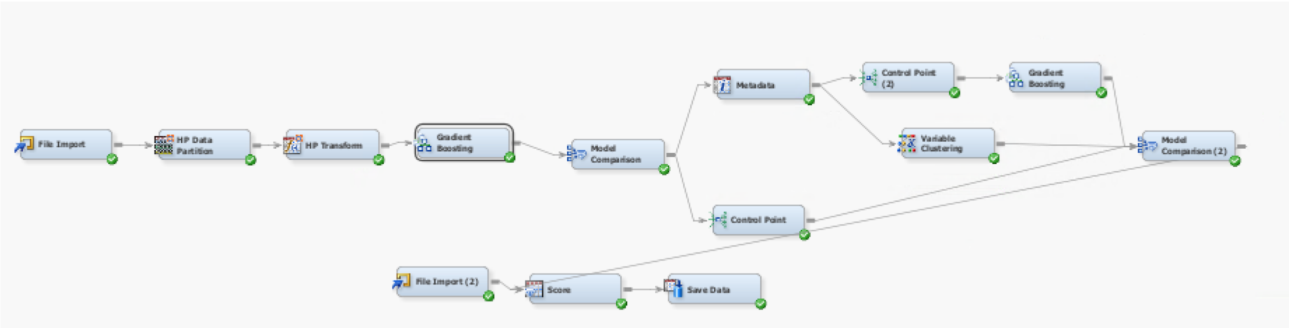
Group Project: MGMT 571 Data Mining

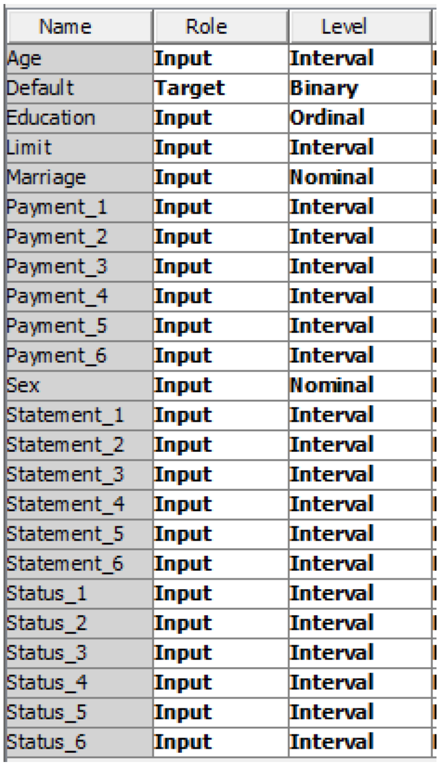
Group Name: Team Machine

Group Members: Li-Ci Chuang, Rachel Fagan, and Yi-Hsuan Hsu

Model #1:

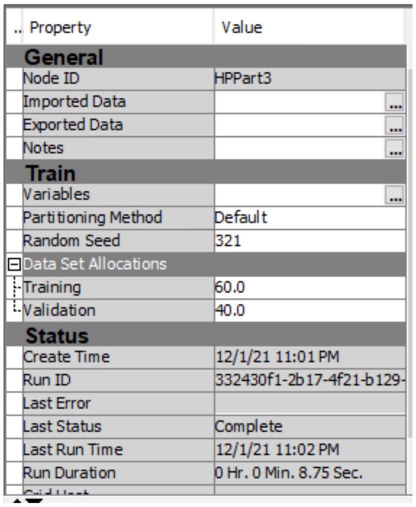


File Import: Use the creditDefault_Train dataset. Set variables as below:



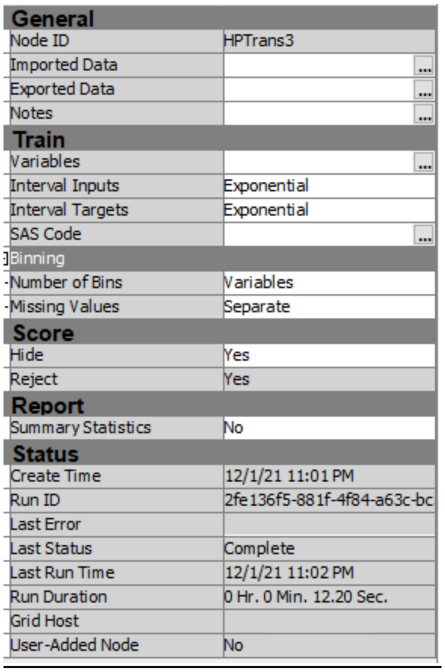
Name	Role	Level
Age	Input	Interval
Default	Target	Binary
Education	Input	Ordinal
Limit	Input	Interval
Marriage	Input	Nominal
Payment_1	Input	Interval
Payment_2	Input	Interval
Payment_3	Input	Interval
Payment_4	Input	Interval
Payment_5	Input	Interval
Payment_6	Input	Interval
Sex	Input	Nominal
Statement_1	Input	Interval
Statement_2	Input	Interval
Statement_3	Input	Interval
Statement_4	Input	Interval
Statement_5	Input	Interval
Statement_6	Input	Interval
Status_1	Input	Interval
Status_2	Input	Interval
Status_3	Input	Interval
Status_4	Input	Interval
Status_5	Input	Interval
Status_6	Input	Interval

HP Data Partition: Default Partitioning Method, Set Seed = 321, 60% Training and 40% Validation



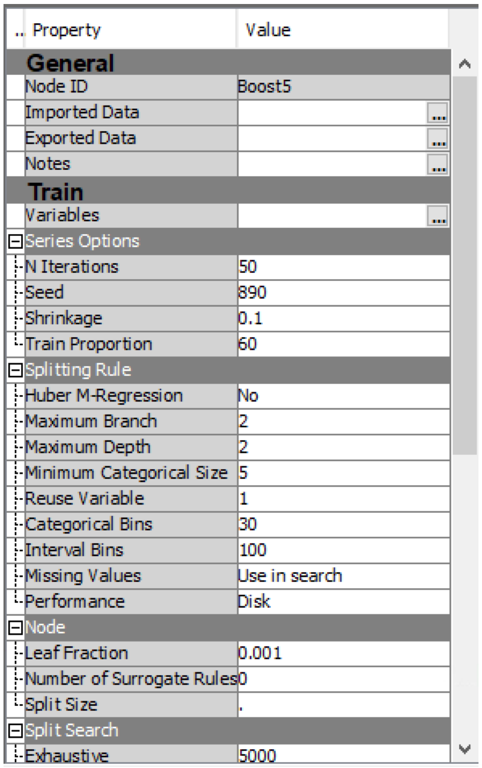
Property	Value
General	
Node ID	HPPart3
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Partitioning Method	Default
Random Seed	321
⊟ Data Set Allocations	
Training	60.0
Validation	40.0
Status	
Create Time	12/1/21 11:01 PM
Run ID	332430f1-2b17-4f21-b129-
Last Error	
Last Status	Complete
Last Run Time	12/1/21 11:02 PM
Run Duration	0 Hr. 0 Min. 8.75 Sec.
Grid Host	

<u>HP Transform</u>: Set Interval Inputs and Interval Targets both equal to Exponential



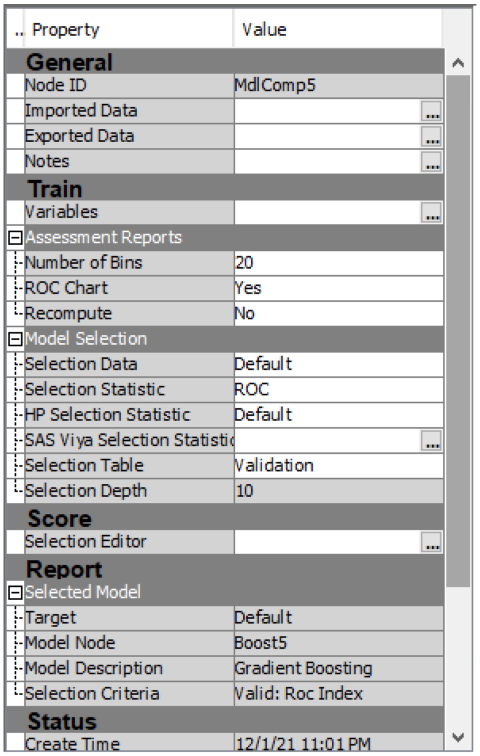
General	
Node ID	HPTrans3
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interval Inputs	Exponential
Interval Targets	Exponential
SAS Code	...
Binning	
Number of Bins	Variables
Missing Values	Separate
Score	
Hide	Yes
Reject	Yes
Report	
Summary Statistics	No
Status	
Create Time	12/1/21 11:01 PM
Run ID	2fe136f5-881f-4f84-a63c-bc
Last Error	
Last Status	Complete
Last Run Time	12/1/21 11:02 PM
Run Duration	0 Hr. 0 Min. 12.20 Sec.
Grid Host	
User-Added Node	No

Gradient Boost (the first one): Set Seed = 890



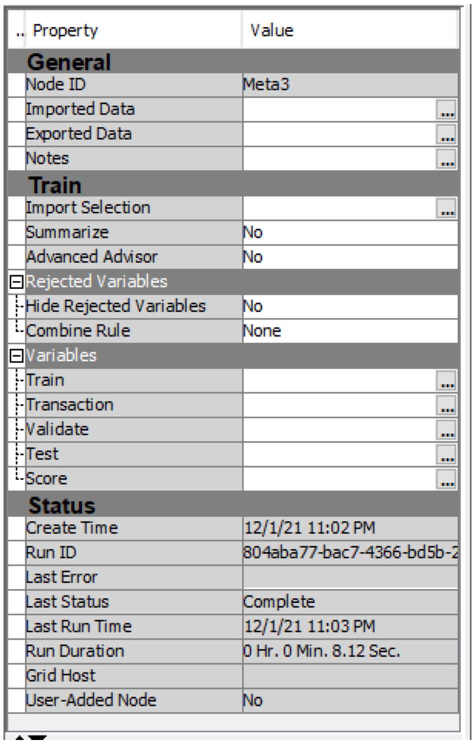
Property	Value
General	
Node ID	Boost5
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟Series Options	
N Iterations	50
Seed	890
Shrinkage	0.1
Train Proportion	60
⊟Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	2
Minimum Categorical Size	5
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Use in search
Performance	Disk
⊟Node	
Leaf Fraction	0.001
Number of Surrogate Rules	0
Split Size	.
⊟Split Search	
Exhaustive	5000

Both Model Comparison Nodes: Set Selection Table = Validation and Selection Statistic = ROC



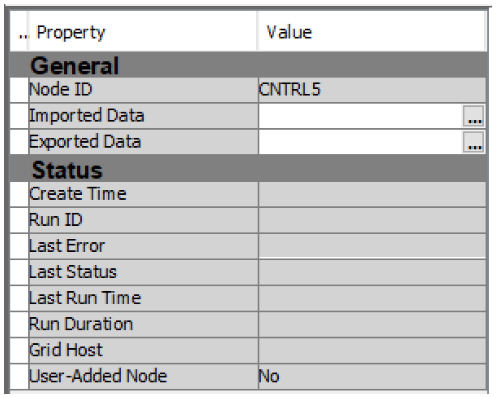
Property	Value
General	
Node ID	MdlComp5
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟Assessment Reports	
Number of Bins	20
ROC Chart	Yes
Recompute	No
⊟Model Selection	
Selection Data	Default
Selection Statistic	ROC
HP Selection Statistic	Default
SAS Viya Selection Statistic	...
Selection Table	Validation
Selection Depth	10
Score	
Selection Editor	...
Report	
⊟Selected Model	
Target	Default
Model Node	Boost5
Model Description	Gradient Boosting
Selection Criteria	Valid: Roc Index
Status	
Create Time	12/1/21 11:01 PM

Metadata: No changes.



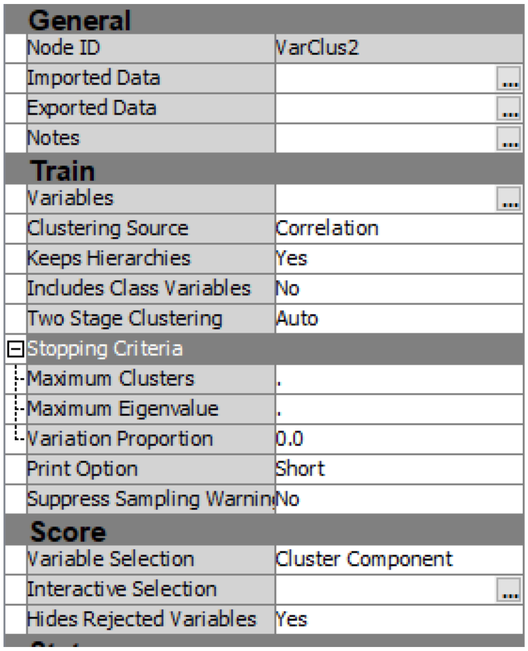
Property	Value
General	
Node ID	Meta3
Imported Data	...
Exported Data	...
Notes	...
Train	
Import Selection	...
Summarize	No
Advanced Advisor	No
⊟ Rejected Variables	
Hide Rejected Variables	No
Combine Rule	None
⊟ Variables	
Train	...
Transaction	...
Validate	...
Test	...
Score	...
Status	
Create Time	12/1/21 11:02 PM
Run ID	804aba77-bac7-4366-bd5b-2
Last Error	
Last Status	Complete
Last Run Time	12/1/21 11:03 PM
Run Duration	0 Hr. 0 Min. 8.12 Sec.
Grid Host	
User-Added Node	No

Both Control Point Nodes: No changes.



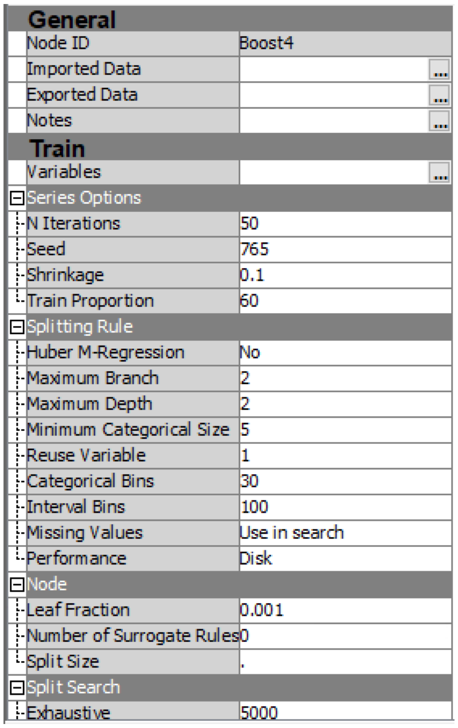
Property	Value
General	
Node ID	CNTRL5
Imported Data	...
Exported Data	...
Status	
Create Time	
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

Variable Clustering: No changes.



General	
Node ID	VarClus2
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Clustering Source	Correlation
Keeps Hierarchies	Yes
Includes Class Variables	No
Two Stage Clustering	Auto
⊟Stopping Criteria	
Maximum Clusters	.
Maximum Eigenvalue	.
Variation Proportion	0.0
Print Option	Short
Suppress Sampling Warnin	No
Score	
Variable Selection	Cluster Component
Interactive Selection	...
Hides Rejected Variables	Yes

Gradient Boost (the second one): Set Seed = 765



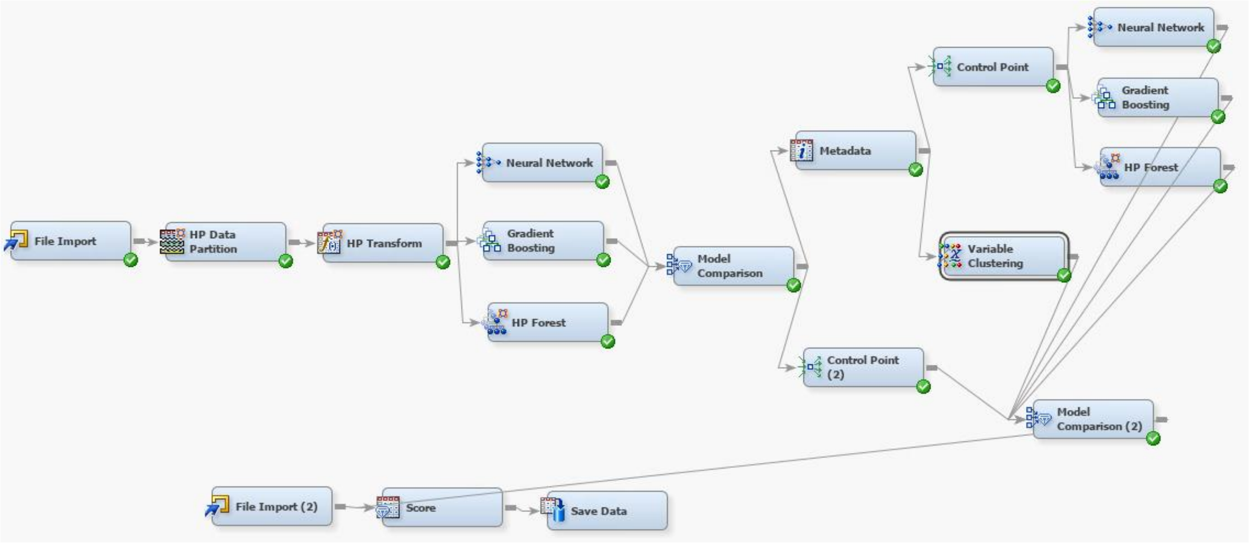
General	
Node ID	Boost4
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟Series Options	
N Iterations	50
Seed	765
Shrinkage	0.1
Train Proportion	60
⊟Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	2
Minimum Categorical Size	5
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Use in search
Performance	Disk
⊟Node	
Leaf Fraction	0.001
Number of Surrogate Rules	0
Split Size	.
⊟Split Search	
Exhaustive	5000

File Import(2): Use the creditDefault_Test_X dataset. Set Role = Score

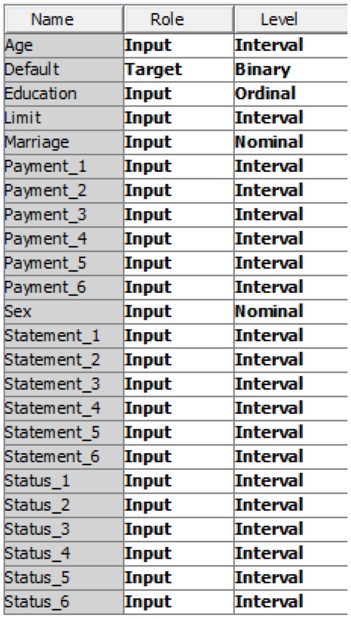
Score: Set Type of Score = Data

Save Data: Set Filename Prefix and Directory. Set File Format = Comma-separated Values (csv)

Model #2:

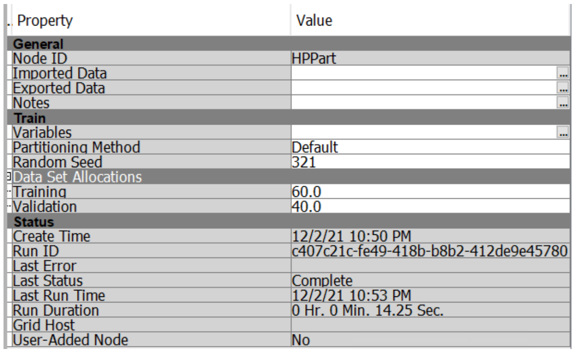


File Import: Use the creditDefault_Train dataset. Set variables as below:



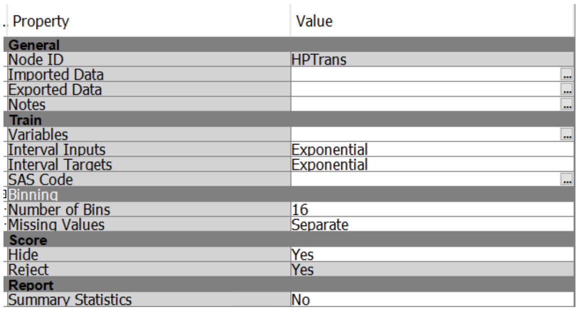
Name	Role	Level
Age	Input	Interval
Default	Target	Binary
Education	Input	Ordinal
Limit	Input	Interval
Marriage	Input	Nominal
Payment_1	Input	Interval
Payment_2	Input	Interval
Payment_3	Input	Interval
Payment_4	Input	Interval
Payment_5	Input	Interval
Payment_6	Input	Interval
Sex	Input	Nominal
Statement_1	Input	Interval
Statement_2	Input	Interval
Statement_3	Input	Interval
Statement_4	Input	Interval
Statement_5	Input	Interval
Statement_6	Input	Interval
Status_1	Input	Interval
Status_2	Input	Interval
Status_3	Input	Interval
Status_4	Input	Interval
Status_5	Input	Interval
Status_6	Input	Interval

HP Data Partition: Default Partitioning Method, Set Seed = 321, 60% Training and 40% Validation



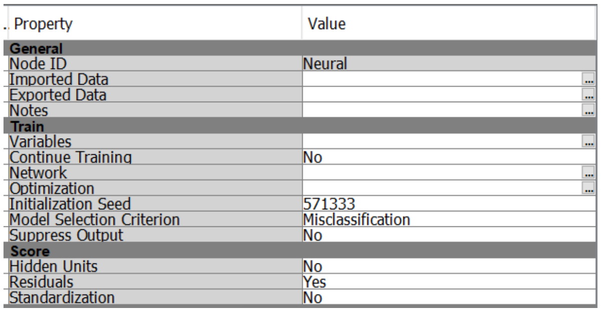
Property	Value
General	
Node ID	HPPart
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Partitioning Method	Default
Random Seed	321
Data Set Allocations	
Training	60.0
Validation	40.0
Status	
Create Time	12/2/21 10:50 PM
Run ID	c407c21c-fe49-418b-b8b2-412de9e45780
Last Error	
Last Status	Complete
Last Run Time	12/2/21 10:53 PM
Run Duration	0 Hr. 0 Min. 14.25 Sec.
Grid Host	
User-Added Node	No

HP Transform: Set Interval Inputs and Interval Targets both equal to Exponential, set Number of Bins = 16



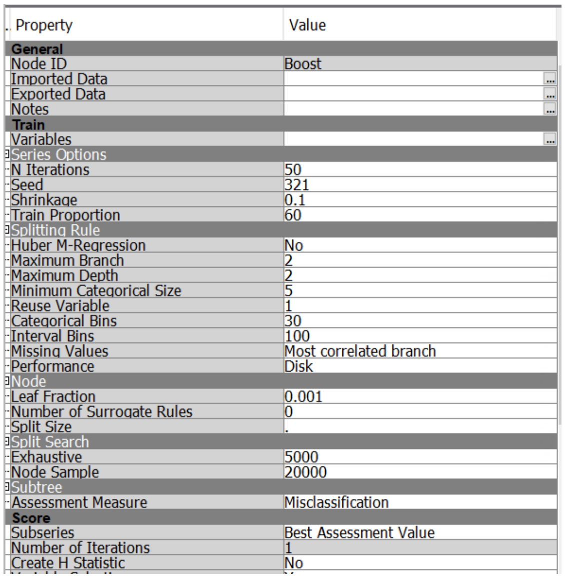
Property	Value
General	
Node ID	HPTrans
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interval Inputs	Exponential
Interval Targets	Exponential
SAS Code	...
Binning	
Number of Bins	16
Missing Values	Separate
Score	
Hide	Yes
Reject	Yes
Report	
Summary Statistics	No

Both Neural Network Nodes: Set initialization seed = 571333, Model Selection Criteria = Misclassification

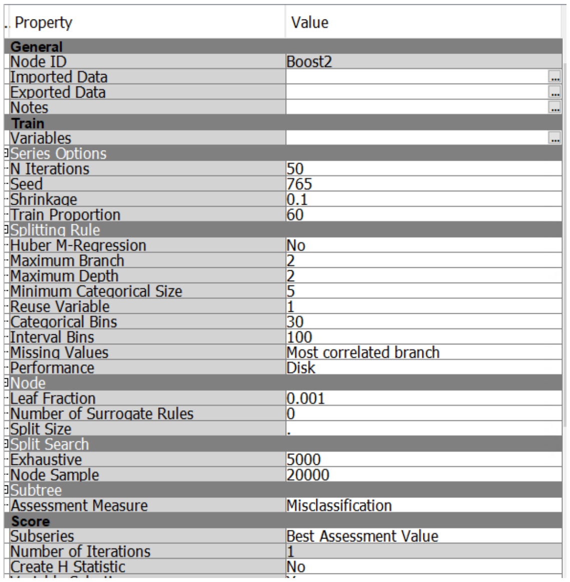


Property	Value
General	
Node ID	Neural
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Continue Training	No
Network	...
Optimization	...
Initialization Seed	571333
Model Selection Criterion	Misclassification
Suppress Output	No
Score	
Hidden Units	No
Residuals	Yes
Standardization	No

Both Gradient Boosting Nodes: Set seed = 321 and 765 separately, Missing Values = Most of correlated branch, Assessment Measure = Misclassification, Leaf fraction = 0.001

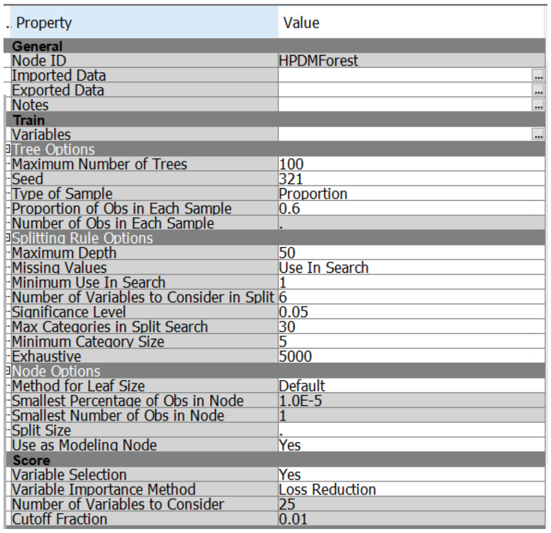


Property	Value
General	
Node ID	Boost
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Series Options	
N Iterations	50
Seed	321
Shrinkage	0.1
Train Proportion	60
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	2
Minimum Categorical Size	5
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Most correlated branch
Performance	Disk
Node	
Leaf Fraction	0.001
Number of Surrogate Rules	0
Split Size	.
Split Search	
Exhaustive	5000
Node Sample	20000
Subtree	
Assessment Measure	Misclassification
Score	
Subseries	Best Assessment Value
Number of Iterations	1
Create H Statistic	No



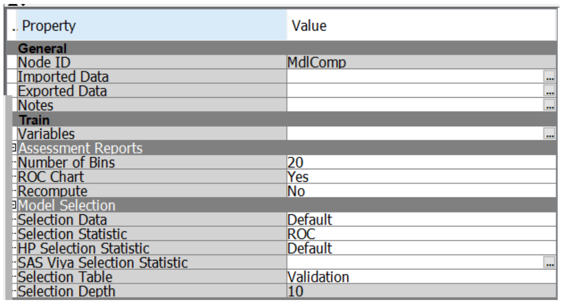
Property	Value
General	
Node ID	Boost2
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Series Options	
N Iterations	50
Seed	765
Shrinkage	0.1
Train Proportion	60
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	2
Minimum Categorical Size	5
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Most correlated branch
Performance	Disk
Node	
Leaf Fraction	0.001
Number of Surrogate Rules	0
Split Size	.
Split Search	
Exhaustive	5000
Node Sample	20000
Subtree	
Assessment Measure	Misclassification
Score	
Subseries	Best Assessment Value
Number of Iterations	1
Create H Statistic	No

<u>Both HP Forest Nodes</u>: Set seed = 321, Number of Variables to Consider in Split Search = 6



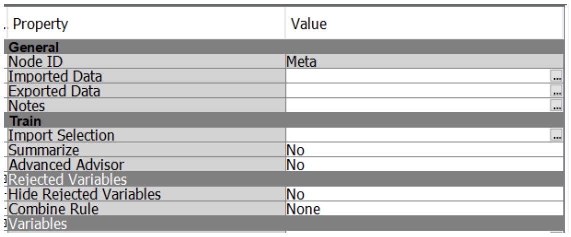
Property	Value
General	
Node ID	HPDMForest
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Tree Options	
Maximum Number of Trees	100
Seed	321
Type of Sample	Proportion
Proportion of Obs in Each Sample	0.6
Number of Obs in Each Sample	.
Splitting Rule Options	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Search	1
Number of Variables to Consider in Split	6
Significance Level	0.05
Max Categories in Split Search	30
Minimum Category Size	5
Exhaustive	5000
Node Options	
Method for Leaf Size	Default
Smallest Percentage of Obs in Node	1.0E-5
Smallest Number of Obs in Node	1
Split Size	.
Use as Modeling Node	Yes
Score	
Variable Selection	Yes
Variable Importance Method	Loss Reduction
Number of Variables to Consider	25
Cutoff Fraction	0.01

<u>Both Model Comparison Nodes</u>: Set Selection Table = Validation and Selection Statistic = ROC



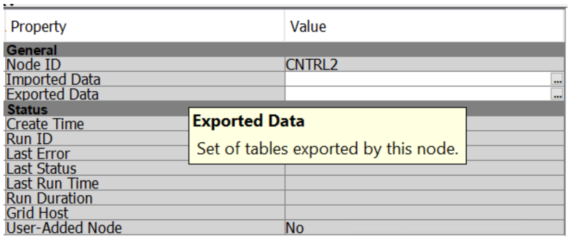
Property	Value
General	
Node ID	MdlComp
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Assessment Reports	
Number of Bins	20
ROC Chart	Yes
Recompute	No
Model Selection	
Selection Data	Default
Selection Statistic	ROC
HP Selection Statistic	Default
SAS Viya Selection Statistic	...
Selection Table	Validation
Selection Depth	10

<u>Metadata</u>: No changes.



Property	Value
General	
Node ID	Meta
Imported Data	...
Exported Data	...
Notes	...
Train	
Import Selection	...
Summarize	No
Advanced Advisor	No
Rejected Variables	
Hide Rejected Variables	No
Combine Rule	None
Variables	

<u>Both Control Point Nodes</u>: No changes.

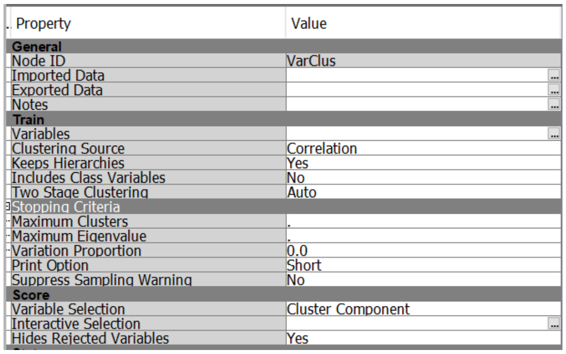


Property	Value
General	
Node ID	CNTRL2
Imported Data	...
Exported Data	...
Status	
Create Time	
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

Exported Data

Set of tables exported by this node.

<u>Variable Clustering</u>: No changes.



Property	Value
General	
Node ID	VarClus
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Clustering Source	Correlation
Keeps Hierarchies	Yes
Includes Class Variables	No
Two Stage Clustering	Auto
Stopping Criteria	
Maximum Clusters	.
Maximum Eigenvalue	.
Variation Proportion	0.0
Print Option	Short
Suppress Sampling Warning	No
Score	
Variable Selection	Cluster Component
Interactive Selection	...
Hides Rejected Variables	Yes

<u>File Import(2)</u>: Use the creditDefault_Test_X dataset. Set Role = Score

<u>Score</u>: Set Type of Score = Data

<u>Save Data</u>: Set Filename Prefix and Directory. Set File Format = Comma-separated Values (csv)

Citation:

We used this source to learn how to use Metadata and Control Point Nodes.

https://github.com/sassoftware/dm-flow/tree/master/EnsembleModeling