

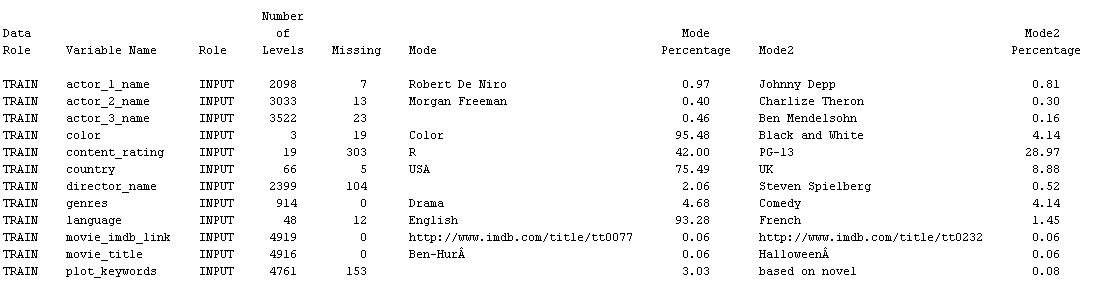
**plot\_keywords:** This attribute contains the keywords about the movie. It conveys what the movie is about with a short number of words.

**Motivation:**

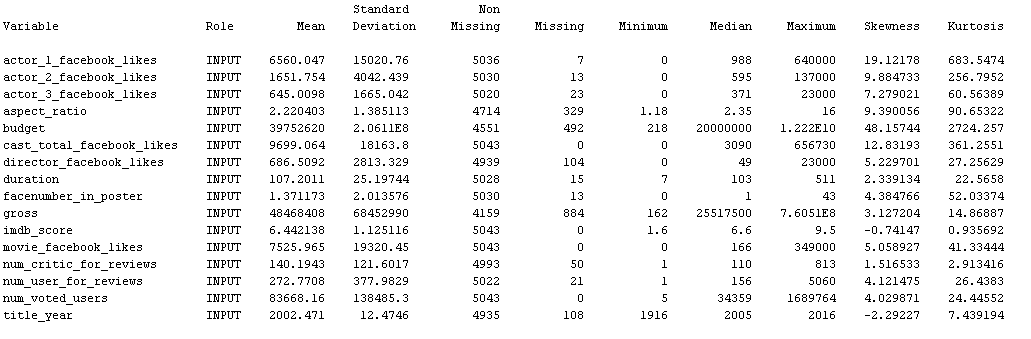
Entertainment have become a large-scale industry comprising of movies, tv-series, advertisements and prints. It also includes smaller segments like radio, music, animation, gaming and visual effects (VFX) and Internet advertising. The U.S. entertainment and media market generated **$479.23 billion** in 2012, representing 29.2 percent of the worldwide revenue of nearly $1.639 trillion. In 2017, the U.S. is expected to account for **$632.09 billion**, or 29.4 percent of the worldwide total of more than **$2.152 trillion**, according to the report [1]. Even in the film industry category, where the hype around China is loudest, the North American market -- which includes the U.S., Canada and Mexico -- will dominate through 2017. Globally, PwC expects a milestone in the filmed-entertainment sector in 2016, when it busts through the $100 billion barrier for the first time in history. “The U.S. remains the largest, most valuable territory in the world for all filmed entertainment,” PwC concludes [1]. The professional services firm says the sector will grow 3.4 percent annually in the U.S. from $31.04 billion in 2013 to $36.35 billion by 2017, while globally, film revenue will grow at a rate of 3.6 percent per year to $106.01 billion by 2017 [1]. The above shows how much the film industry is crucial to the entertainment revenue of a country or continent. For making a huge revenue it is crucial to first predict the feasibility of a film whether it will be benefiting or will be a loss. The advancement in the field of analytics can help us do this. Hence, we will be trying to predict the feasibility of a film before releasing with the help of analytics concepts like clustering for deciding the threshold value from legacy data, decision tree for making classification, linear regression analysis for making the model for prediction and neural network for checking how the BIAS factor comes into play as it involves a human factor.

**Data Preprocessing**:

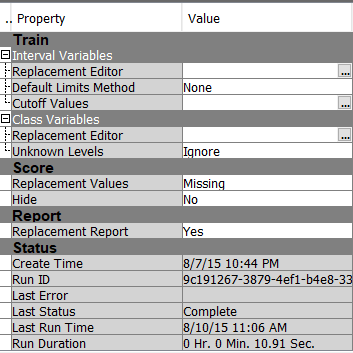
We loaded the entire dataset into the SAS enterprise miner and changed the role of all the variables as input and ran the stat explorer node over it to get the missing value of all the class variables and the interval variables.



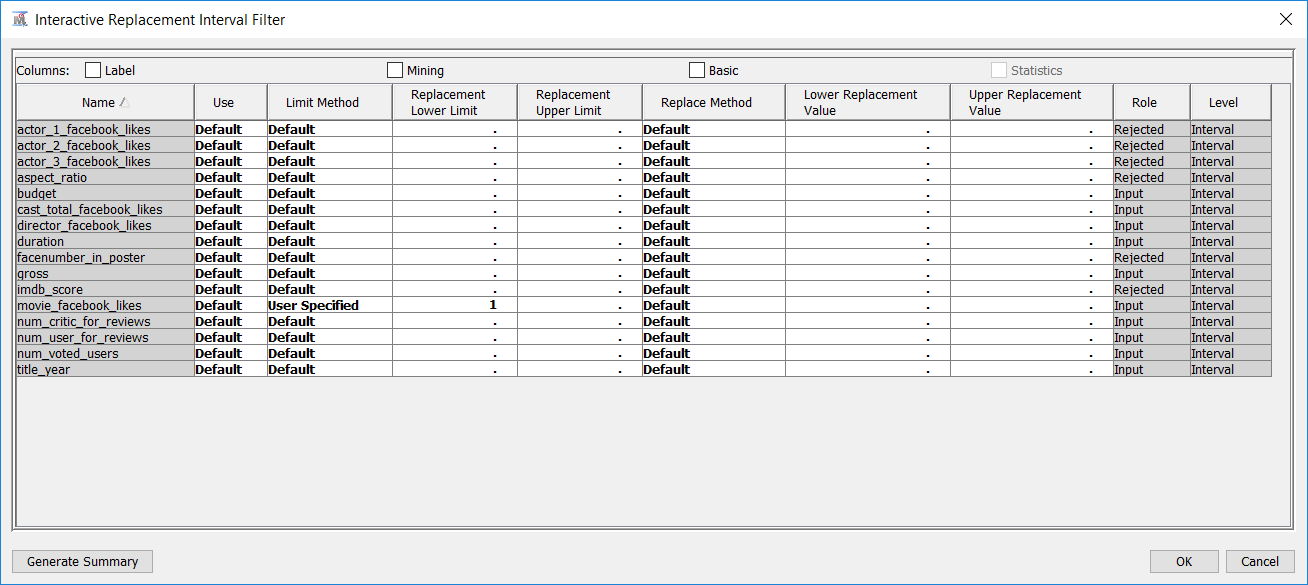
For the Class variables we replaced the blank variables with mode values from each of the columns in the excel file itself.



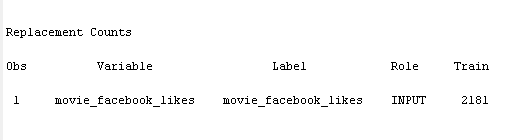
For the interval variables we used the replacement node and the impute node to eliminate the zero and blank values in SAS enterprise miner.



The replacement value was set to “Missing” and the replacement editor was opened to specify the columns which had 0 values and that are to be replaced with the “Missing” values.



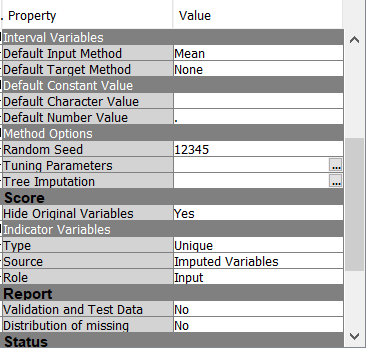
In our dataset, only the movie\_facebook\_likes contained ‘0’ values so the condition was set only for that column and all the cells with the value less than the 1 were replaced with the missing values.

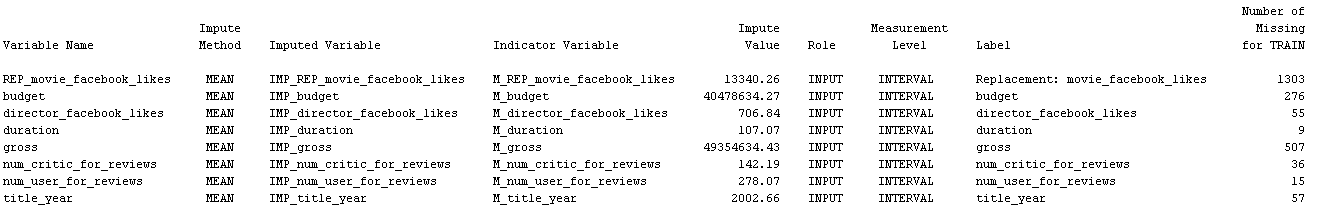


This is the output of the “Replacement” node. Totally 2181 ‘0’ values were replaced with the “Missing” values.

**Imputation:**

The imputation node replaces the “Missing” values with the Mean, median or mode of the entire column of data. In our project we have used the impute node to impute the missing values with the mean of the column.

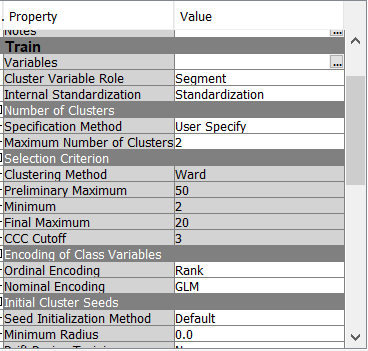


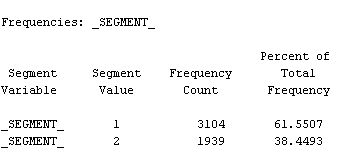


In the impute node we have specified the type of the indicator variables are Unique and the role as “Input”, which means that there is another copy of the imputed variable which is used as the input in the model used. The new imputed variable is renamed with a prefix of IMP.

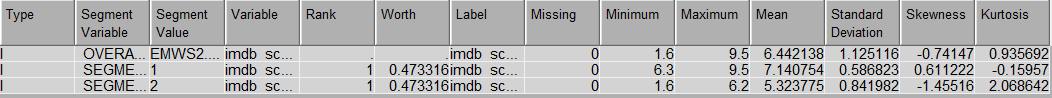
**CLUSTERING:**

Clustering, the first data mining technique that was carried out in our project. We did clustering based on the rating column. Our data set didn’t contain the ID column, so we added another column to our data set and named it ID. Then clustering is done on the IMDB rating column, which didn’t not contain any missing or ZERO values. So, no replacement or imputation is needed for clustering to be carried out. The number of clusters is made 2 because for the project we need to predict the success of a movie, whether the movie is good or bad or it is successful or not.

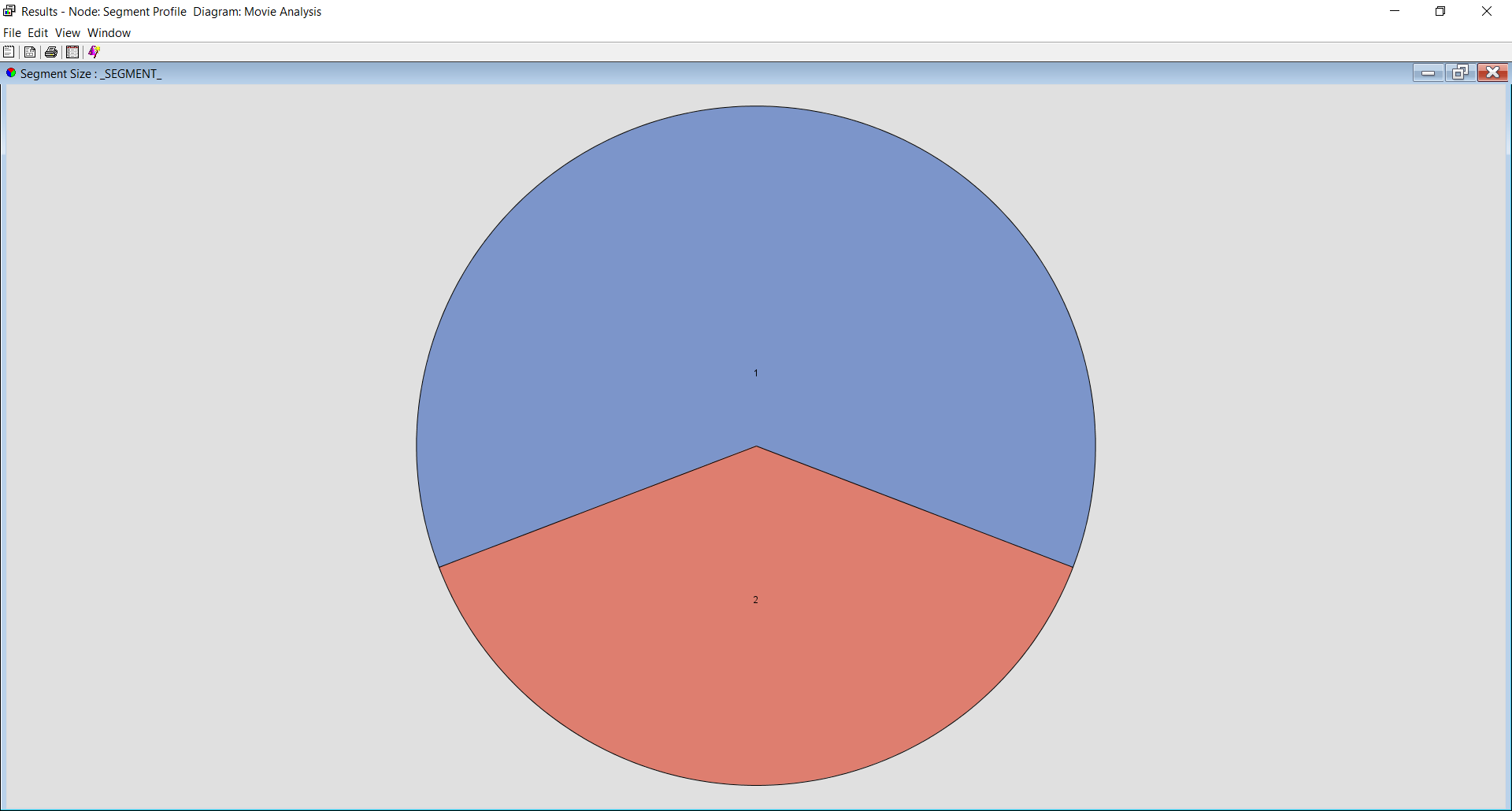




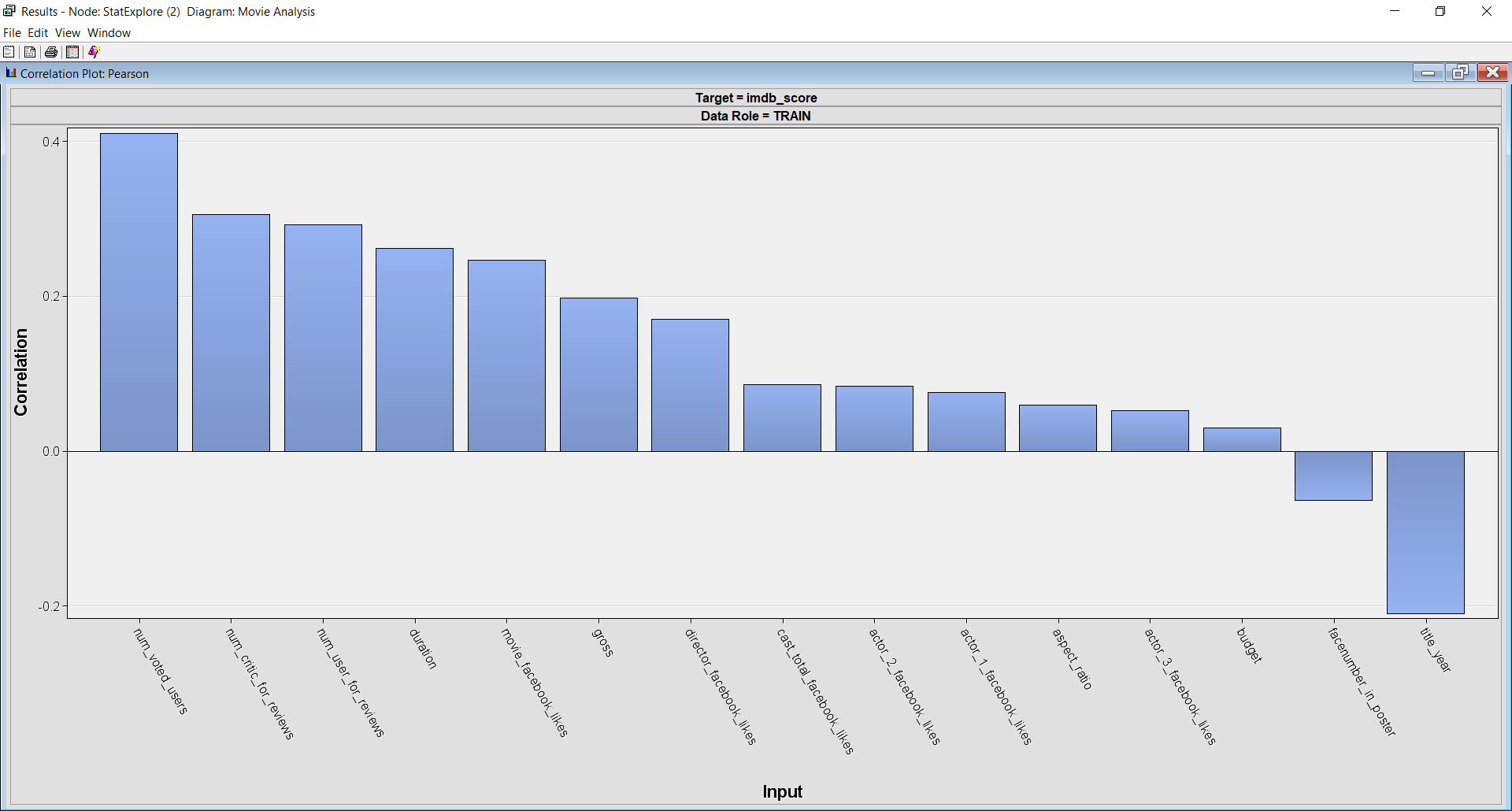
The result of the clustering is specified above, there were totally 2 clusters. One cluster contains approximately 61% of the data and the other segment contains approximately 38% of the data.



The first segment contains 61% of the data and the minimum rating in that cluster is 6.3, which means that 6.3 is the cutoff point and movies which has the rating above this value can be classifies and good movie and can be deemed successful. The movies in the second cluster makes about 38% of the total data and the maximum rating in this cluster is 6.2, which means the movies having the rating as 6.2 or below can be classified as bad movies and can be deemed unsuccessful. The variable that has been created in our data set for this purpose is Movie\_Categorization. With this, the supervised learning can be carried out.

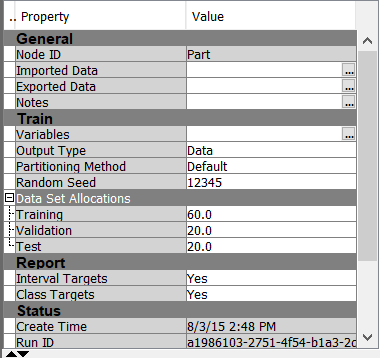


Variable selection for the input:



The above plot is the correlation plot, which shows the correlation of each of the input variables to the target variable (IMDB\_Rating). Num\_voted\_users has the highest correlation, with correlation value higher than 0.4. Title\_year has negative correlation with the target variable with 0.2 value. Among these variables not all the variable can be used in the model. Some variables are neglected. Actor\_1\_Facebook\_likes, Actor\_2\_Facebook\_likes, Actor\_3\_Facebook\_likes, director\_facebook\_likes are neglected because cast\_total\_facebook\_likes has been included which contains actor, director and many other people. Also, another reason why the director facebook likes is rejected is that less likes for the director doesn’t necessarily mean he is a bad director and has movie will not be good. There are exceptional directors whose work has been amazing, yet their likes are less.

**Data Partition:**

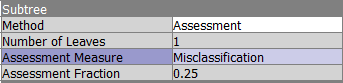


We divided the entire data set in to three parts namely TRANING, VALIDATION and TEST. The training data set consists of 60% of the data, validation data has 20% and test data has 20% of the data in it. Since our model is used for prediction purposes we need to train the data a lot to have a very good validation and test results, so we have assigned 60% of our data to training node. And this node is connected to various predictive models like decision tree, regression and neural network.

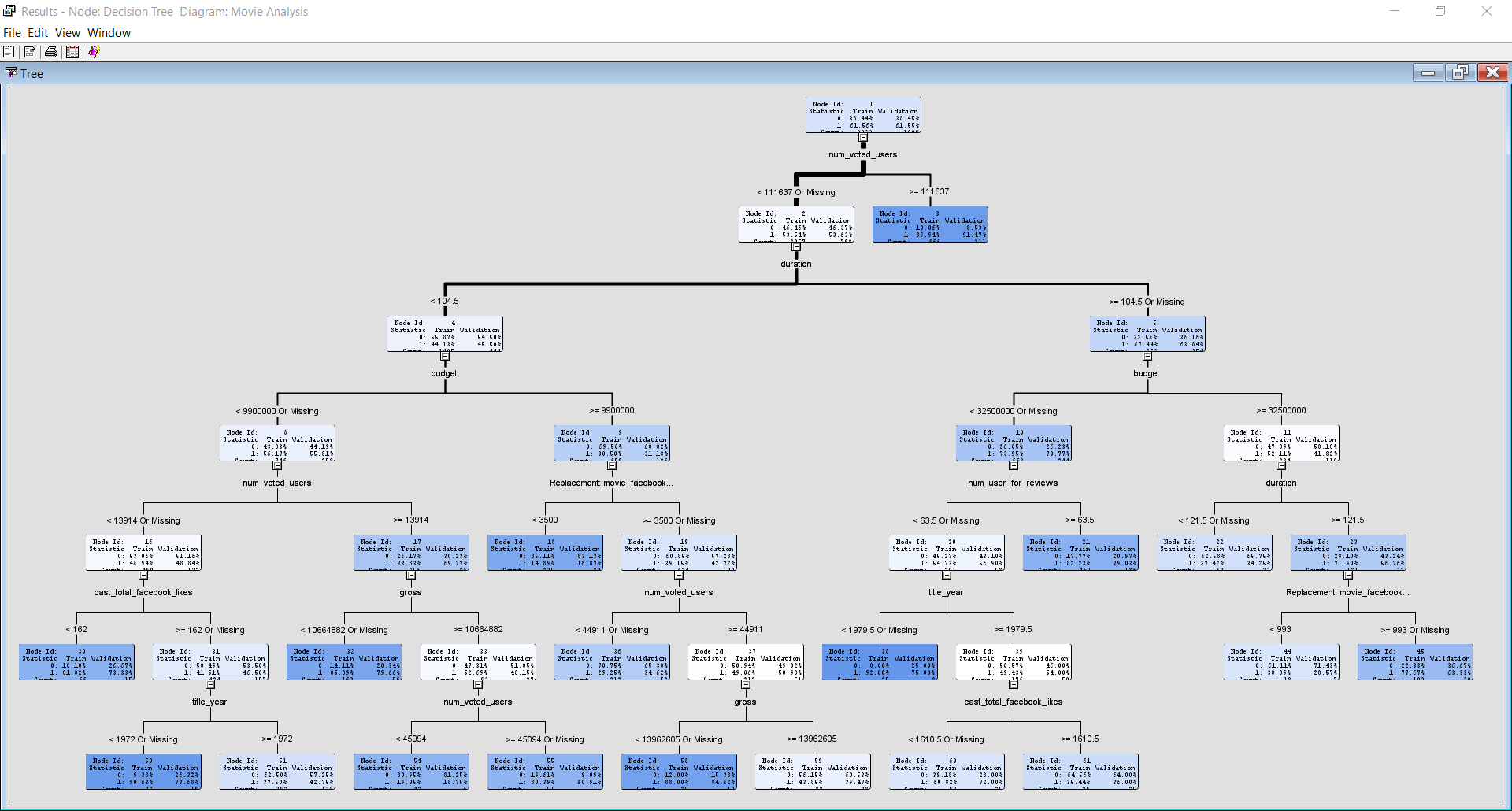
**Decision tree:**

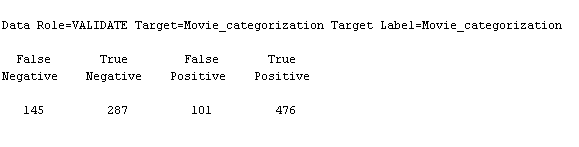
Decision tree is one of the models that was used in our project, the target variable in our project is binary, either successful or unsuccessful. Hence, decision tree is considered, as it is one of the most efficient models for predicting binary values.

**Optimal Decision Tree:**



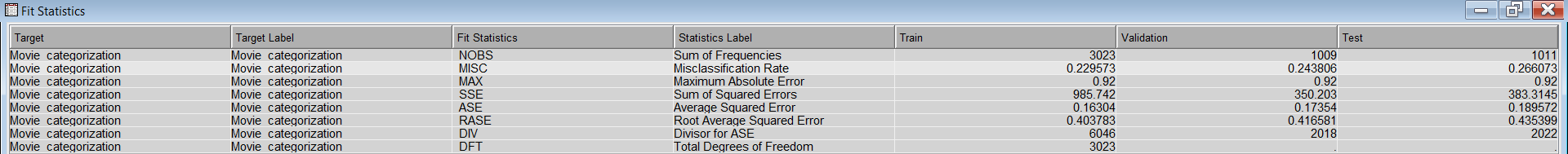
The Assessment measure of the decision tree is kept as misclassification, as our target variable is binary the assessment measure is changed to misclassification.



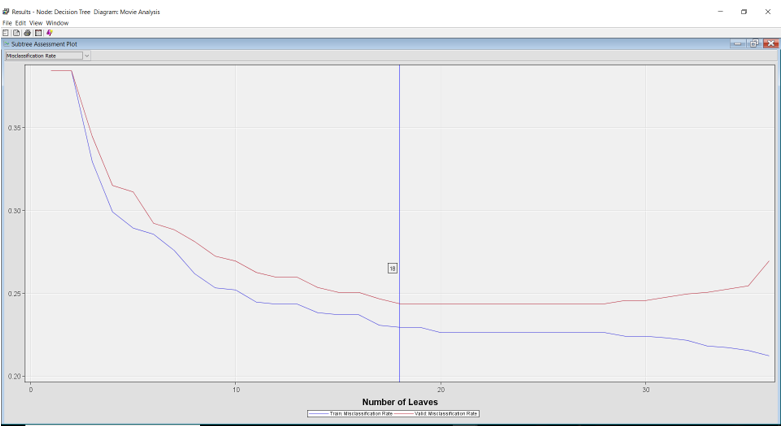


The above table is Event classification table, it shows the number of predictions that falls in each of the category. True- Positive and False-Negative are two of the categories that we would feel happy about, as these are the two categories in which the value of the target variable has been correctly predicted. Whereas, the True-Negative and False-Positive are the two categories which contains the wrongly predicted values.

The decision tree can be interpreted using the misclassification rate, the misclassification rate of the validation data is 0.243806, which means approximately 75% of the data has been correctly classified.



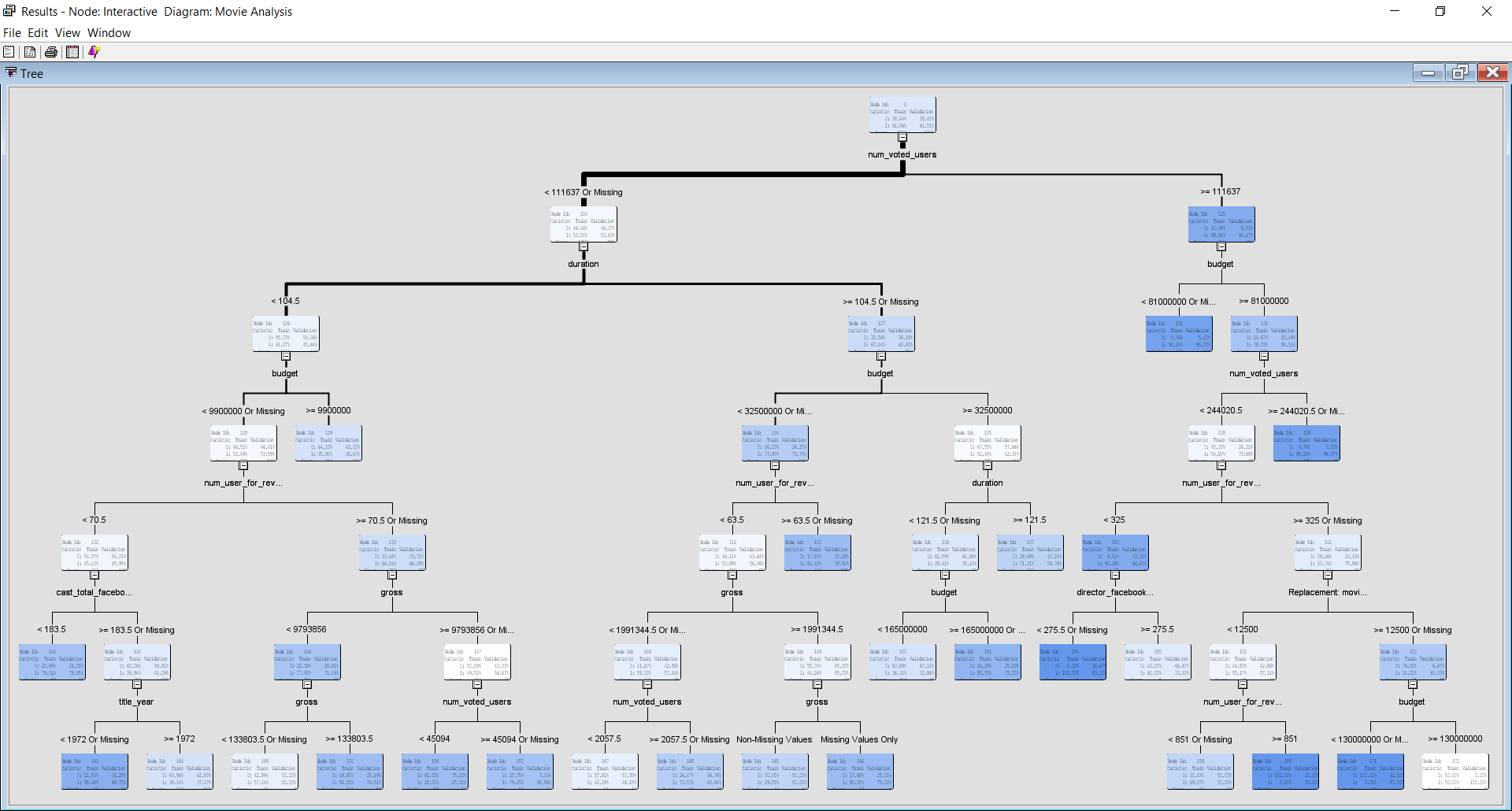
**Sub tree assessment plot:**



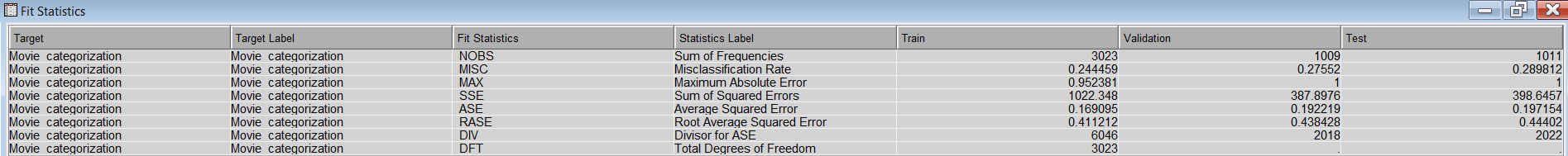
From the subtree plot we conclude that the optimal number of leaves is 18, which means that if we split the node above 18 leaves there are high chances that overfitting problem will occur. Also, from the graph we can see that the difference in the misclassification rate for validation and training data increases beyond the optimal number of leaves. At this optimal point, the misclassification rate for the validation data is 0.2468 and for training data it is 0.2296, from this we can infer that the deviation of the misclassification rate of validation data from training data is approx. 0.02.

**Interactive Decision Tree:**

In interactive decision tree the root node is split using the logworth value i.e. log(P). So, for each of the split the variable with the highest logworth value is chosen to make the split.



The decision tree is built by splitting based on the logworth value. In our decision tree the first split was based on the variable “num\_voted\_users” which has the highest logworth value. This variable based on which the first split is done, is same as the optimal decision tree.



The misclassification rate for interactive decision tree is found to be 0.27552, which is relatively higher than the optimal decision tree’s misclassification rate.

**Logistic Regression:**

The second model in our project is logistic regression. Logistic regression is used when the target variable is categorical. There are three ways in which the regression can be carried out, namely, stepwise, forward and backward. The mathematical model for the regression can be show in an equation which consists of dependent variable, independent variables and the intercept.

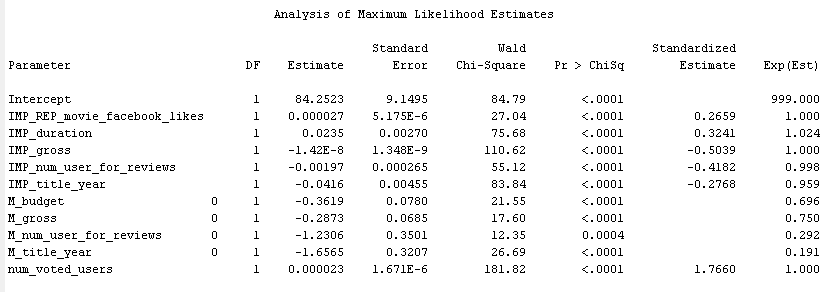
Ln(p/1-p) = a+Bx+ E

Where, p is the probability of the event Y occurring

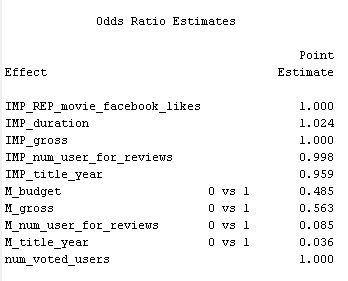
p/1-p is the odds

Ln(p/1-p) is the natural logarithm of the odds or logit

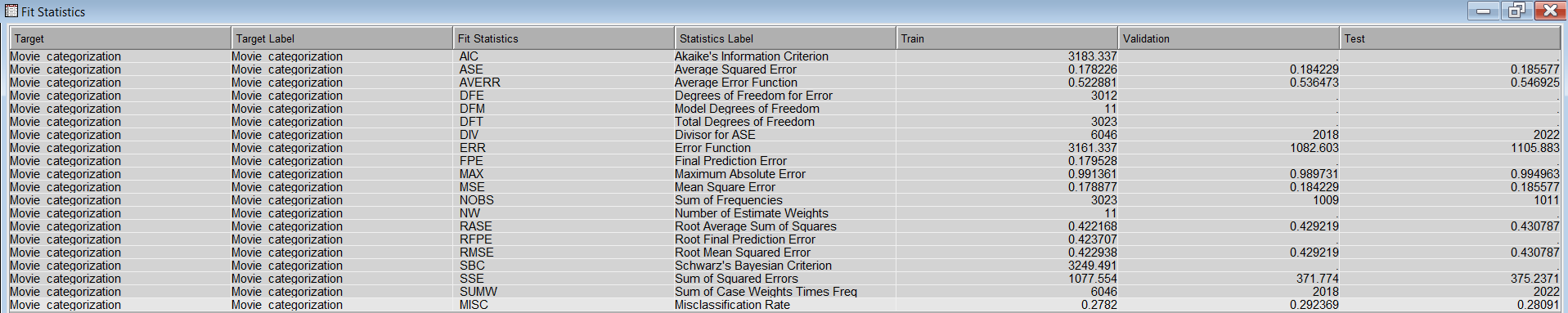
**Stepwise:**



The output of the stepwise regression is interpreted through the likelihood estimates. It shows the likelihood of the event (Y) to happen, based on the input values. The estimates are the value of the coefficients based on which the regression equation can be formed.

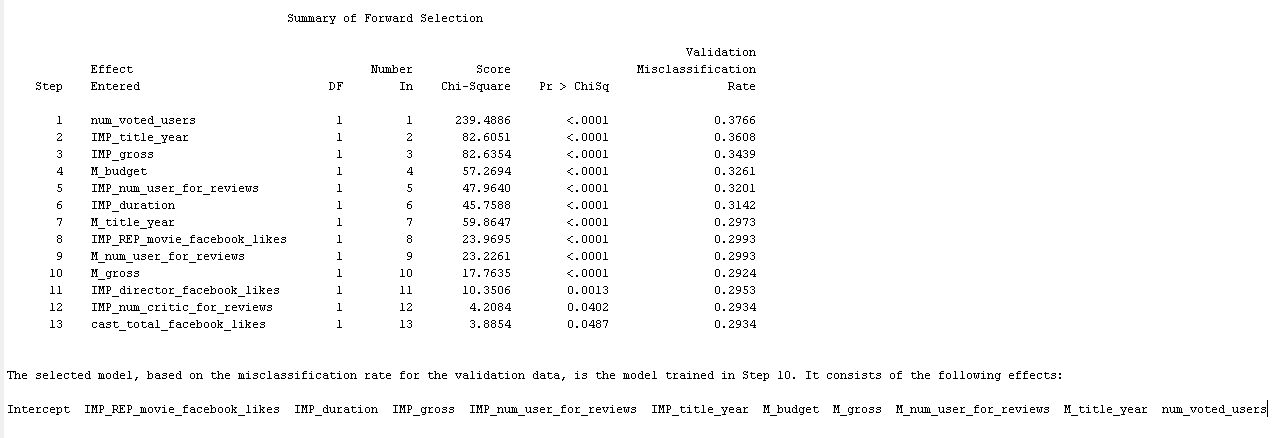


The odds ratio explains the impact on the target variable based on the variation in the input parameters. For instance, when all the other variables are considered to be constant and there is a unit increase in the duration of the movie, the odds the movie will be successful is 2.4%.



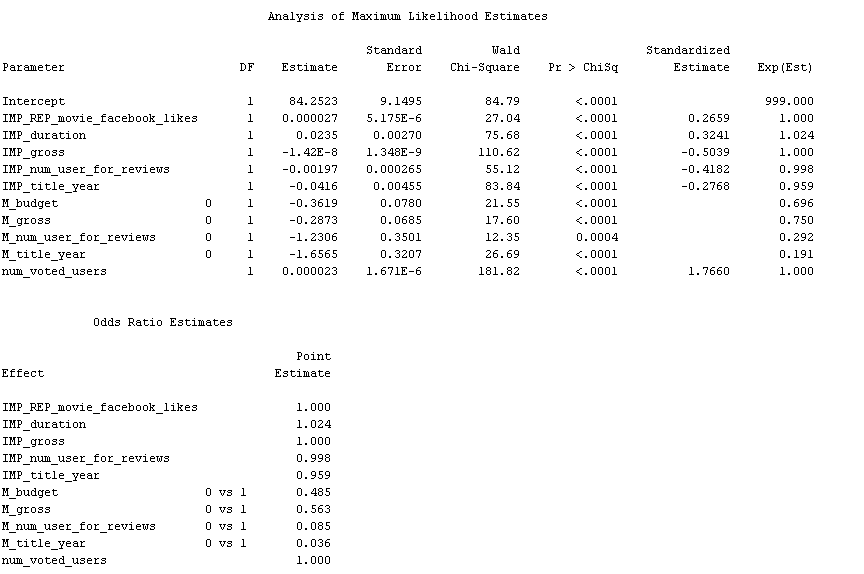
From the fit statistics window, the misclassification rate of the stepwise regression model is 0.2923.

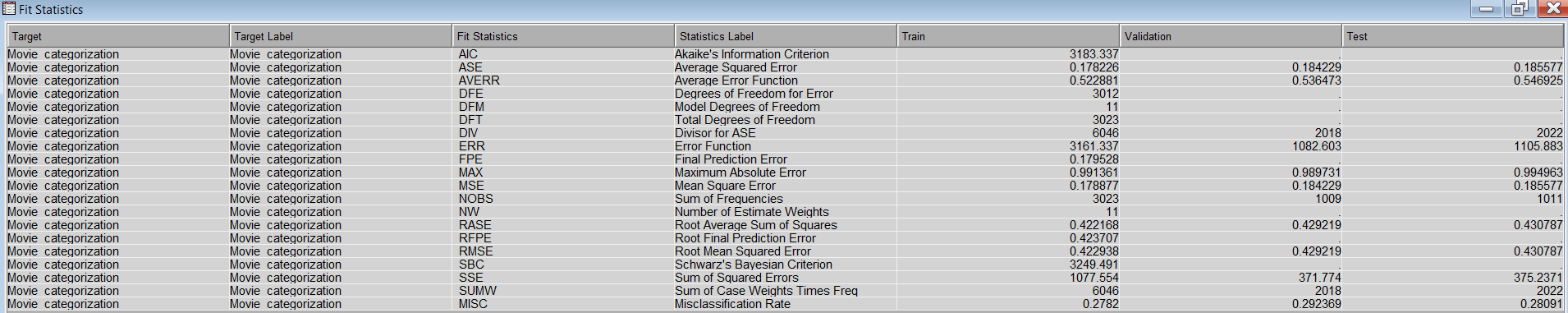
**Forward Logistic Regression:**



From the output window we can conclude that the forward regression model is trained at the 10th step with the following effects,

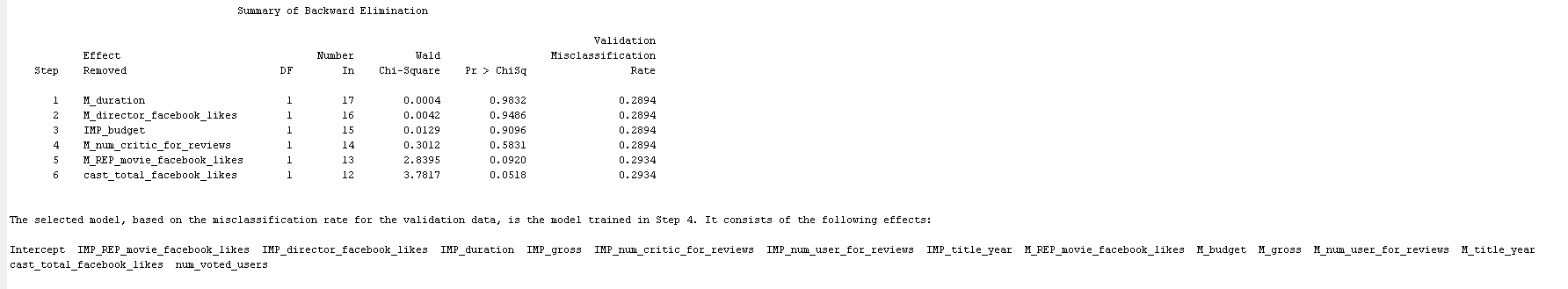
Intercept, IMP\_REP\_movie\_facebook\_likes, IMP\_duration, IMP\_gross, IMP\_num\_user\_for\_reviews, IMP\_title\_year, M\_budget, M\_gross, M\_num\_user\_for\_reviews, M\_title\_year num\_voted\_users





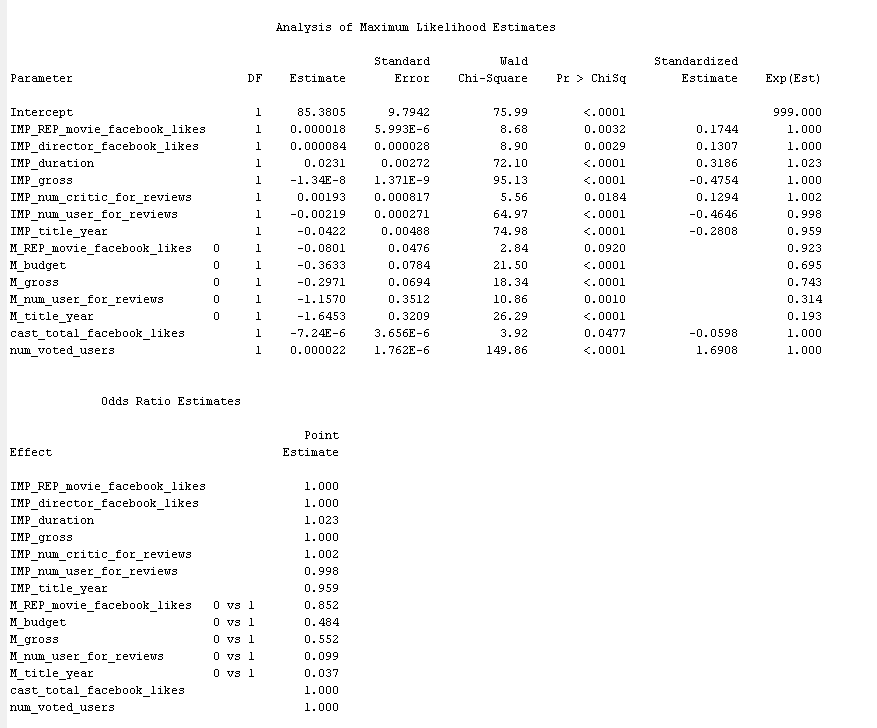
From the fit statistics window, the misclassification rate of the forward regression model is 0.2923.

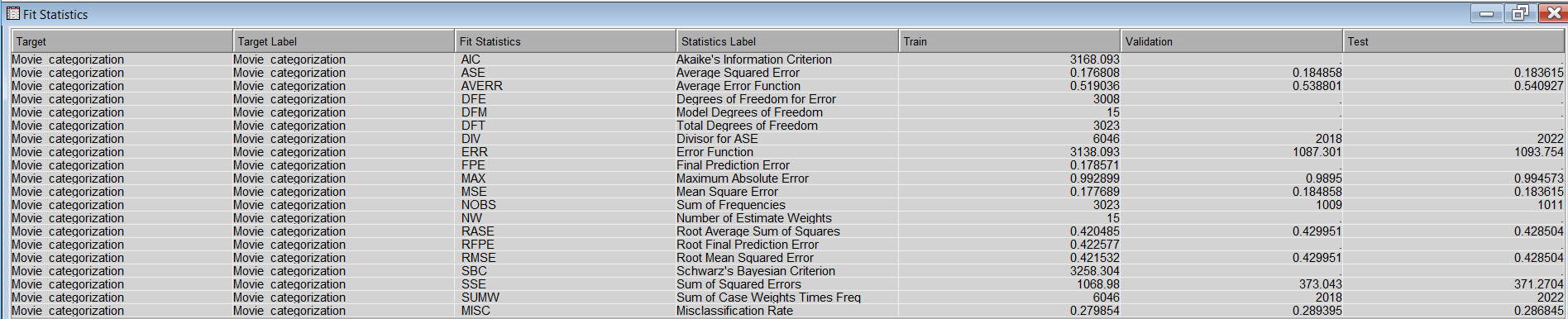
**Backward Logistic Regression:**



From the output window we can conclude that the backward regression model is trained at the 4th step with the following effects,

Intercept,IMP\_REP\_movie\_facebook\_likes,IMP\_director\_facebook\_likes,IMP\_duration, IMP\_gross,IMP\_num\_critic\_for\_reviews,IMP\_num\_user\_for\_reviews,IMP\_title\_year, M\_REP\_movie\_facebook\_likes,M\_budget,M\_gross,M\_num\_user\_for\_reviews,M\_title\_year,cast\_total\_facebook\_likes,num\_voted\_users

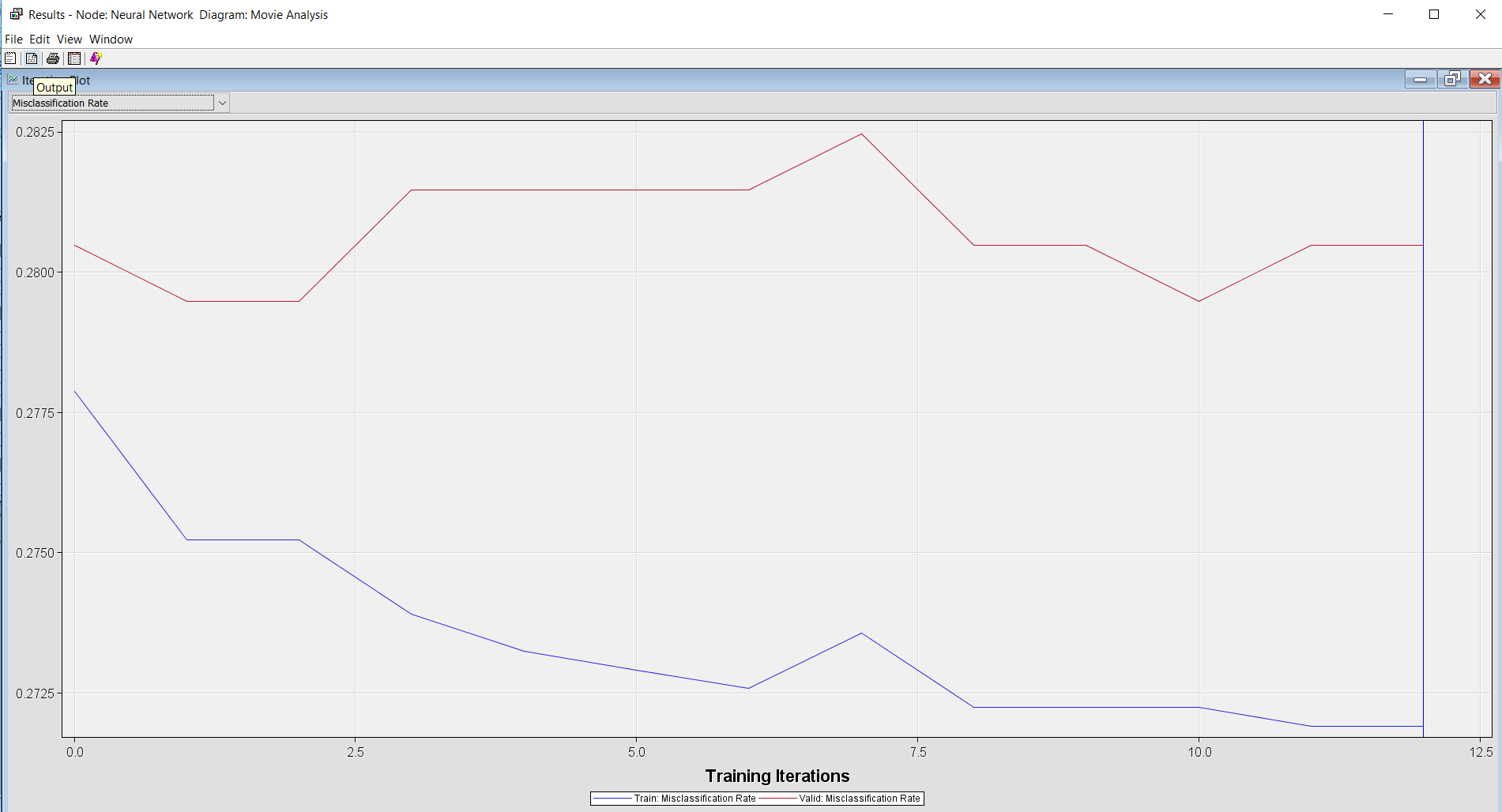




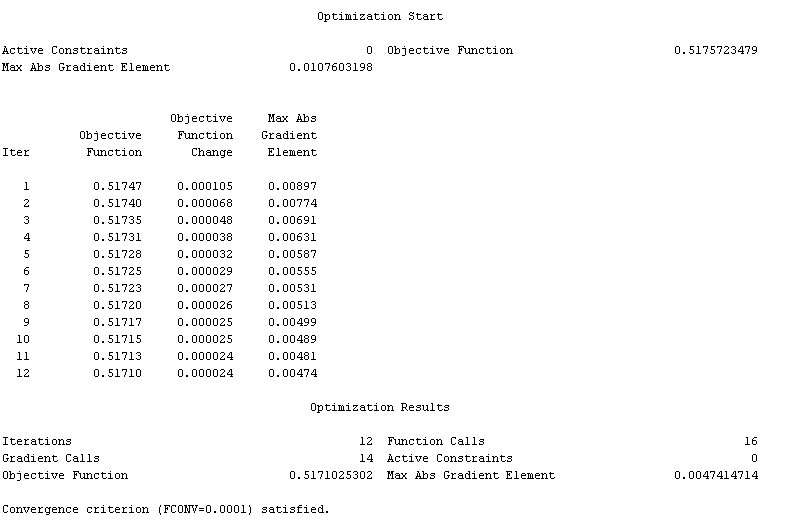
From the fit statistics window, the misclassification rate of the backward regression model is 0.2893.

**Neural Network:**

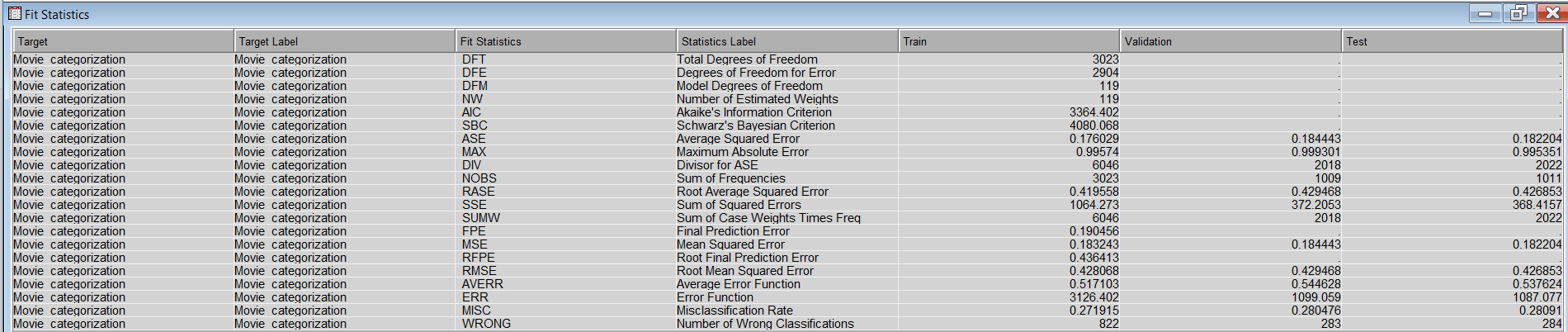
So far, we have seen two prediction models like decision tree and logistic regression, now we are using another predictive model to determine the success of the movie which is neural network. Neural network generally functions on binary and continuous data. Since, our target being binary variable neural network can be implemented. In general, neural network is very tough to understand but we consider this to have a stronger predictive model for rest of the model comparison. There are two main parameters considered in neural network firstly we must define the neural connection and in second phase we may need to consider the number of iterations involved so that the model is converged at the end. We ran the neural network and inferred that the model took 12 iterations to converge.

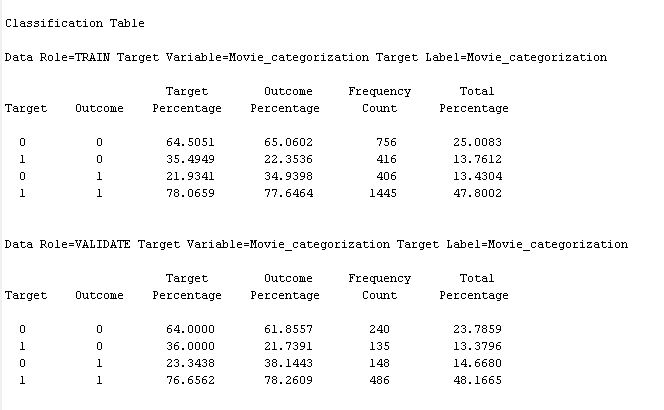


We can interpret the output of the neural network using misclassification rate and also with event classification table.



From the above table we can conclude that the model has taken 12 iterations to converge and with a misclassification rate of 0.2804

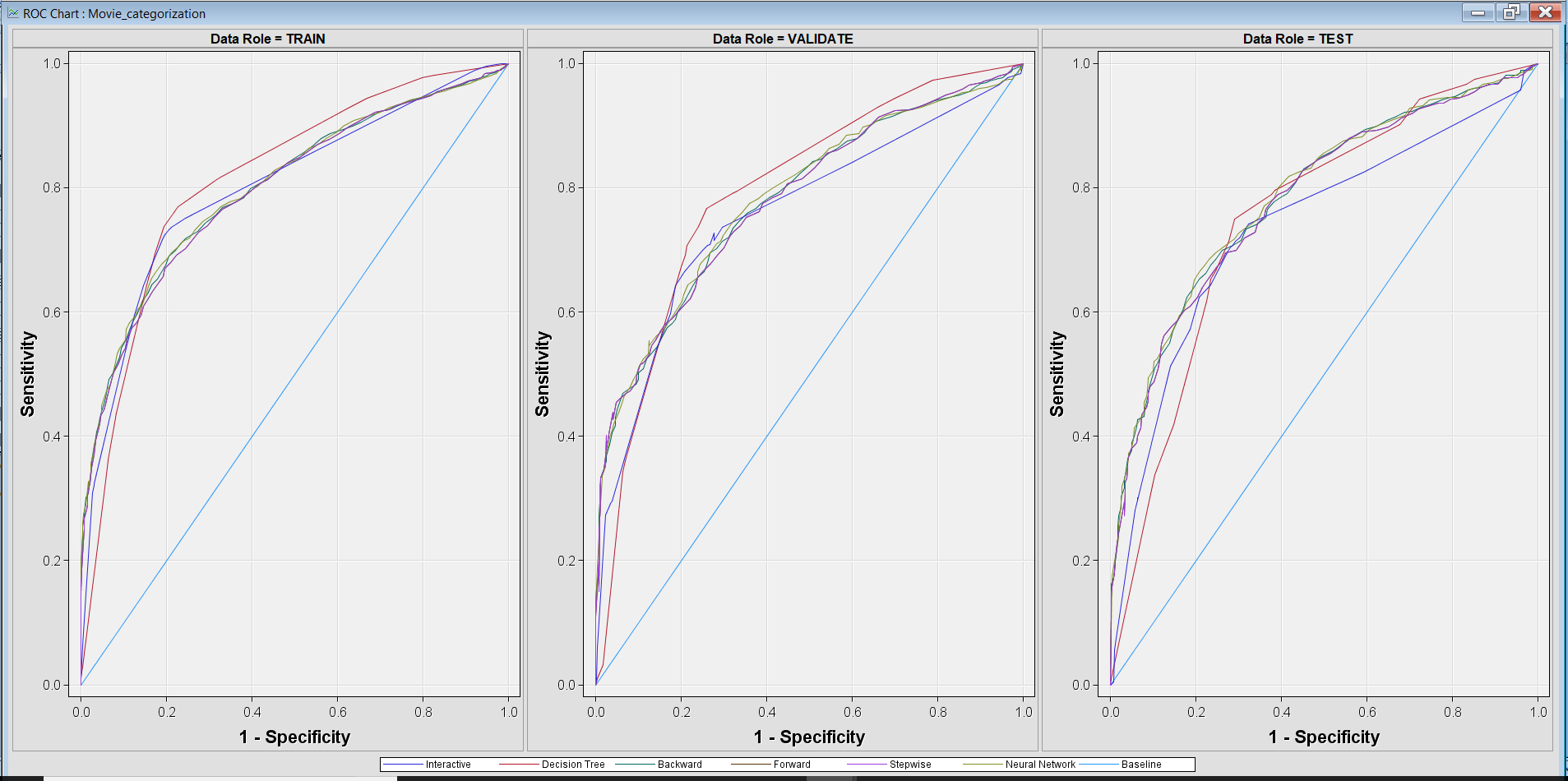


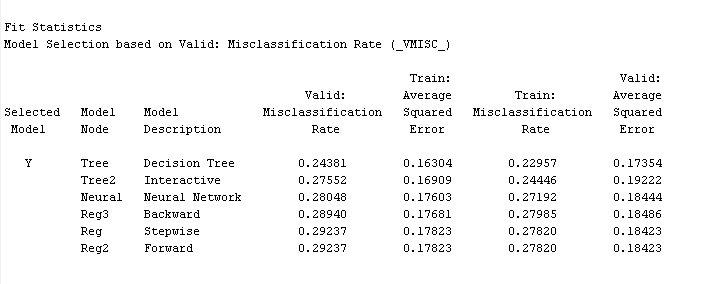


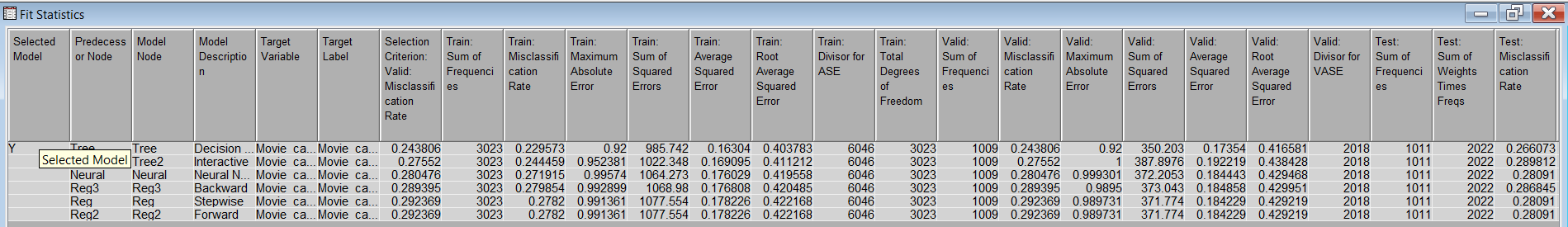
From the classification table we can infer that prediction of false negative data is 23.7% and true positive is 48.16% which is in total 71.8% of the total data predicted in correct.

**Model Comparison:**

It the human tendency to know which is the best model, with SAS enterprise miner we can compare different models using “Model comparison” node. Connect all the output of the predictive models as an input to the model comparison node and make the Selection Statistic to “Misclassification Rate” to compare the best among them. The output of model comparison is as shown below.

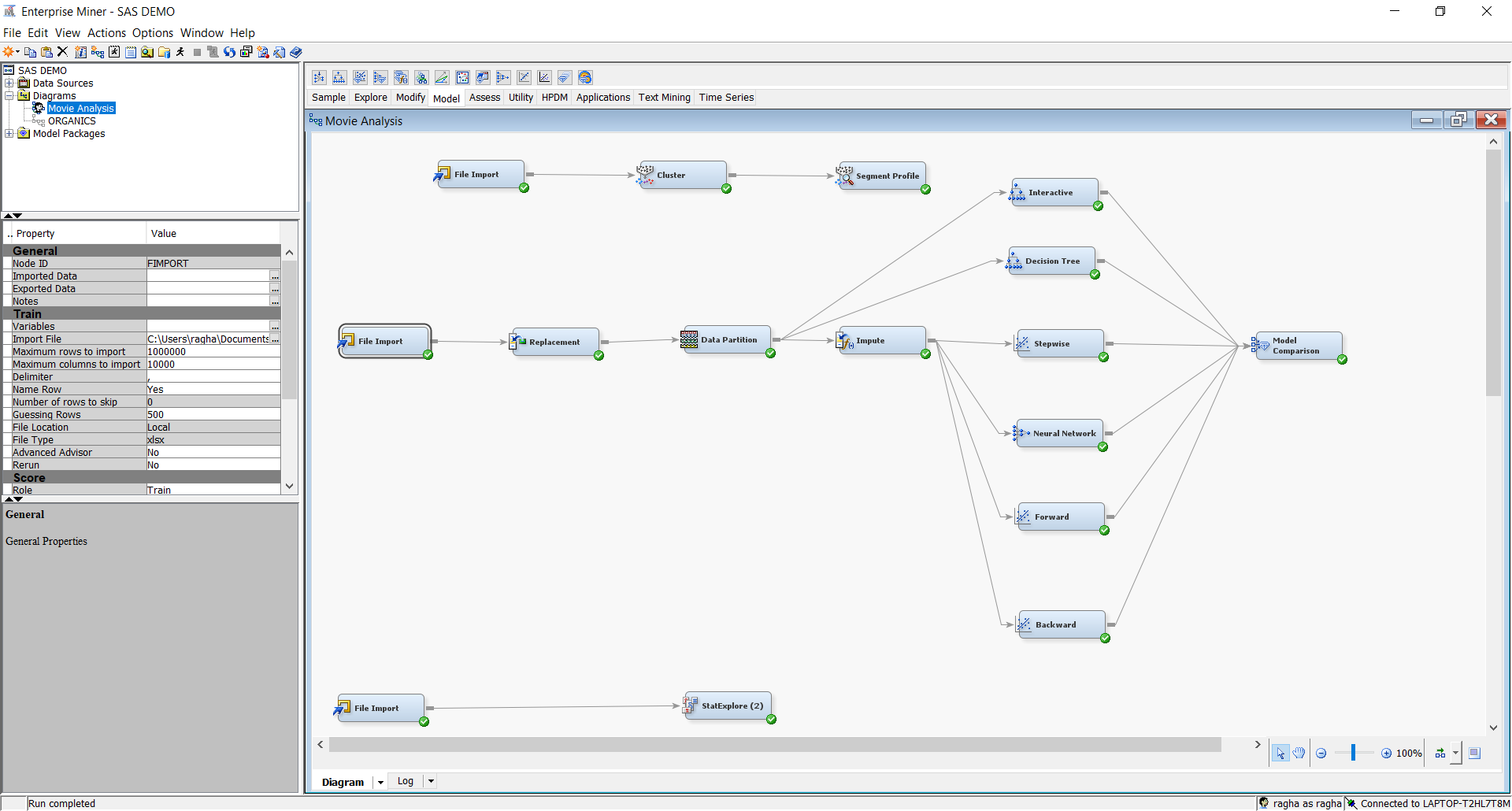






The misclassification rate of Decision tree with 0.24381 is the best when compared to all.

**SAS E-MINER DIAGRAM:**



**Conclusion:**

There are different models that can be considered in predicting the success of a movie based on the parameters that has the highest correlation with the target variable. The best model can be picked based on the efficiency to predict the maximum correct values.

The director or the team can focus more on specific parameters which are deemed to produce a positive effect on the movie i.e. they can focus more of the parameters that will push the success of the movie higher in the ladder.

These models will be useful for the production houses to assess their movie’s success just before it gets released. They can also be used by the Ad agencies and the short film producing companies to estimate the success of their commercials using the same model.

**References:**

1. Applied Analytics using SAS Enterprise Miner, ISBN: 978-1-61290-139-8.
2. Video and class lectures of Professor Zhe Zhang.
3. Video lectures on analytics concepts from “Analytical University”,

<https://www.youtube.com/channel/UC2XO4HDxzfMOZIV1l795g1Q>

1. <http://www.billboard.com/biz/articles/news/global/1565728/study-global-entertainment-industry-poised-to-top-2-trillion-in>

**Acknowledgement:**

We would like to thank Professor Zhe Zhang on his valuable inputs and guidance throughout the project.