Tan Yin Yen (WQD180108)

Data Mining Project: Milestone 4 (Interpretation of Data)

Analysis Goal

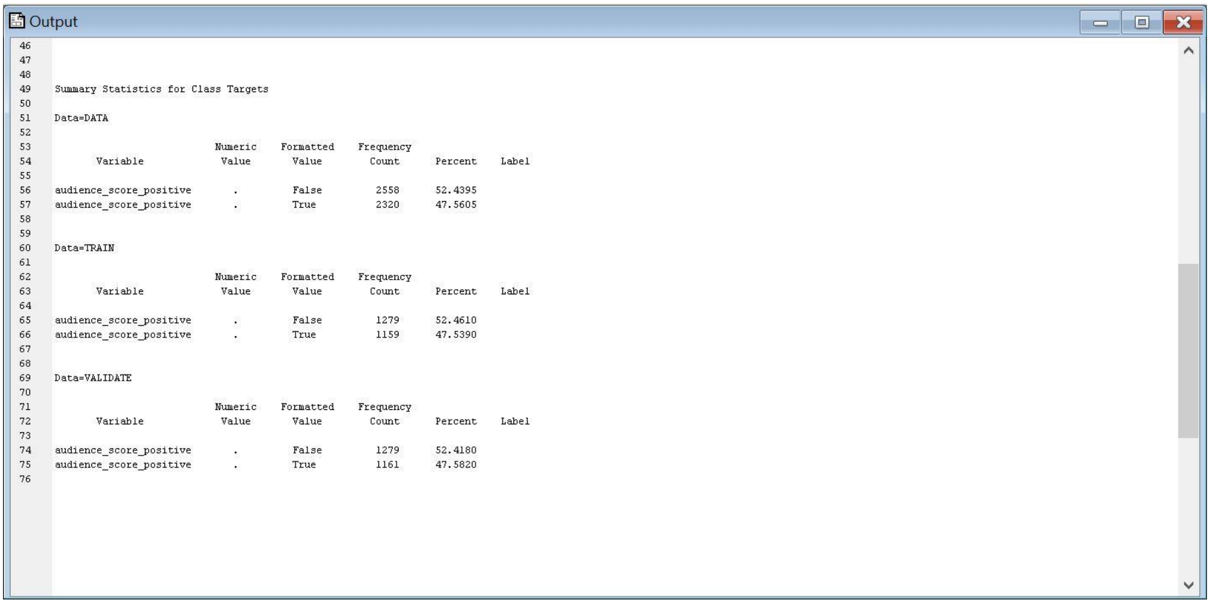
A movie streaming company (Netflix) seeks to maximize customer's retention by recommending highly rated movies with DVD or streaming options available to their users. Use sentiment score of user reviews on a movie, movie information and box office data to predict the user ratings of a movie.

By predicting the user ratings of a movie based on its reviews and box office achievement, the movie streaming company can filter out latest movies with DVD or streaming options available that are highly rated and recommend them to its users. Customers who are satisfied with the movie recommendations are more likely to subscribe to the movie streaming service in the next month.

1 Creating training and validation data

Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Metho	Default
Random Seed	12345
Data Set Allocatio	
Training	50.0
Validation	50.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	16/11/19 00:14
Run ID	e02d840b-241f-4

The training data set and validation data set are created by partitioning the raw analysis data. Since more data devoted to training partition results in more stable predictive results, and test partition is only used for calculating fit statistics after modeling and model selection is complete, the partitioning forgoes a test partition. An equal number of cases (50% for each partition) are assigned to both the training and validation partition.

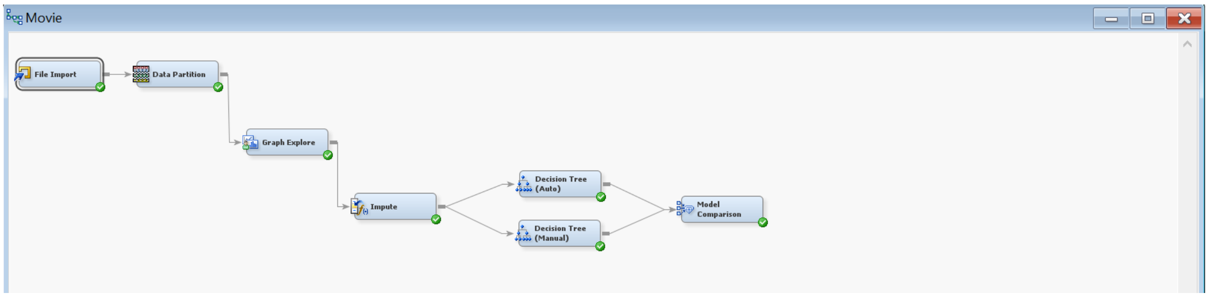


```
Summary Statistics for Class Targets

Data=DATA

                          Numeric    Formatted    Frequency
       Variable            Value      Value        Count       Percent    Label

audience_score_positive      .         False        2558       52.4395
audience_score_positive      .         True         2320       47.5605


Data=TRAIN

                          Numeric    Formatted    Frequency
       Variable            Value      Value        Count       Percent    Label

audience_score_positive      .         False        1279       52.4610
audience_score_positive      .         True         1159       47.5390


Data=VALIDATE

                          Numeric    Formatted    Frequency
       Variable            Value      Value        Count       Percent    Label

audience_score_positive      .         False        1279       52.4180
audience_score_positive      .         True         1161       47.5820
```

The proportions are not exactly the same in both the training and validation partitions due to an odd number of False and True cases in the audience_score_positive column. For both the training and validation partitions, there are 52% of False value and 47% of True values.

2 Constructing A Decision Tree Predictive Model (Interactive Decision Tree)

2.1 Build decision tree

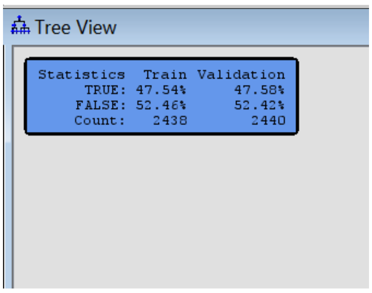


A decision tree is used to model the movie dataset from previous milestone to predict whether the movie is good or bad.

2.2 Decision Tree Node Train Properties: Splitting Rule

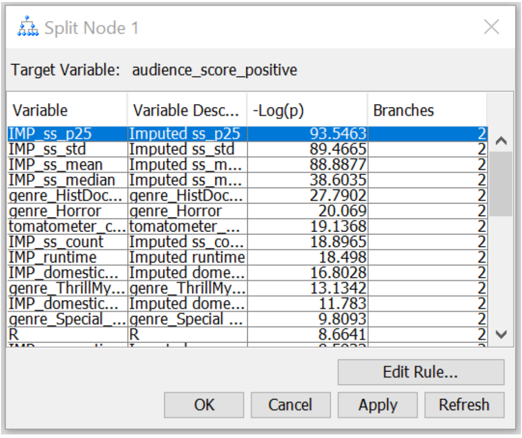
The interactive decision tree is built based on the information gain from the variables of the data, where the data with highest information gain is selected. The interactive decision tree is also built based on interpretability of the variables, which is the variable that is easier to understand tends to be selected.

2.3 Launching the Interactive Decision Tree Application

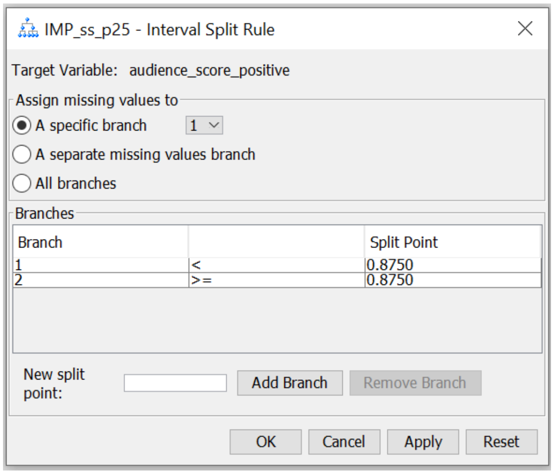


The interactive decision tree application is launched as shown above.

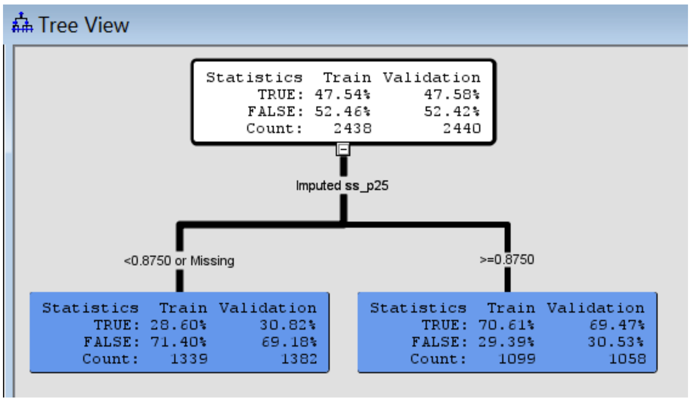
2.4 Split Nodes



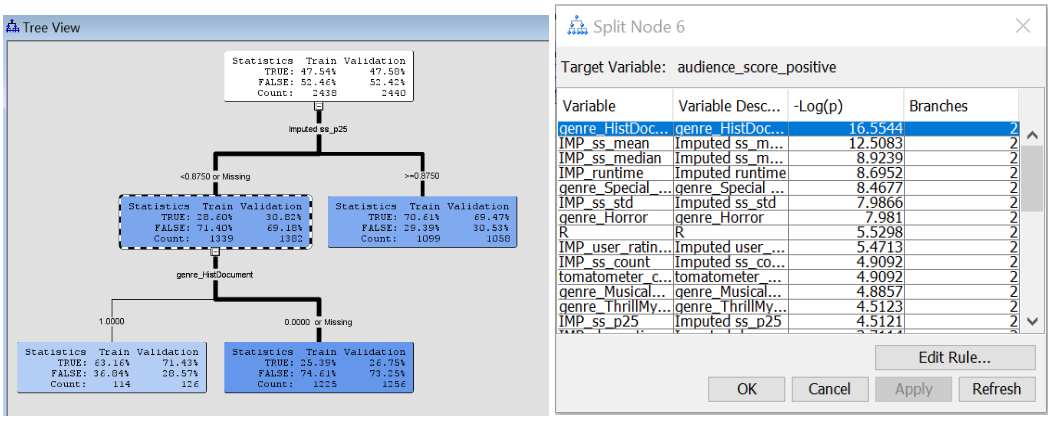
The Split Node dialog box shows the relative value, -Log(p) or logworth, of partitioning the training data using the indicated input. As the logworth increases, the partition better isolates cases with identical target values.



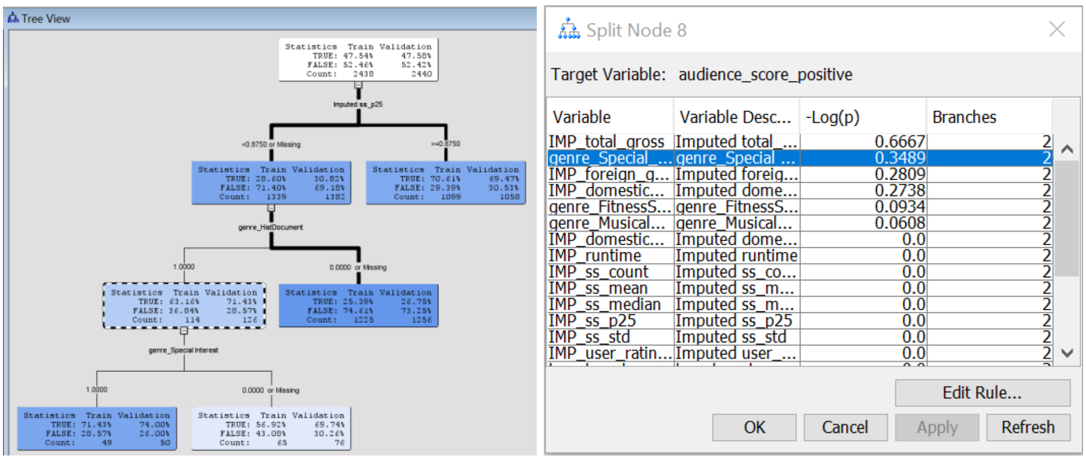
The dialog box shows how the training data is partitioned using the input IMP_ss_p25.



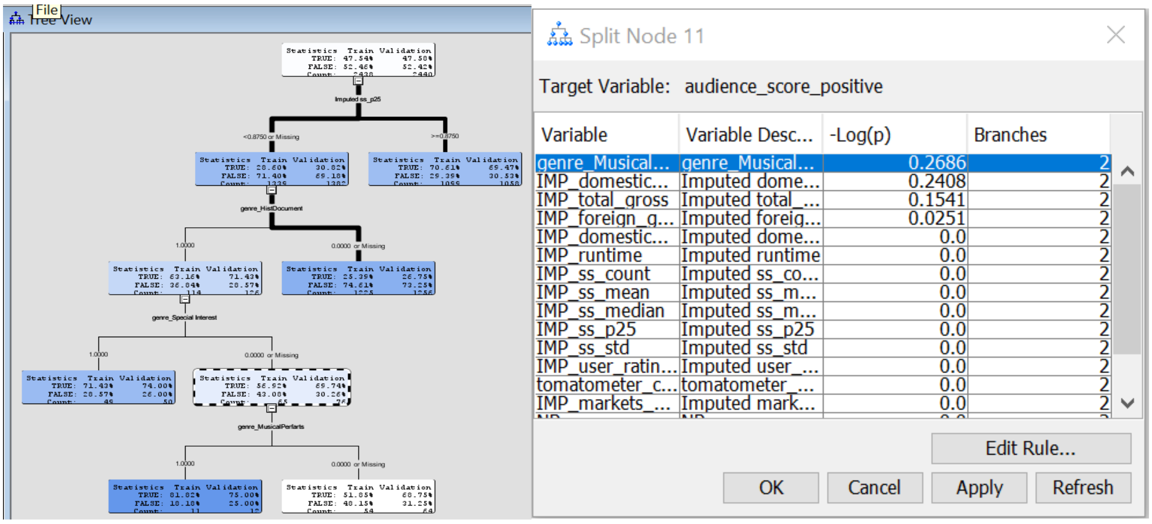
The first split is based on the IMP_ss_p25 as it has the highest logworth value. The training data is partitioned into two subsets. The first subset, corresponding to movies with first quartile sentiment scores less than 0.875 has a higher than average concentration of Target=False (bad movie) whereas the second subset, corresponding to movies with first quartile sentiment scores greater or equal to 0.875 has a higher than average concentration of Target=True (good movie).



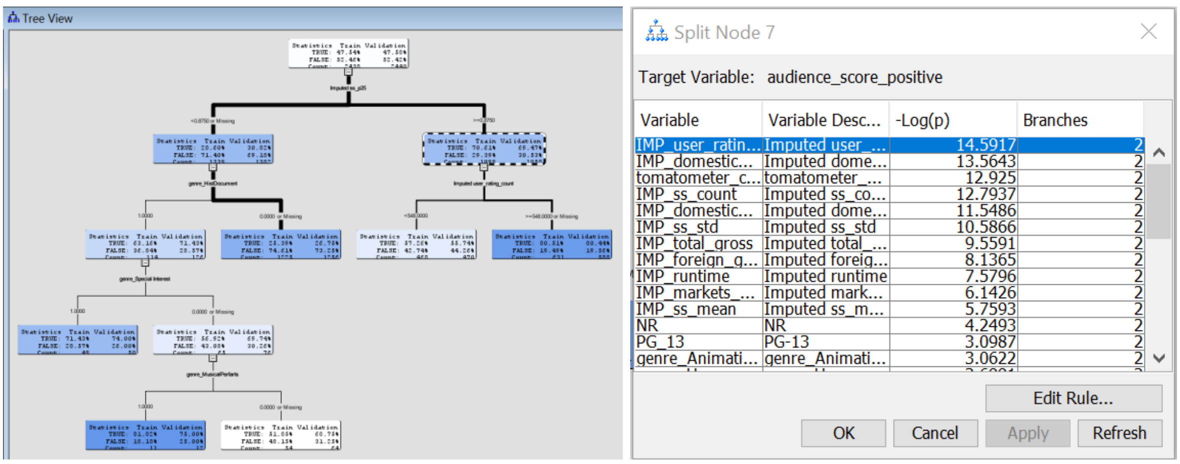
The first subset is further split into movies with genre of history or documentary as it has the highest logworth value. The movie which is not with genre of history or documentary has a concentration of higher than 70% of Target=False for both the training and validation partitions, hence is more likely to be a bad movie. The movie with genre of history or documentary requires further split.



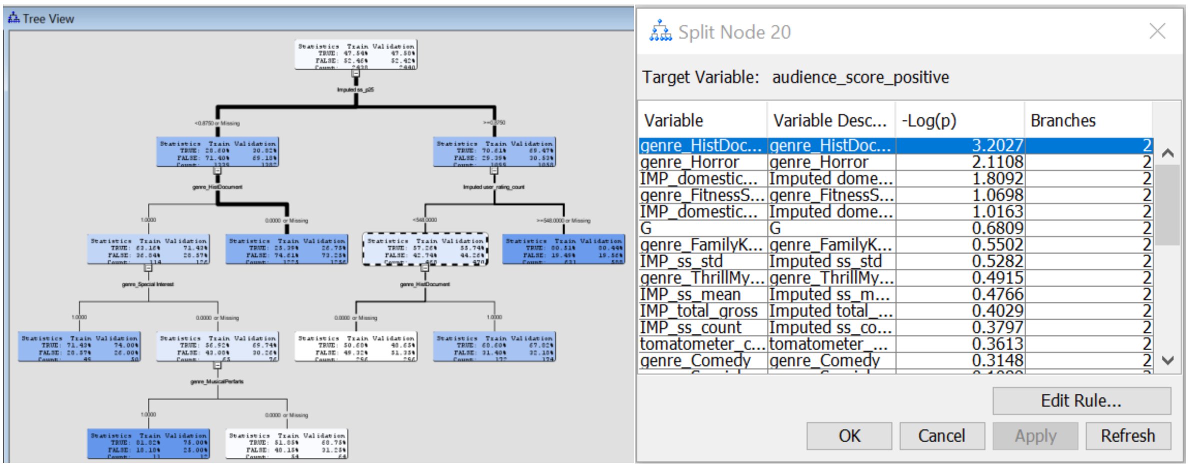
The next split is based on movie with genre of special interest. It is selected as it has second highest value of logworth and easier to understand. The movie with genre of special interest has a concentration of higher than 70% of Target =True and is more likely to be a good movie whereas the movie which is not with genre of special interest requires further split.



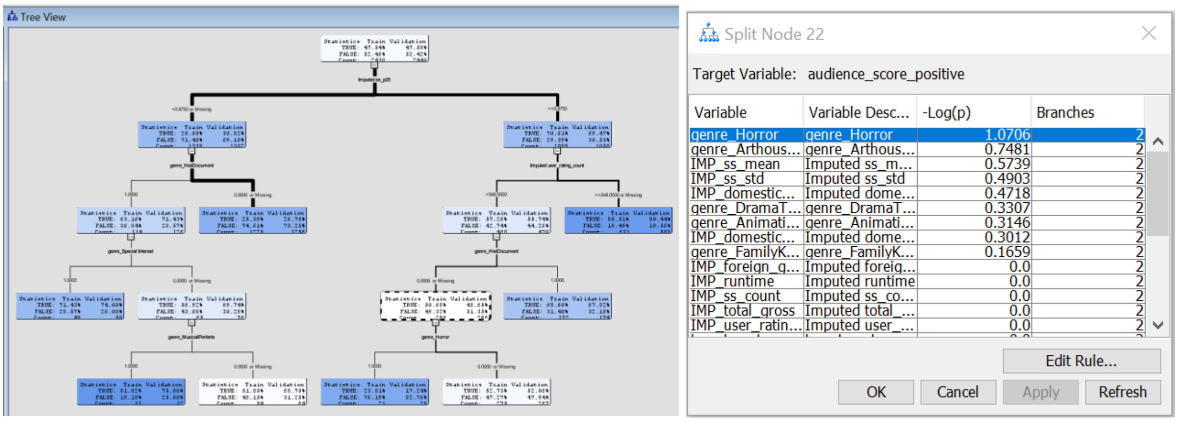
The movie with genre of special interest is further split into genre of musical or performing arts as it has the highest logworth value. The movie with genre of musical or performing arts has a higher concentration of Target=True, showing that it is a good movie when compared to the movie which is not with genre of musical or performing arts.



The second subset is further split into number of ratings given by verified users in rottentomatoes. com, which has the highest logworth value. The movie which is rated by more than or equal to 548 verified users has a concentration of higher than 80% of Target=True, indicating that it is a good movie. The movie which is rated by less than 548 users has similar concentration of about 50% for both targets and thus needs further split.



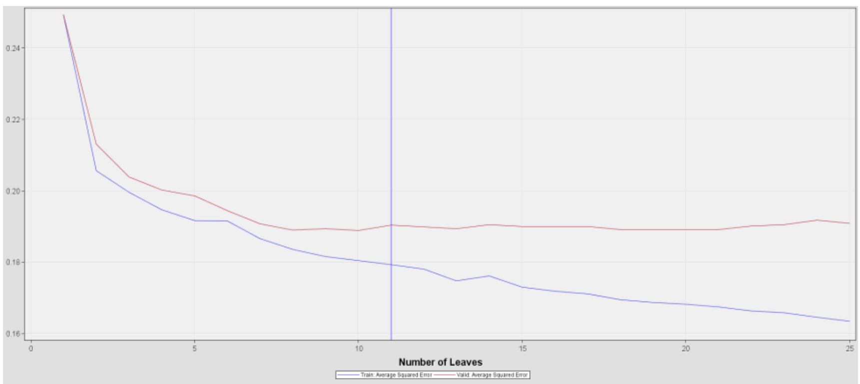
The IMP_user_rating_count is further branched into movie with genre of history and documentary. The movie with genre of history and documentary has a concentration of near to 70% of Target=True. Thus, it is highly possible that it is a good movie. The movie which is not with genre of history or documentary has a concentration of about 50% for both targets and needs further split.



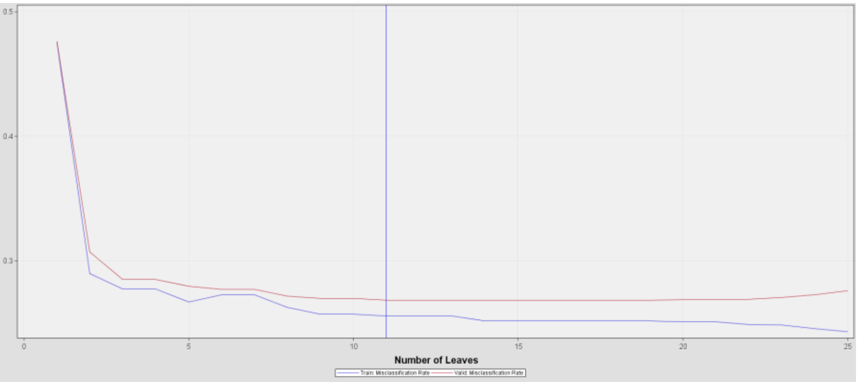
The movie with genre of history and documentary is further branched into genre of horror. The movie with genre of horror has a higher concentration of Target=True, showing that it is a good movie when compared to the movie which is not with genre of horror.

2.5 Subtree Assessment Plot

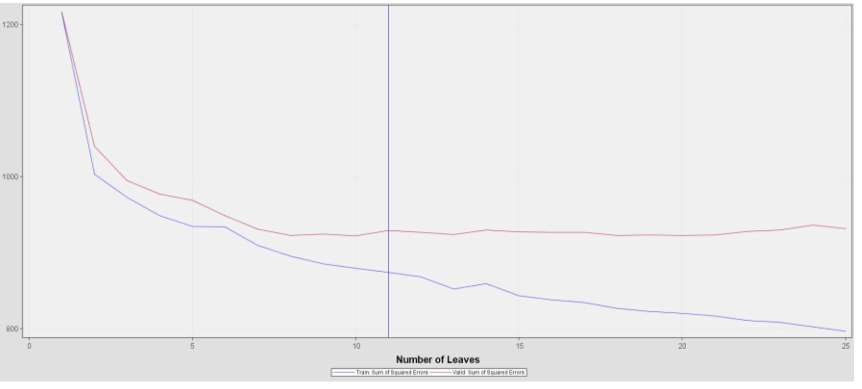
Average Square Error



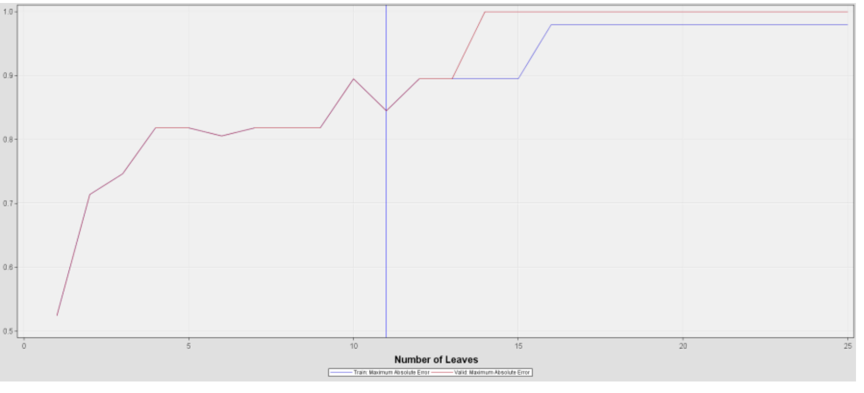
Misclassification Rate



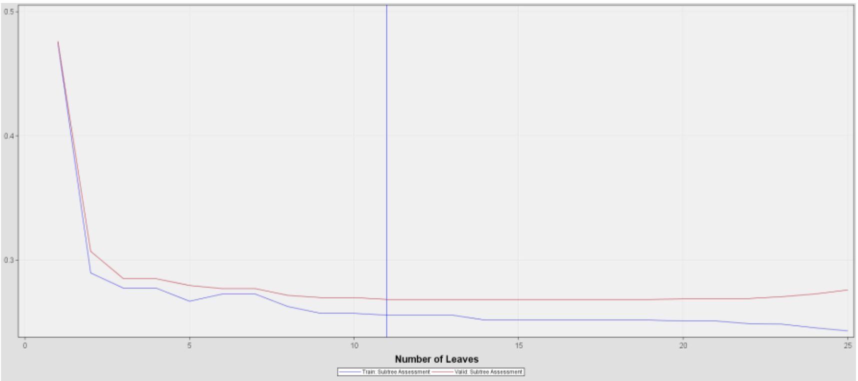
Sum of Square Errors



Maximum Assessment Error



Subtree Assessment



Generally, all the subtree assessment plots show that the model performance is getting better with the increase of leaves, but the model performance is worse for validation data when compared to training data. All the subtree assessment plots above suggested that the optimal number of leaves are 11.

2.6 Fit Statistics

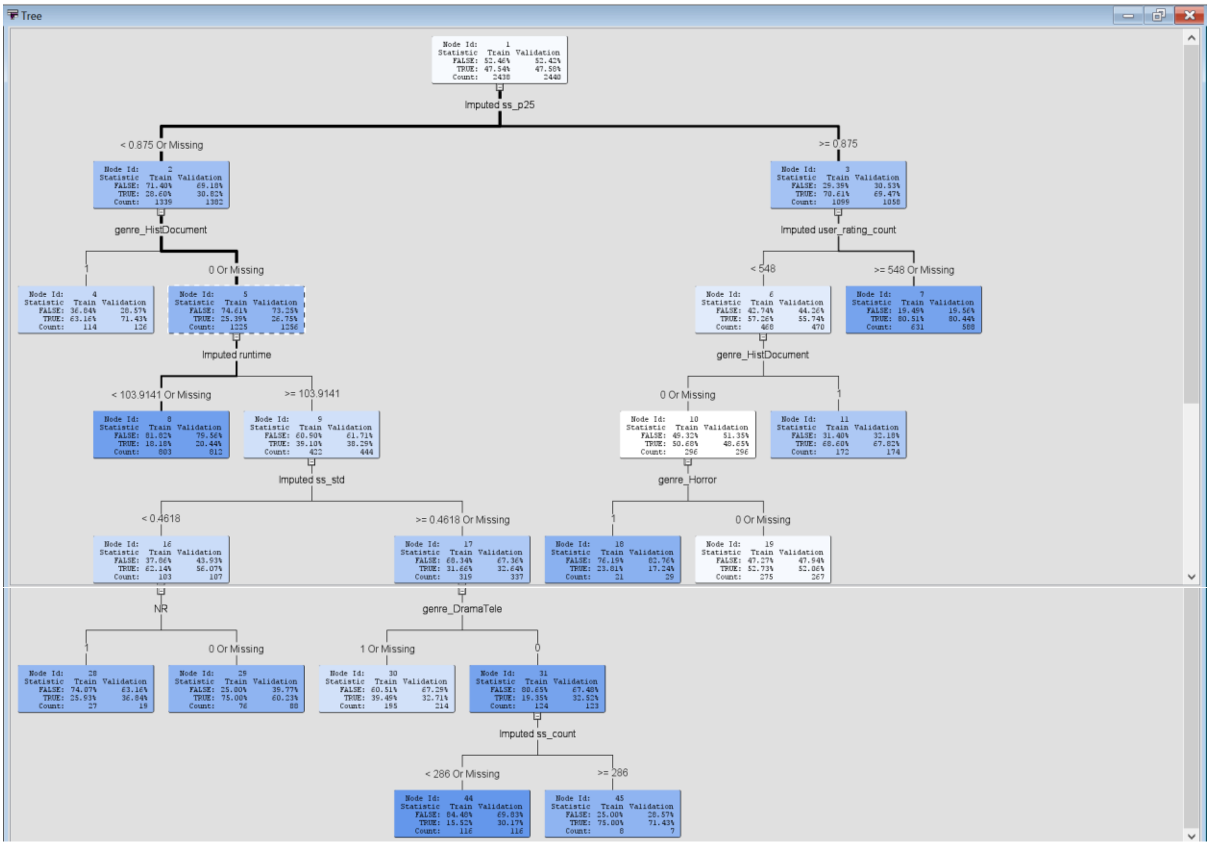
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
audience score positive		NOBS	Sum of Frequencies	2438	2440	.
audience score positive		MISC	Misclassification Rate	0.272765	0.277049	.
audience score positive		MAX	Maximum Absolute Error	0.818182	0.818182	.
audience score positive		SSE	Sum of Squared Errors	931.1744	950.5857	.
audience score positive		ASE	Average Squared Error	0.190971	0.194792	.
audience score positive		RASE	Root Average Squared Error	0.437002	0.441353	.
audience score positive		DIV	Divisor for ASE	4876	4880	.
audience score positive		DFT	Total Degrees of Freedom	2438	.	

From the fit statistics, the misclassification rate is 0.272765 for training data and 0.277049 for validation data. The average squared error is 0.190971 for training data and 0.194792 for validation error. All the statistics show that the model performs well.

3 Constructing A Decision Tree Predictive Model (Interactive Decision Tree)

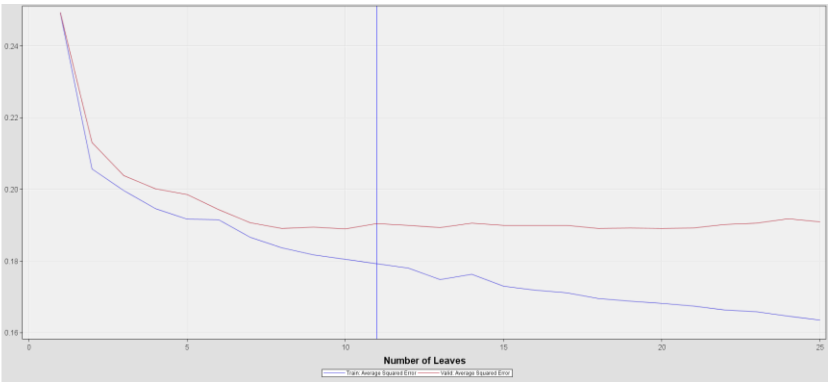
3.1 Launching the Non-Interactive Decision Tree Application



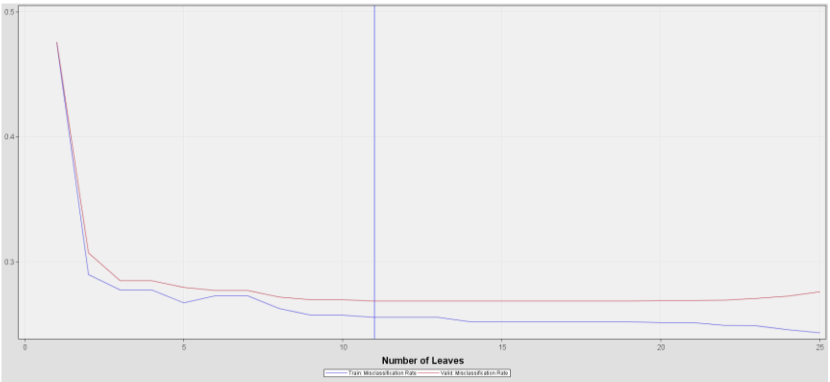
The non-interactive decision tree application is launched and trained as shown above.

3.2 Subtree Assessment Plot

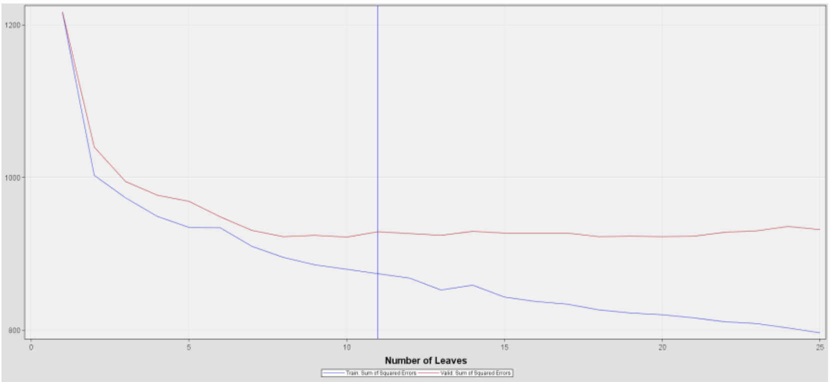
Average Square Error



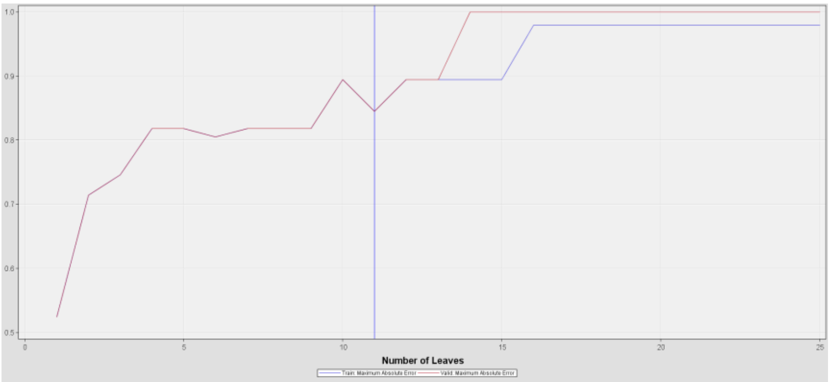
Misclassification Rate



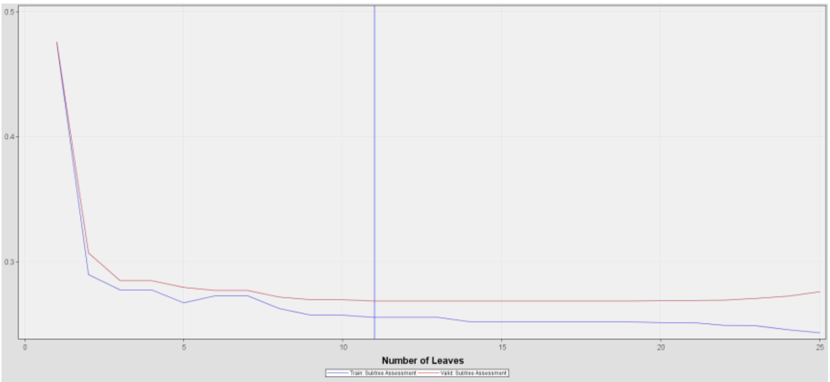
Sum of Squared Errors



Maximum Assessment Error



Subplot Assessment



Generally, all the subtree assessment plots for non-interactive decision tree also show that the model performance is getting better with the increase of leaves, but the model performance is worse for validation data when compared to training data. All the subtree assessment plots above also suggested that the optimal number of leaves are 11.

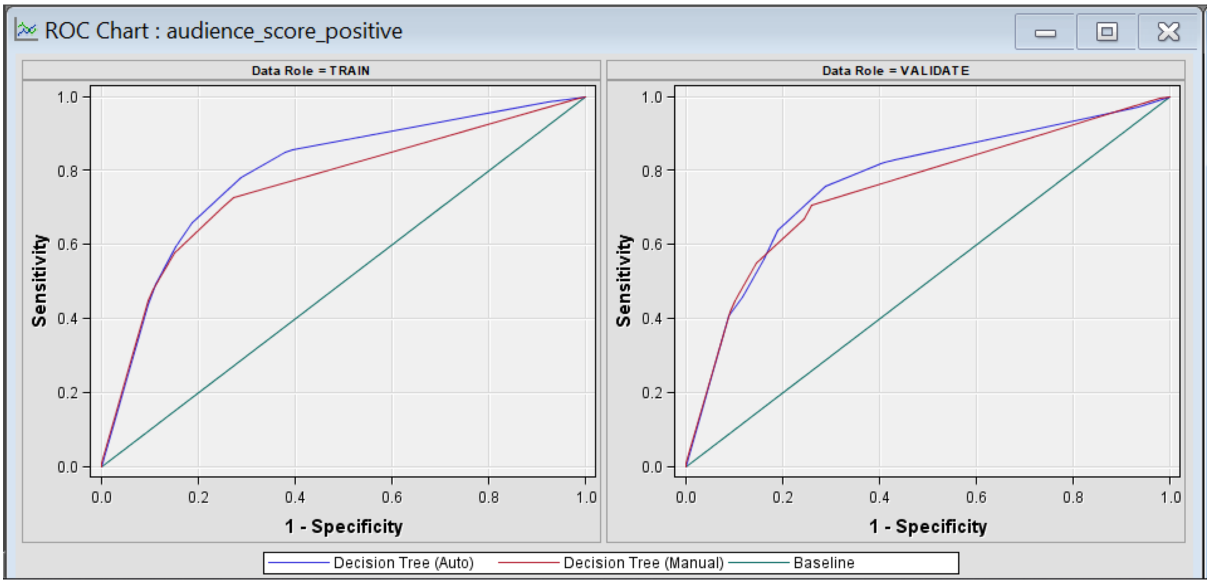
3.3 Fit Statistics

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
audience score positive		NOBS	Sum of Frequencies	2438	2440	.
audience score positive		MISC	Misclassification Rate	0.255537	0.268443	.
audience score positive		MAX	Maximum Absolute Error	0.844828	0.844828	.
audience score positive		SSE	Sum of Squared Errors	874.2862	929.1307	.
audience score positive		ASE	Average Squared Error	0.179304	0.190396	.
audience score positive		RASE	Root Average Squared Error	0.423443	0.436343	.
audience score positive		DIV	Divisor for ASE	4876	4880	.
audience score positive		DFT	Total Degrees of Freedom	2438	.	.

From the fit statistics, the misclassification rate is 0.255537 for training data and 0.268443 for validation data. The average squared error is 0.179304 for training data and 0.190396 for validation error.

4 Model comparison



The interactive decision tree is compared with the non-interactive decision tree is compared and the ROC chart for both the models is shown as above. It can be seen that the non-interactive decision tree performs slightly better than the interactive decision tree for both the training and validation data.