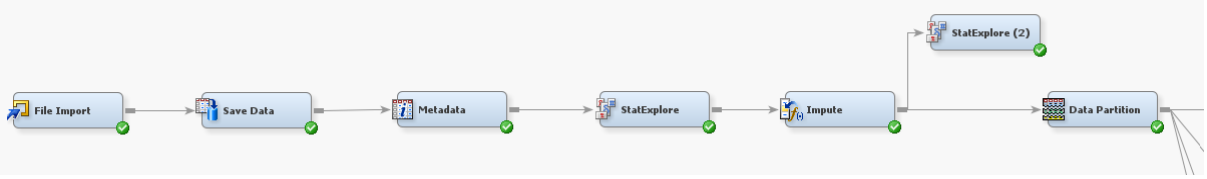
5.0 Decision Tree Modelling using SAS Enterprise Miner

5.1 Data Partition

Specify the ratio of training/validation data using “Data Partition” node.

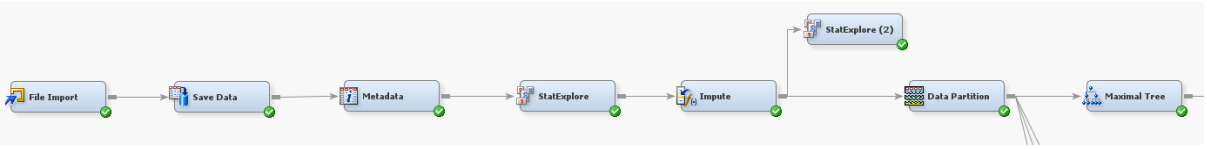


The ratio of training and validation data is 70/30.

Property	Value
General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	1/6/24 4:11 PM
Run ID	efb9ef80-6514-9e4d-bffd-c
Last Error	
Last Status	Complete
Last Run Time	1/7/24 5:27 AM
Run Duration	0 Hr. 0 Min. 3.46 Sec.
Grid Host	
User-Added Node	No

5.2 Maximal Decision Tree

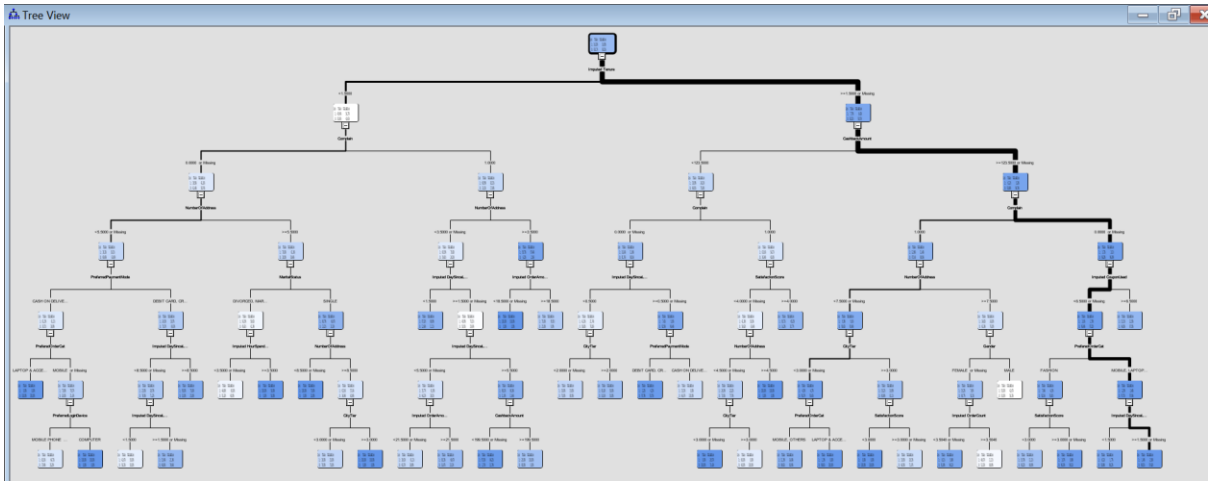
Create the maximal tree using “Decision Tree” node.



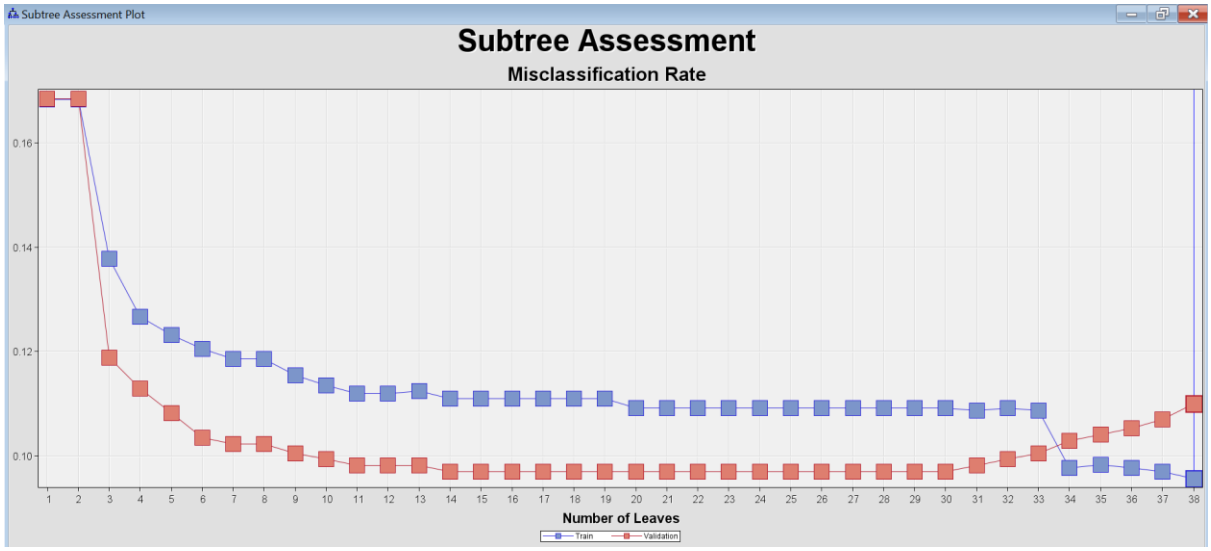
Click on the “…” button at “Interactive” row to open the Interactive Decision Tree tool.

Property	Value
General	
Node ID	Tree
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Interactive	
Import Tree Model	No
Tree Model Data Set	
Use Frozen Tree	No
Use Multiple Targets	No

Right click on the root node of the tree and select Train Node. This will grow the tree until stopping rules prohibited further growth. Figure below shows the maximal tree with 38 leaves.



Based on the Subtree Assessment Plot, it appears that the maximal, 38-leaf tree gives a lower misclassification rate than any of its simpler predecessors. However, it is misleading because it applies to training data only. Further optimization is therefore required.



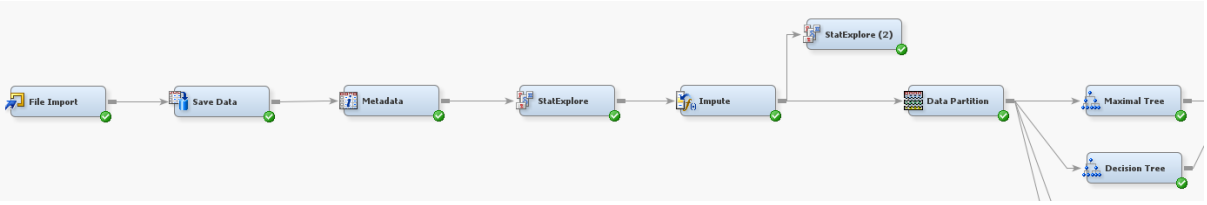
Based on Fit Statistics, misclassification rate is 0.0957 for training dataset and 0.1099 for validation dataset.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		NOBS	Sum of Frequencies	3939	1691	.
Churn		MISC	Misclassification Rate	0.09571	0.109994	.
Churn		MAX	Maximum Absolute Error	0.986111	1	.
Churn		SSE	Sum of Squared Errors	555.5301	267.5079	.
Churn		ASE	Average Squared Error	0.070517	0.079098	.
Churn		RASE	Root Average Squared Error	0.26555	0.281243	.
Churn		DIV	Divisor for ASE	7878	3382	.
Churn		DFT	Total Degrees of Freedom	3939	.	.

5.3 Pruned Decision Tree

Create a decision tree using “Decision Tree” node.

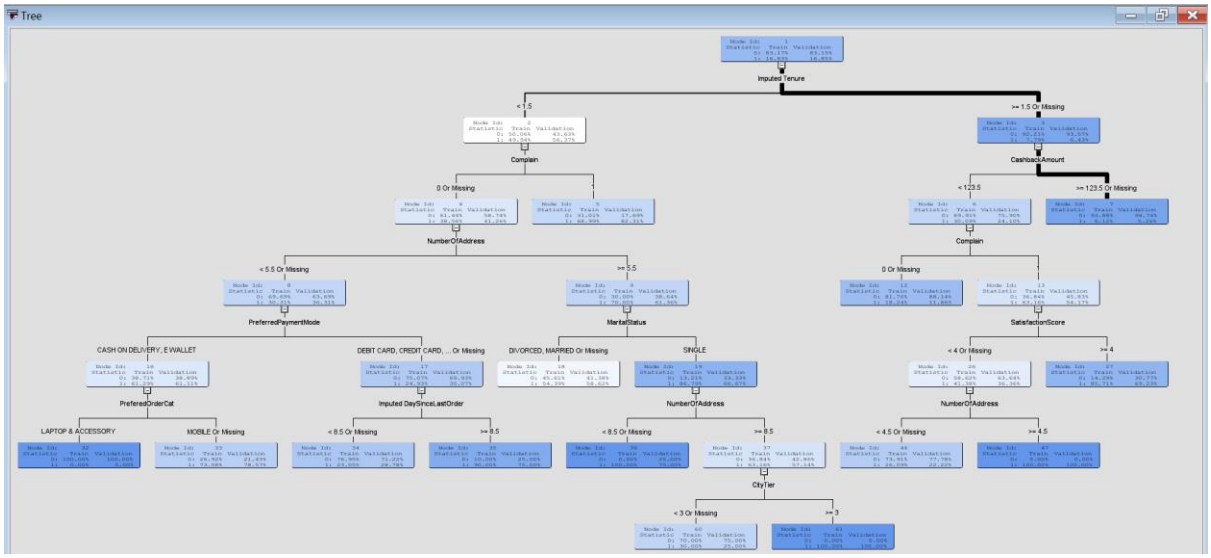


Go to the “Subtree” section of the properties table to specify the tree pruning properties. The method used to prune the maximal tree is Assessment. This means that the algorithms choose the best tree based on the optimality measure specificized by the Assessment Measure. By setting Assessment Measure as Decision, the algorithms will choose a tree that is optimized for making the best decisions (as opposed to best rankings or best probability estimates). Keep other settings as default.

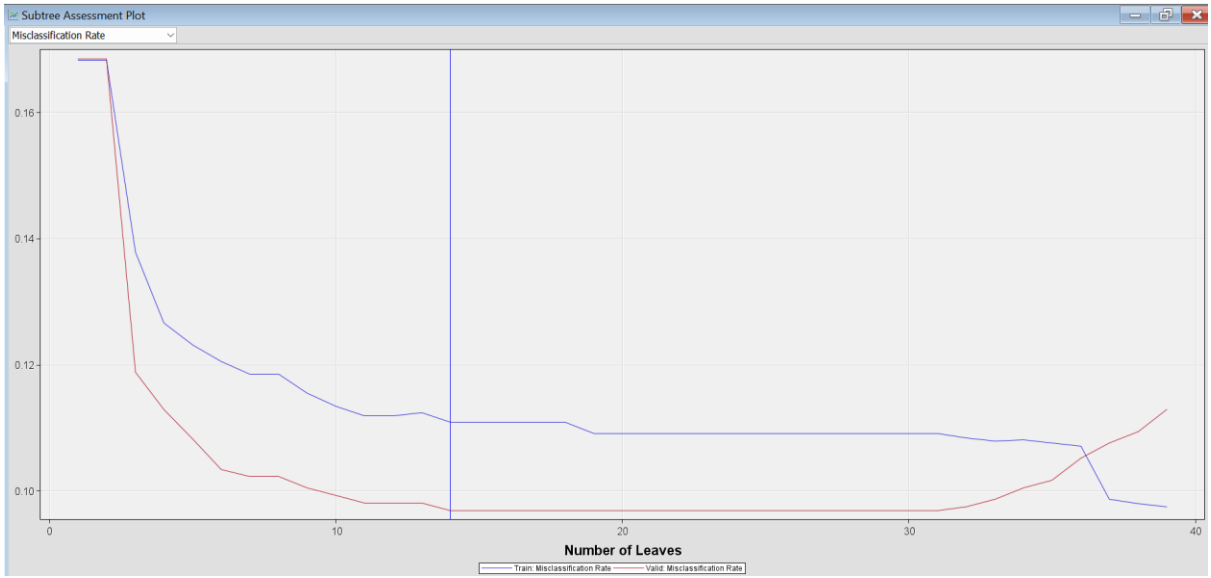
Property	Value
General	
Node ID	Tree2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Interactive	
Import Tree Model	No
Tree Model Data Set	
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000

Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importance	
Observation Based Importance	No
Number Single Var Importance	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustment	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	
Leaf Variable	Yes
Interactive Sample	
Create Sample	Default
Sample Method	Random
Sample Size	10000
Sample Seed	12345
Performance	Disk
Score	
Variable Selection	Yes
Leaf Role	Segment
Report	
Precision	4
Tree Precision	4
Class Target Node Color	Percent Correctly Classi
Interval Target Node Color	Average
Node Text	

Figure below shows the pruned decision tree with 14 leaves.



Based on the Subtree Assessment Plot, it appears that misclassification rate is most optimized when the number of leaves equals 14. The validation misclassification rate plateaued out at 0.097 when number of leaves increased from 15 to 31. Beyond 31, validation misclassification rate increases. Therefore, 14 leaves give the most optimized misclassification rate.



Based on Fit Statistics, misclassification rate is 0.1109 for training dataset and 0.09698 for validation dataset.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		NOBS	Sum of Frequencies	3939	1691	.
Churn		MISC	Misclassification Rate	0.110942	0.096984	.
Churn		MAX	Maximum Absolute Error	0.938846	1	.
Churn		SSE	Sum of Squared Errors	706.6919	278.7777	.
Churn		ASE	Average Squared Error	0.089704	0.08243	.
Churn		RASE	Root Average Squared Error	0.299507	0.287106	.
Churn		DIV	Divisor for ASE	7878	3382	.
Churn		DFT	Total Degrees of Freedom	3939	.	.

The Variable Importance Plot displays the importance of each predictor variable in the model. Only 10 out of 18 input variables are important to the pruned decision tree model.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
IMP_Tenure	Imputed Tenure	1	1.0000	1.0000	1.0000
Complain		2	0.4765	0.4814	1.0102
NumberOfAddress		3	0.3905	0.1960	0.5019
CashbackAmount		1	0.3130	0.1963	0.6273
PreferredPaymentMode		1	0.2434	0.2114	0.8688
PreferedOrderCat		1	0.1949	0.2351	1.2064
IMP_DaySinceLastOrder	Imputed DaySi...	1	0.1922	0.0988	0.5140
MaritalStatus		1	0.1564	0.0000	0.0000
SatisfactionScore		1	0.1541	0.0820	0.5320
CityTier		1	0.1403	0.1201	0.8557
NumberOfDeviceRegistered		0	0.0000	0.0000	.
PreferredLoginDevice		0	0.0000	0.0000	.
IMP_CouponUsed	Imputed Coupo...	0	0.0000	0.0000	.
IMP_OrderAmountHikeFromlastYear	Imputed Order...	0	0.0000	0.0000	.
IMP_HourSpendOnApp	Imputed HourS...	0	0.0000	0.0000	.
Gender		0	0.0000	0.0000	.
IMP_OrderCount	Imputed Order...	0	0.0000	0.0000	.
IMP_WarehouseToHome	Imputed Wareh...	0	0.0000	0.0000	.