

Steps to SAS e-Miner

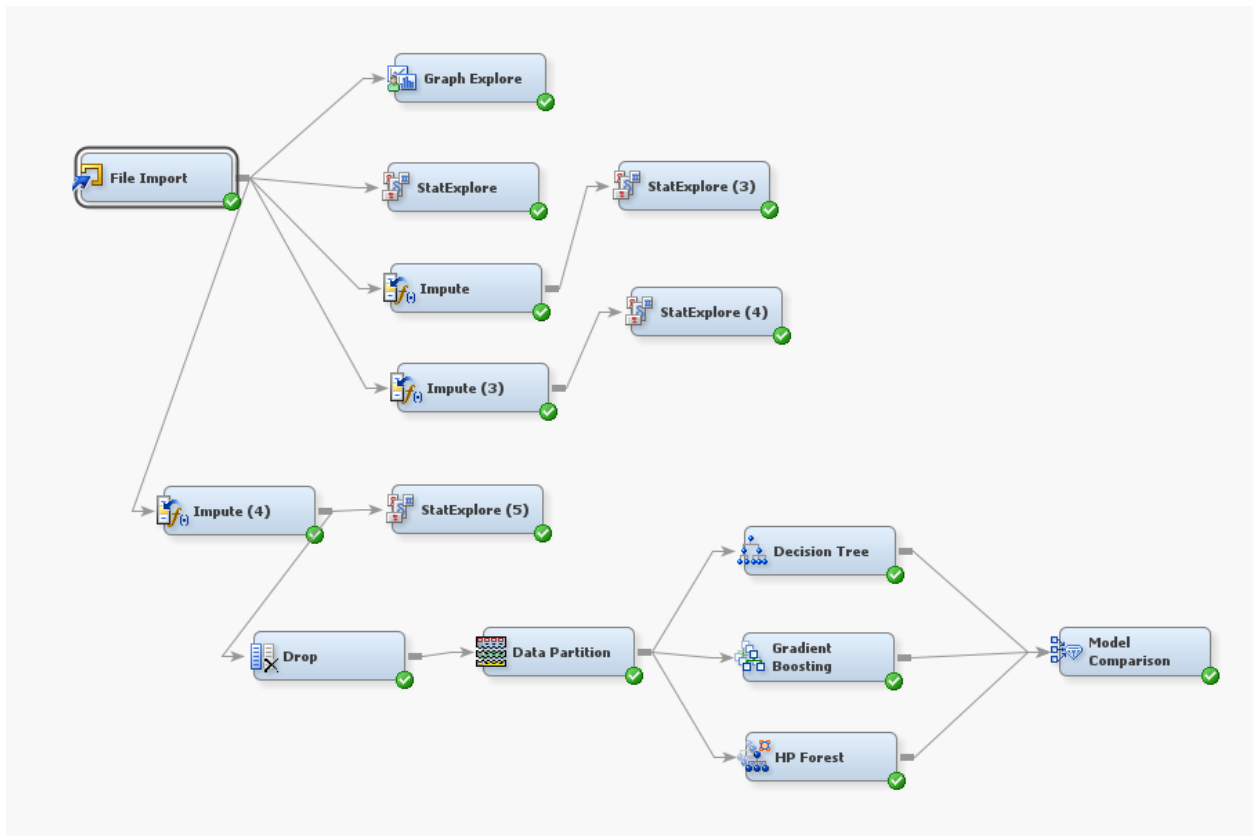


Figure 1: Overview of Workflow in SAS e-Miner

1. File Import Settings.

The File Imported was the one prepped with Knime.

Property

Value

General

Node ID: FIMPORT3
Imported Data: ...
Exported Data: ...
Notes: ...

Train

Variables: ...
Import File: C:\Users\grace\Data Science\...
Maximum Rows to Import: 1000000
Maximum Columns to Import: 10000
Delimiter: ,
Name Row: Yes
Number of Rows to Skip: 0
Guessing Rows: 500
File Location: Local
File Type: csv
Advanced Advisor: No
Rerun: No

Score

Role: Train

Report

Summarize: No

Status

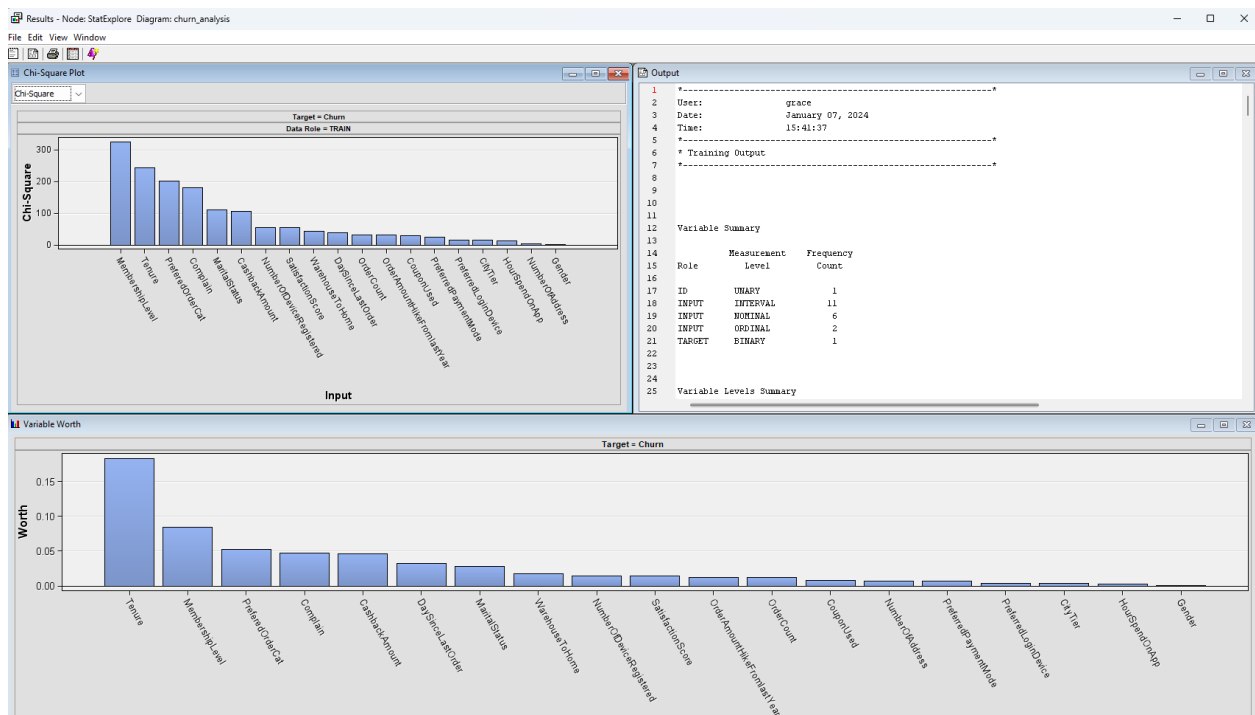
Crash Time: 1/7/24 3:38 PM

☐ Label
☐ Mining
☐ Basic
☐ Statistics

Name	Role	Level	Report	Order	Drop	L
CashbackAmount	Input	Interval	No		No	
Churn	Target	Binary	No		No	
CityTier	Input	Ordinal	No		No	
Complain	Input	Interval	No		No	
CouponUsed	Input	Interval	No		No	
CustomerID	ID	Unary	No		No	
DaySinceLastOrder	Input	Interval	No		No	
Gender	Input	Nominal	No		No	
HourSpendOnApp	Input	Interval	No		No	
MaritalStatus	Input	Nominal	No		No	
MembershipLevel	Input	Nominal	No		No	
NumberOfAddresses	Input	Interval	No		No	
NumberOfDevices	Input	Interval	No		No	
OrderAmountHistory	Input	Interval	No		No	
OrderCount	Input	Interval	No		No	
PreferredOrderC	Input	Nominal	No		No	
PreferredLoginD	Input	Nominal	No		No	
PreferredPayment	Input	Nominal	No		No	
SatisfactionScore	Input	Ordinal	No		No	
Tenure	Input	Interval	No		No	
WarehouseToDel	Input	Interval	No		No	

Figure 2: File Import Settings

2. Added StatExplore



Read the summary statistics to understand the distribution of the data and what might the distribution mean. The explanation was written in the report.

3. Added Graph Explore

Property	Value
General	
Node ID	GrfExpl2
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
<input checked="" type="checkbox"/> Sample Properties	
Method	Stratify
Size	Max
Random Seed	12345
Report	
Target	Yes
Group by Target	Yes
Status	
Create Time	1/7/24 3:39 PM
Run ID	1abd46f3-4ba3-4ba6-90b5-f881ad
Last Error	
Last Status	Complete
Last Run Time	1/7/24 3:41 PM
Run Duration	0 Hr. 0 Min. 3.92 Sec.
Grid Host	
User-Added Node	No

Figure 3: Graph Explore Setting Size Changed to Max

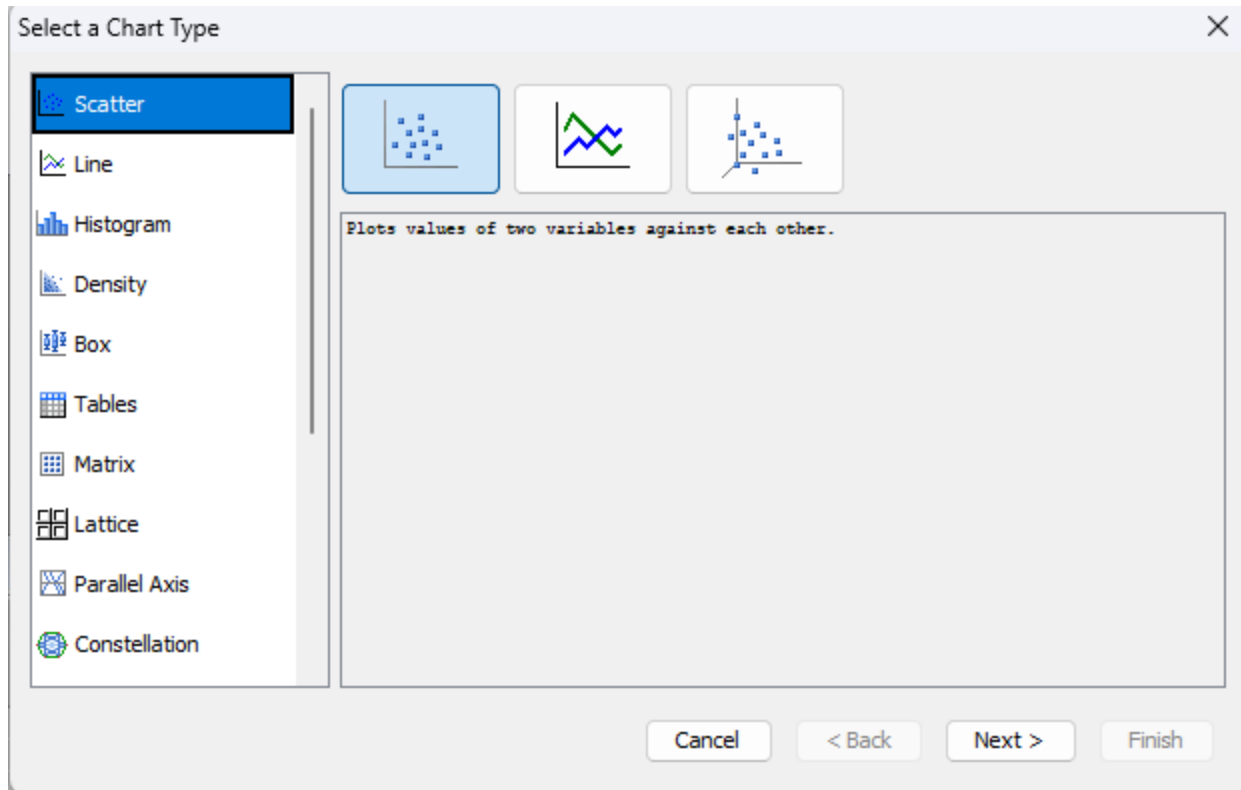


Figure 4: Tested Plots for Visualization of Data

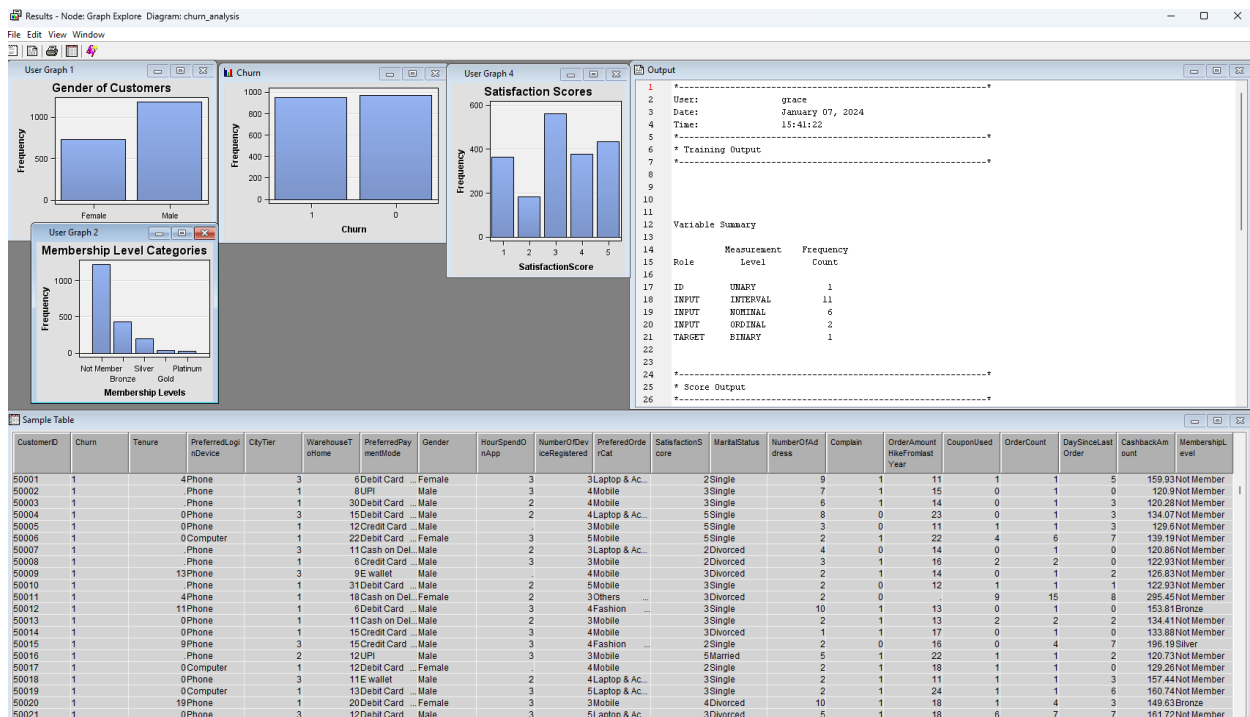


Figure 5: Current GraphExplore Results Display

Plots were added to get a better visualization and understanding of the data, but were not added to the case study report as the focus of the case study was on training a good model.

4. Added First Impute Node

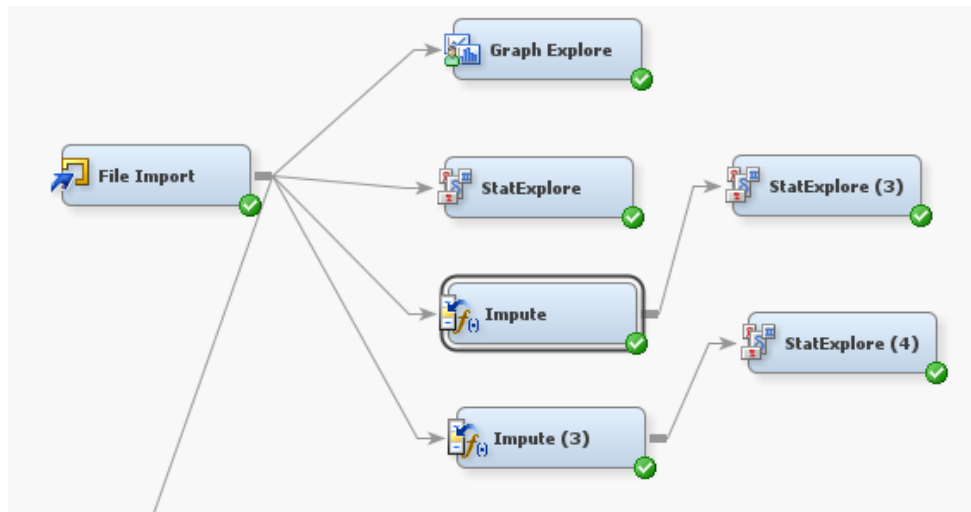


Figure 6: Impute Node

.. Property	Value
General	
Node ID	Impt2
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Nonmissing Variables	No
Missing Cutoff	50.0
<input checked="" type="checkbox"/> Class Variables	
Default Input Method	None
Default Target Method	None
Normalize Values	Yes
<input checked="" type="checkbox"/> Interval Variables	
Default Input Method	Median
Default Target Method	None
<input checked="" type="checkbox"/> Default Constant Value	
Default Character Value	
Default Number Value	.
<input checked="" type="checkbox"/> Method Options	
Random Seed	12345
Tuning Parameters	...
Tree Imputation	...
Score	

Figure 7: First Impute Node Setting Default Input Method as Median

5. Linked a StatExplore Node to First Impute Node

- A summary statistic to check the mean and standard deviation if impute is median

6. Added Second Impute Node

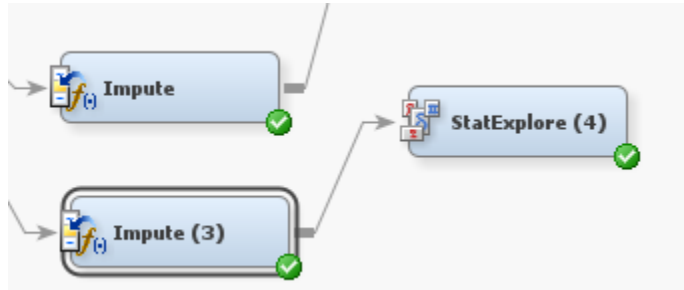


Figure 7: Second Impute Node

Property	Value
General	
Node ID	Impt3
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Count
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Tree Surrogate
Default Target Method	None
Default Constant Value	
Default Character Value	
Default Number Value	.
Method Options	
Random Seed	12345
Tuning Parameters	...
Tree Imputation	...

Figure 8: Second Impute Node's Input Method Default Set to Tree Surrogate

7. Linked a StatExplore Node to Second Impute Node

- a. A summary statistic to check the mean and standard deviation if impute is Tree Surrogate
8. After comparison of the summary statistics of the two imputation methods, added a third impute node and configured the final setting of a third impute node.

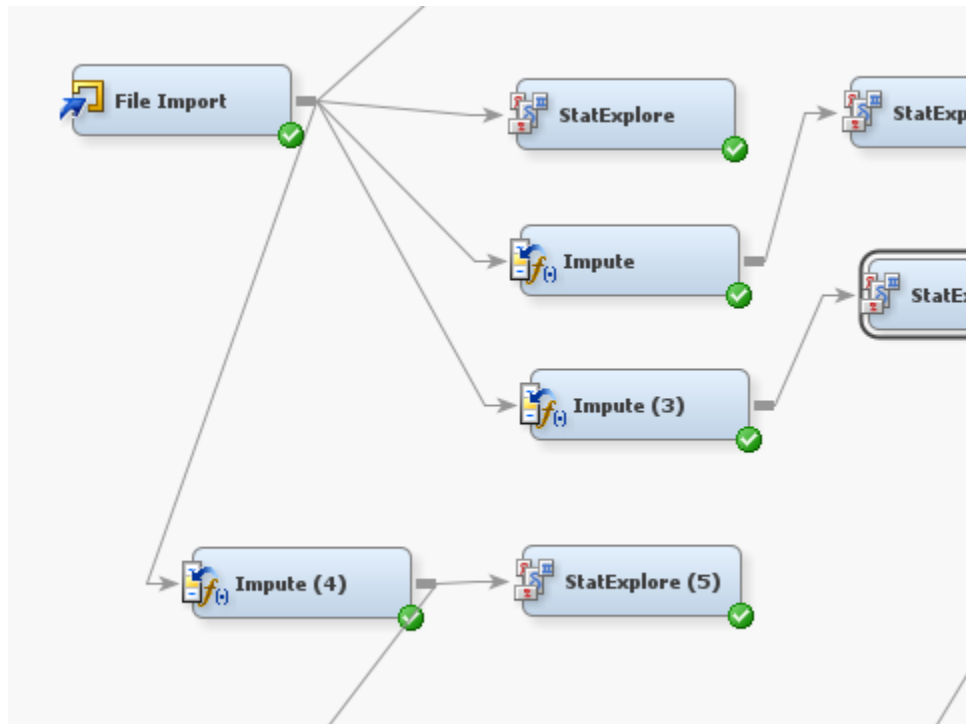


Figure 9: Impute (4) being the third impute node.

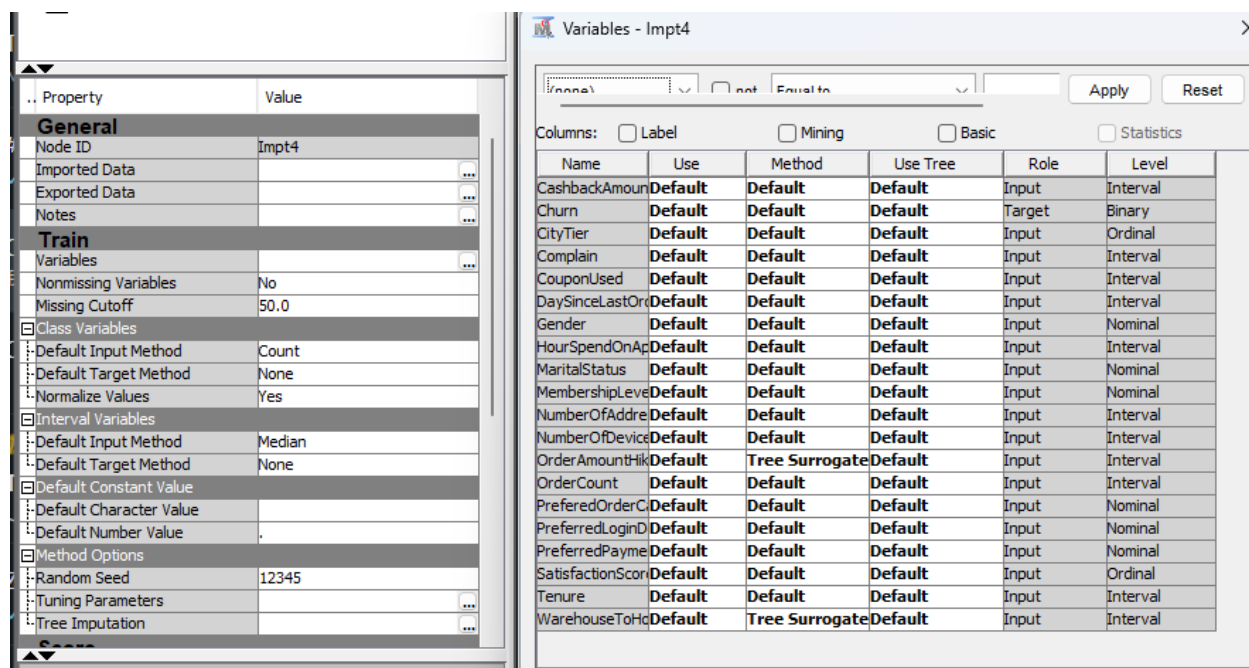


Figure 10: Final Setting of Imputation

9. Linked a Summary Statistics Node to Final Impute Node to confirm distribution and proportions

10. Added a Drop Node

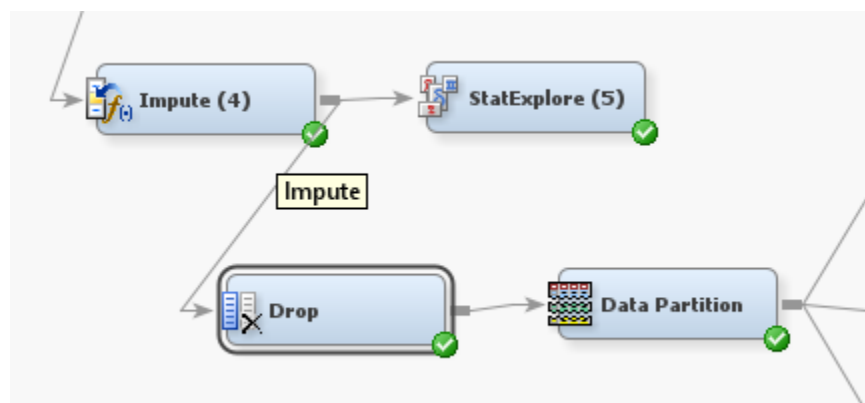


Figure 11: Drop Node

Variables - Drop
 ✕

(none)

not

Equal to

Apply

Reset

Columns:

☐ Label
 ☐ Mining
 ☐ Basic
 ☐ Statistics

Name	Drop	Role	Level
CashbackAmount	Default	Input	Interval
Churn	Default	Target	Binary
CityTier	Default	Input	Ordinal
Complain	Default	Input	Interval
CustomerID	Default	ID	Unary
Gender	Yes	Input	Nominal
IMP_CouponUse	Default	Input	Interval
IMP_DaySinceLa	Default	Input	Interval
IMP_HourSpend	Default	Input	Interval
IMP_OrderAmou	Default	Input	Interval
IMP_OrderCoun	Default	Input	Interval
IMP_Tenure	Default	Input	Interval
IMP_Warehouse	Default	Input	Interval
MaritalStatus	Default	Input	Nominal
MembershipLeve	Default	Input	Nominal
NumberOfAddre	Default	Input	Interval
NumberOfDevice	Default	Input	Interval
PreferredOrderC	Default	Input	Nominal
PreferredLoginD	Default	Input	Nominal
PreferredPayme	Default	Input	Nominal
SatisfactionScor	Default	Input	Ordinal
W/ARN	Default	Assessment	Nominal

Explore...

Update Path

OK

Cancel

Figure 12: Drop Node Setting

11. Added Data Partition Node

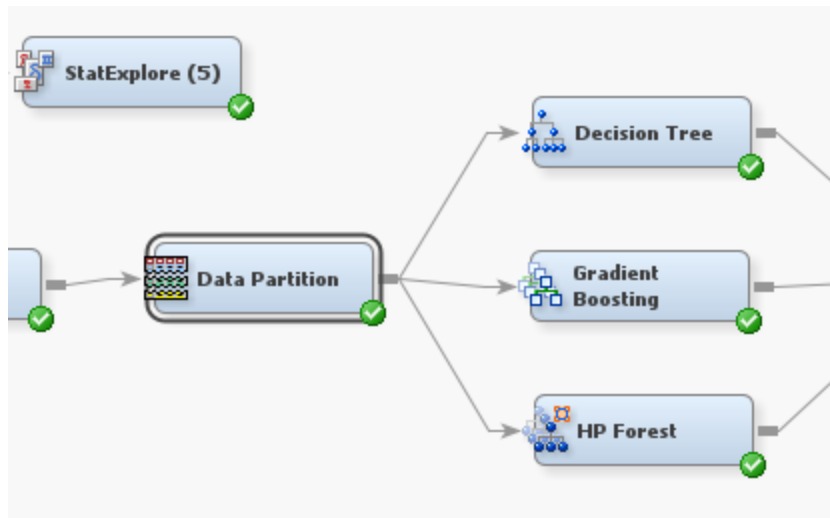


Figure 13: Added Data Partition

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	1/7/24 5:28 PM
Run ID	c378bd89-a199-47d8-ab06-005-
Last Error	
Last Status	Complete
Last Run Time	1/7/24 7:06 PM
Run Duration	0 Hr 0 Min 2.65 Sec

Figure 14: Data Partition Setting

12. Added Model Nodes

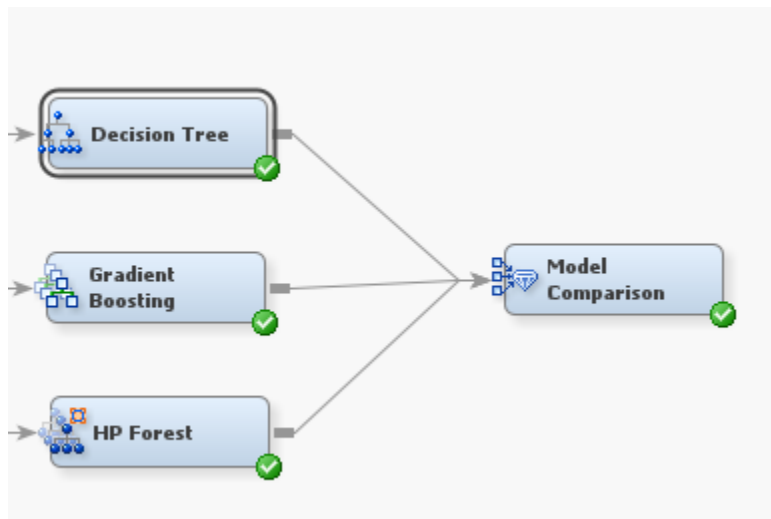


Figure 15: Model Nodes: Decision Tree, Gradient Boosting, and HP Forest

The three nodes used for modeling are the Decision Tree, Gradient Boosting, and HP Forest. The configuration of each model will be explained in the following sections.

Property	Value
Minimum Categorical Size	5
Node	
Leaf Size	8
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	Yes
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importance	
Observation Based Importance	No
Number Single Var Importance	5

Figure 16 Setting of Decision Tree

The input data will be modeled using different decision trees. Firstly, a full decision tree will be trained and configured to have a maximum depth property of 10 and a leaf size of 8. A maximum depth of 10 allows the decision tree to have a reasonably complex structure, capturing intricate patterns in the data.

Property	Value
General	
Node ID	Boost
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Series Options	
N Iterations	50
Seed	12345
Shrinkage	0.1
Train Proportion	60
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	10
Minimum Categorical Size	5
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Use in search
Performance	Disk
Node	
Leaf Fraction	0.001

Figure 17: Setting of Gradient Boosting

There is little risk of overfitting as the dataset is large enough. The leaf size of 8 would add regularization and prevent the model from being too sensitive to noise in the training data as the training data has 20 features. Gradient Boosting was set to have a maximum depth of 10. With Gradient Boosting often combining weak learners, this should also capture a more complex pattern than the default depth of 2 and should be less prone to overfitting than the decision tree.

Property	Value
General	
Node ID	HPDMForest
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Tree Options	
Maximum Number of Trees	100
Seed	12345
Type of Sample	Proportion
Proportion of Obs in Each Sample	0.6
Number of Obs in Each Sample	.
Splitting Rule Options	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Search	1
Number of Variables to Consider	.
Significance Level	0.05
Max Categories in Split Search	30
Minimum Category Size	5
Exhaustive	5000
Node Options	
Method for Leaf Size	Default

Figure 18: Setting of HP Random Forest.

The HP Random Forest setting was kept the same since it is more robust to overfitting as compared to decision trees and this should be a good starting point for a prediction of customer churn.

13. Added Model Comparison Node

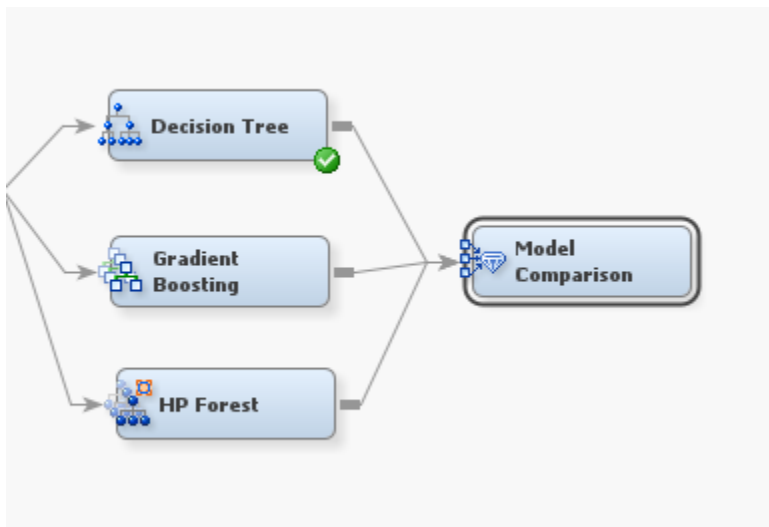


Figure 19: Addition of Model Comparison Node

To assess and compare the models, a model comparison node was added to the diagram.

Fit Statistics						
Model Selection based on Valid: Misclassification Rate (_VMISC_)						
Selected			Valid:	Train:		Valid:
Model	Model Node	Model Description	Misclassification	Average	Train:	Average
			Rate	Squared	Misclassification	Squared
				Error	Rate	Error
Y	Boost	Gradient Boosting	0.08666	0.00336	0.00000	0.07459
	Tree	Decision Tree	0.18544	0.13573	0.16990	0.14661
	HPDMForest	HP Forest	0.18891	0.12096	0.16915	0.13986

Figure 20: Fit Statistics of Three Models