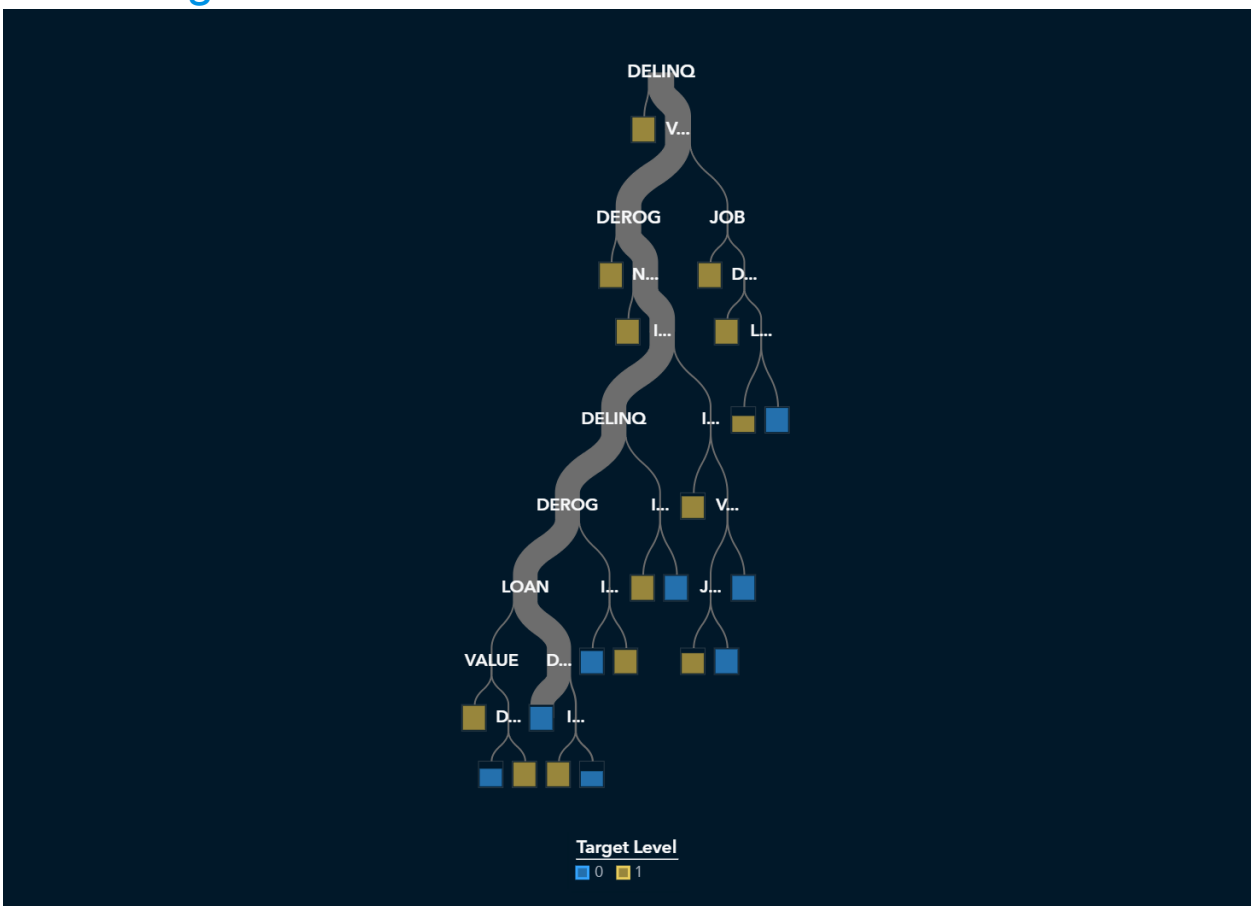
Machine Learning Analytic

"Decision Tree" Results

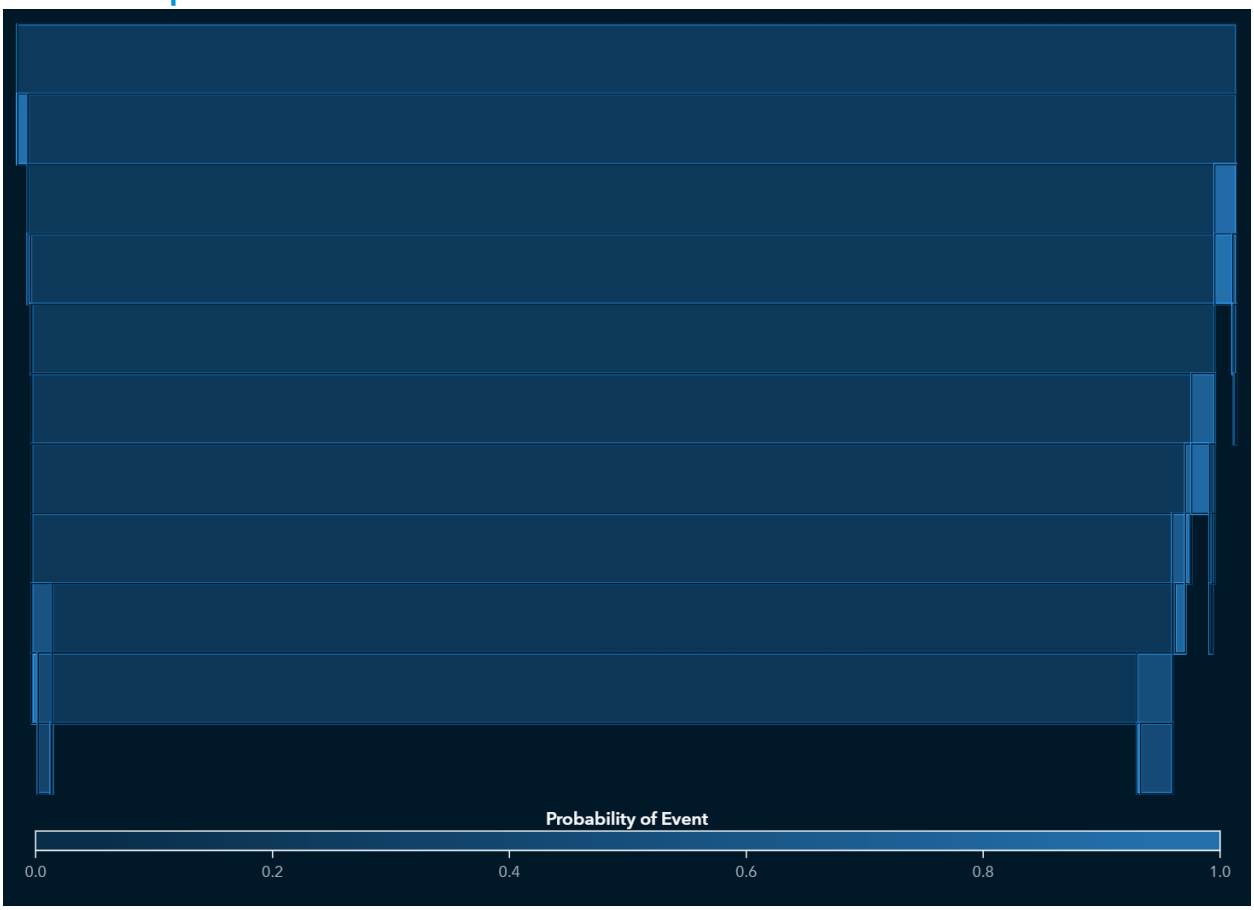
by: jbae7@ncsu.edu

Contents

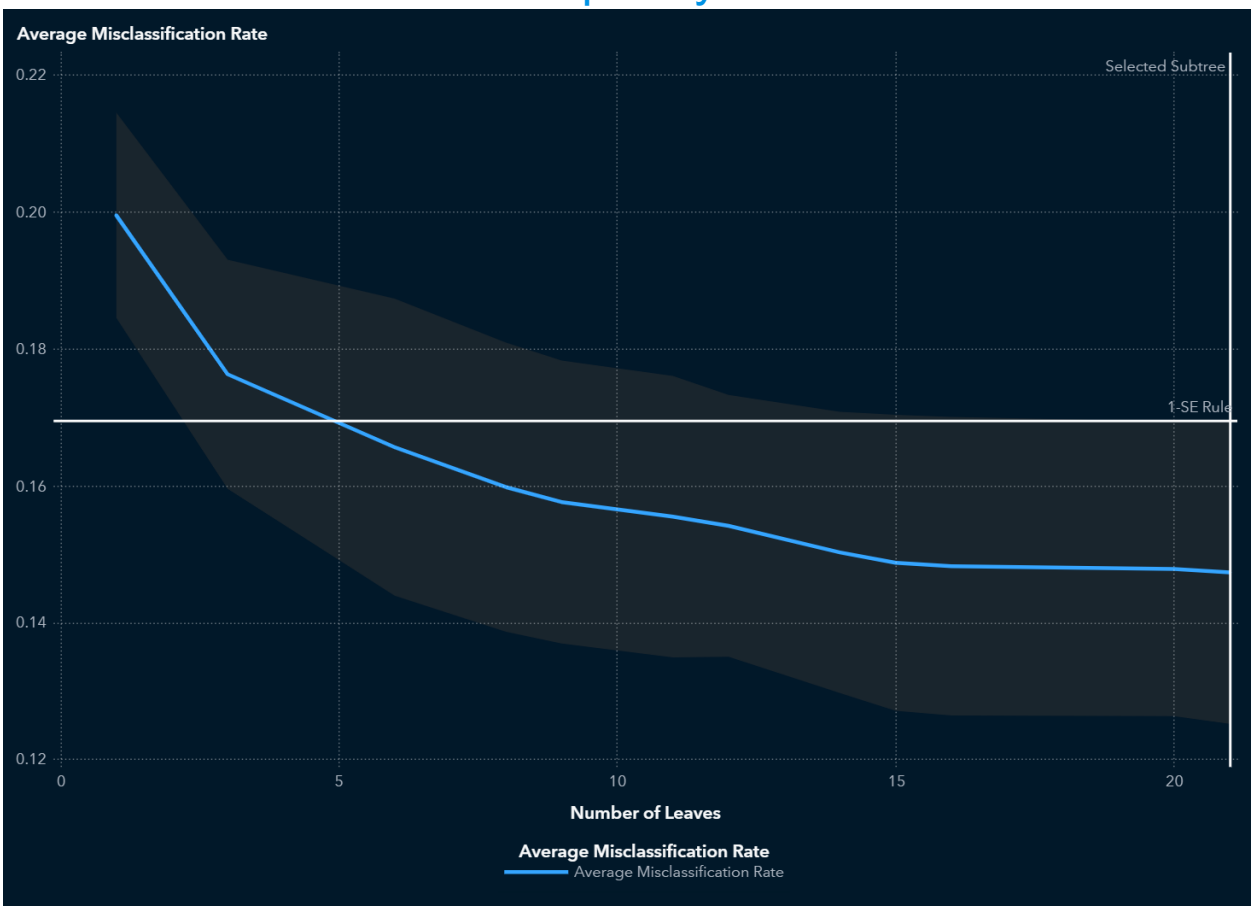
Tree Diagram



Treemap



Cross Validation Cost-Complexity



This plot shows how the average of the misclassification rate across folds changes for subtrees, which are created by cost-complexity pruning of the full decision tree to various numbers of leaves based on cross validation. The band around the line ranges from the average misclassification rate minus one standard error (SE) to the average misclassification rate plus one SE. The reference line for the 1-SE Rule occurs at the value of 0.17, the minimum average misclassification rate plus one SE. When the property for the 1-SE rule is selected, the smallest subtree for which the average misclassification rate is less than this value is used; otherwise, the subtree with the minimum average misclassification rate is used. For this decision tree model, the selected subtree has 21 leaves with an average misclassification rate across folds of 0.147.

Variable Importance

Variable Name	Training Relative Importance	Count	Training Importance
DELINQ	1	3	133.1236
VALUE	0.9919	3	132.0395
IM_DEBTINC	0.7060	2	93.9868
DEROG	0.5083	4	67.6619
LOAN	0.2816	2	37.4862
IM_MORTDUE	0.1587	2	21.1258
IM_CLAGE	0.1309	1	17.4319
JOB	0.0622	2	8.2839
NINQ	0.0611	1	8.1380

Score Inputs

Name	Role	Variable Level	Type
DELINQ	INPUT	NOMINAL	N
DEROG	INPUT	NOMINAL	N
IM_CLAGE	INPUT	INTERVAL	N
IM_DEBTINC	INPUT	INTERVAL	N
IM_MORTDUE	INPUT	INTERVAL	N
JOB	INPUT	NOMINAL	C
LOAN	INPUT	INTERVAL	N
NINQ	INPUT	NOMINAL	N
VALUE	INPUT	INTERVAL	N

Variable Type	Variable Label	Variable Format	Variable Length
double			8
double			8
double			8
double			8
double			8
char			7
double			8
double			8
double			8

Score Outputs

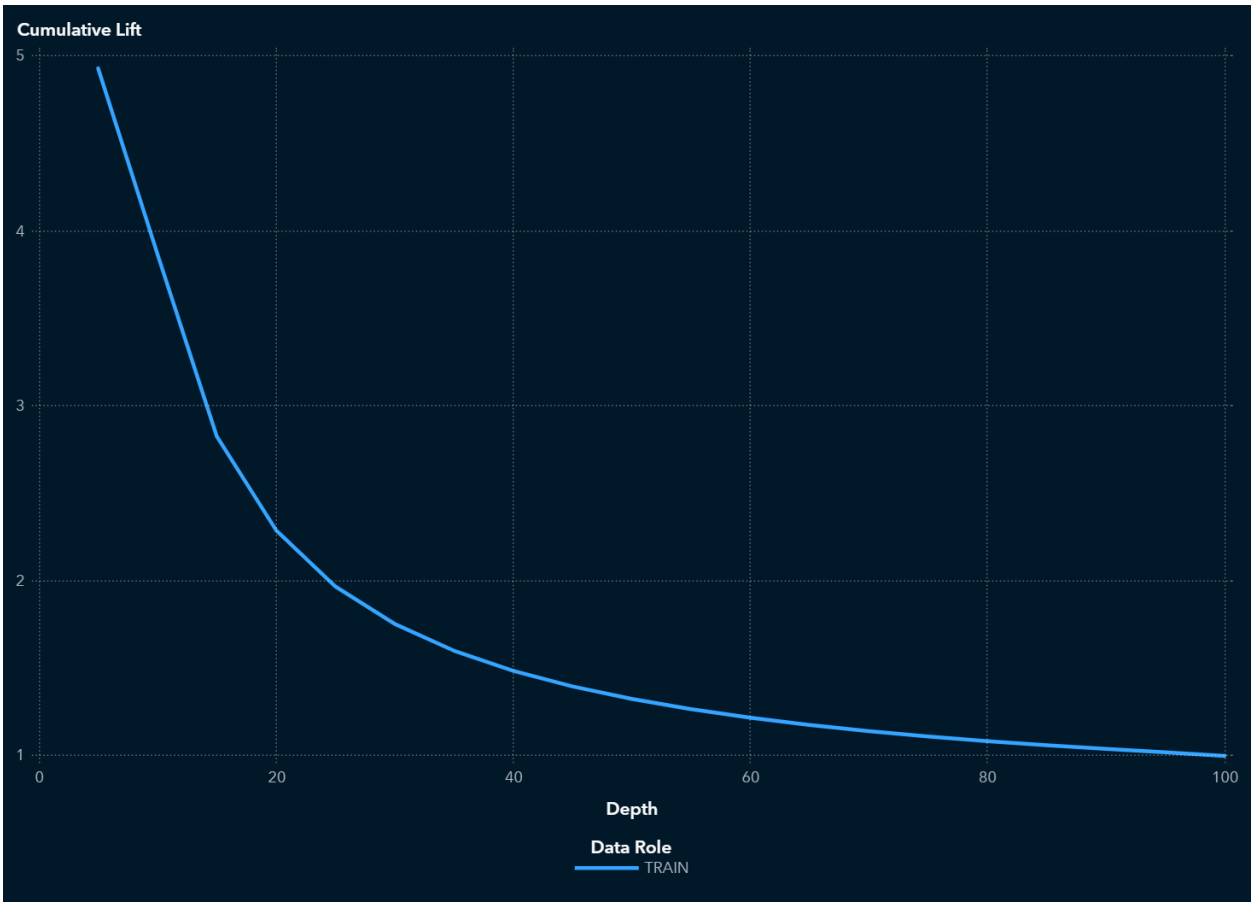
Name	Role	Type	Variable Type
EM_CLASSIFICATION	CLASSIFICATION	C	char
EM_EVENTPROBABILITY	PREDICT	N	double
EM_PROBABILITY	PREDICT	N	double
I_BAD	CLASSIFICATION	C	char
P_BAD0	PREDICT	N	double
P_BAD1	PREDICT	N	double
WARN	ASSESS	C	char

Variable Label	Variable Format	Variable Length	Creator
Predicted for BAD		12	tree
Probability for BAD=1		8	tree
Probability of Classification		8	tree
Into: BAD		32	tree
Predicted: BAD=0		8	tree
Predicted: BAD=1		8	tree
Warnings		4	tree

Function	Creator GUID
CLASSIFICATION	41d8d726-660d-49c5-89fd-78c991fd2dff
PREDICT	41d8d726-660d-49c5-89fd-78c991fd2dff
PREDICT	41d8d726-660d-49c5-89fd-78c991fd2dff

Function	Creator GUID
CLASSIFICATION	41d8d726-660d-49c5-89fd-78c991fd2dff
PREDICT	41d8d726-660d-49c5-89fd-78c991fd2dff
PREDICT	41d8d726-660d-49c5-89fd-78c991fd2dff
ASSESS	41d8d726-660d-49c5-89fd-78c991fd2dff

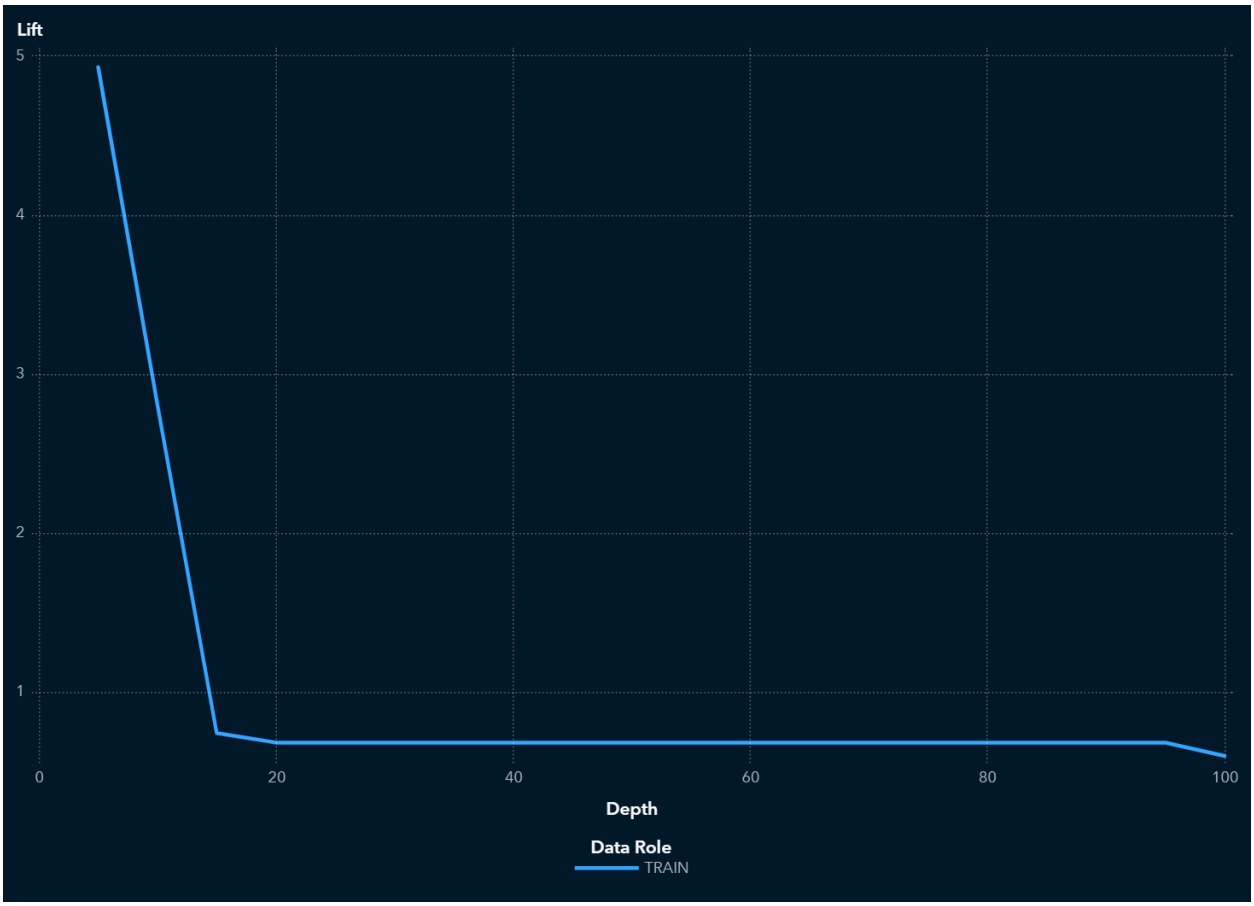
Cumulative Lift



The TRAIN partition has a Cumulative Lift of 3.87 in the 10% quantile (depth of 10) meaning there are 3.87 times more events in the first two quantiles than expected by random (10% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition.

Cumulative lift is calculated by sorting each partition in descending order by the predicted probability of the target event P_BAD1, which represents the predicted probability of the event "1" for the target BAD. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. The cumulative lift for a particular quantile is the ratio of the number of events across all quantiles up to and including the current quantile to the number of events that would be there at random, or equivalently, the ratio of the cumulative response percentage to the baseline response percentage. The cumulative lift at depth 10 includes the top 10% of the data, which is the first 2 quantiles, which would have 10% of the events at random. Thus, cumulative lift measures how much more likely it is to observe an event in the quantiles than by selecting observations at random.

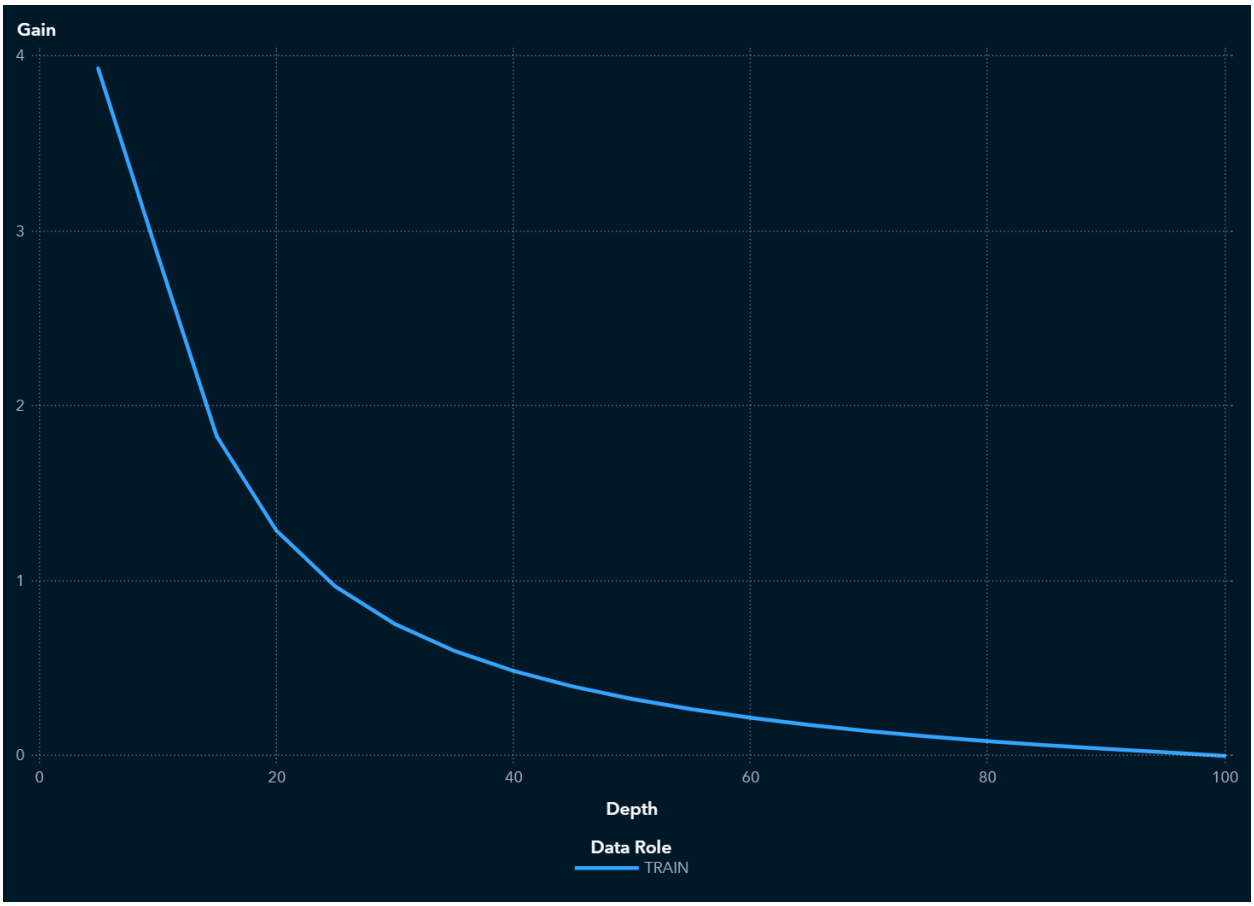
Lift



The TRAIN partition has a Lift of 4.93 in the 5% quantile (depth of 5) meaning there are 4.93 times more events in that quantile than expected by random (5% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition.

Lift is calculated by sorting each partition in descending order by the predicted probability of the target event P_BAD1, which represents the predicted probability of the event "1" for the target BAD. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. Lift is the ratio of the number of events in that quantile to the number of events that would be there at random, or equivalently, the ratio of the response percentage to the baseline response percentage. With 20 quantiles, it is expected that 5% of the events occur in each quantile. Thus, Lift measures how much more likely it is to observe an event in each quantile than by selecting observations at random.

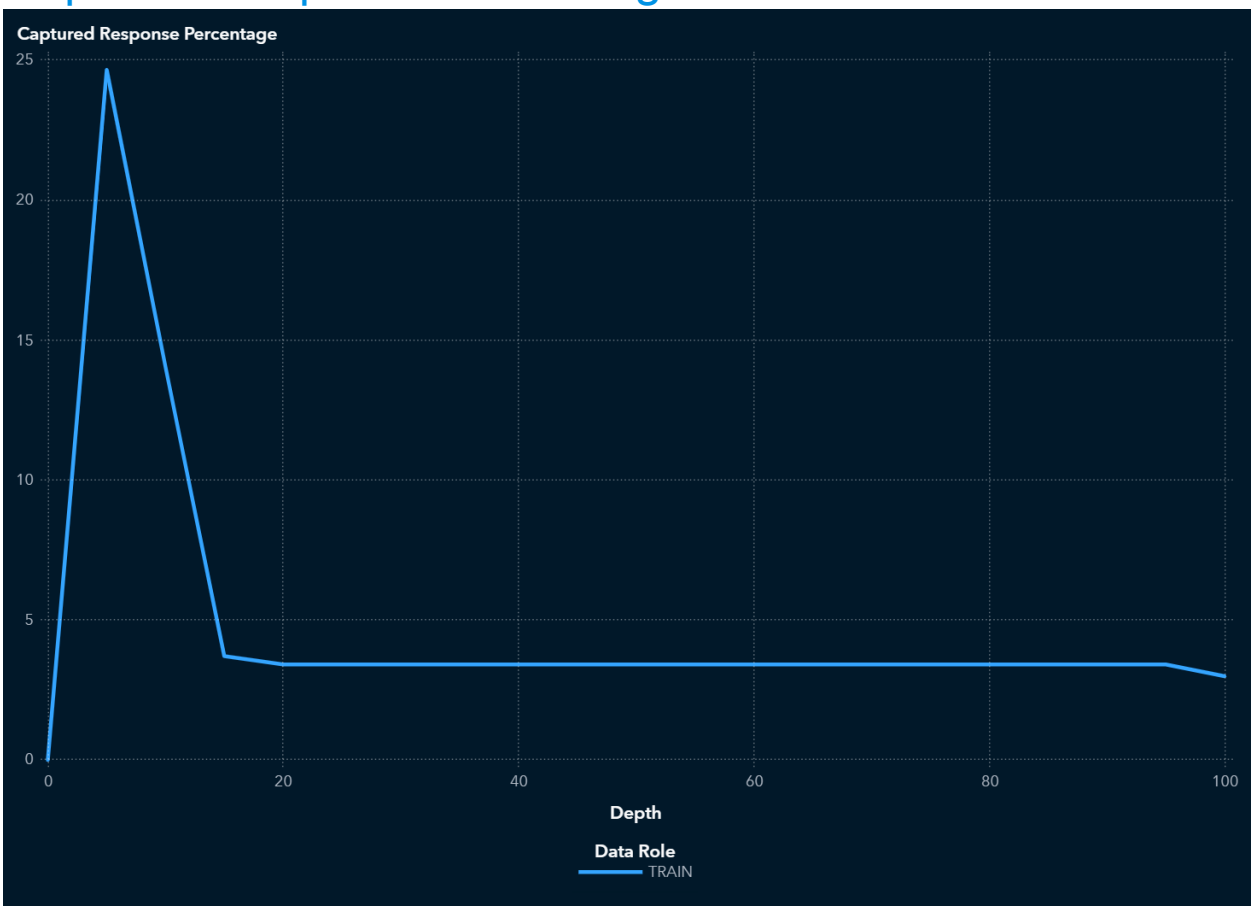
Gain



The TRAIN partition has a Gain of 2.9 at the 10% quantile (depth of 10). Because this value is greater than 0, it is better to use your model to identify responders than no model, based on the selected partition. The best possible value of Gain for this partition at depth 10 is 4.01.

Gain is calculated by sorting each partition in descending order by the predicted probability of the target event P_BAD1, which represents the predicted probability of the event "1" for the target BAD. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. Gain is a cumulative measure for the quantiles up to an including the current one and is calculated as (number of events in the quantiles) / (number of events expected by random) - 1. With 20 quantiles, it is expected that 5% of the events occur in each quantile. Note that the value of Gain is the same as the value of Cumulative Lift - 1. If the value of Gain is greater than 0, then your model is better at identifying events than using no model.

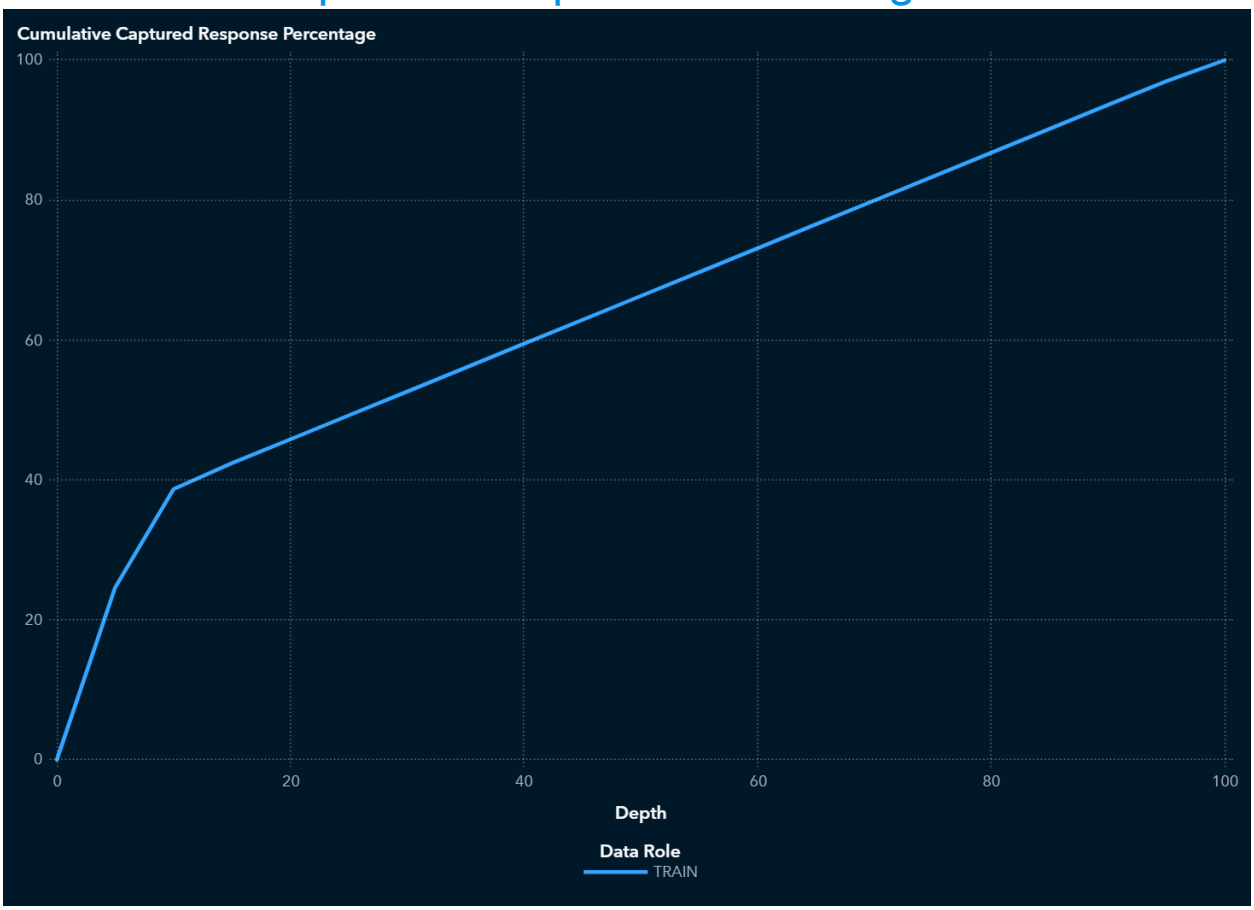
Captured Response Percentage



At the 5% quantile (depth of 5), the TRAIN partition has a Captured response percentage of 24.7 (compared to the expected value of 5 for no model). The best possible value of Captured response percentage for this partition at depth 5 is 25.06.

Captured response percentage is calculated by sorting each partition in descending order by the predicted probability of the target event P_BAD1, which represents the predicted probability of the event "1" for the target BAD. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. Captured response percentage is the percentage of the total number of events that are in that quantile. With no model, it is expected that 5% of the events are in each quantile.

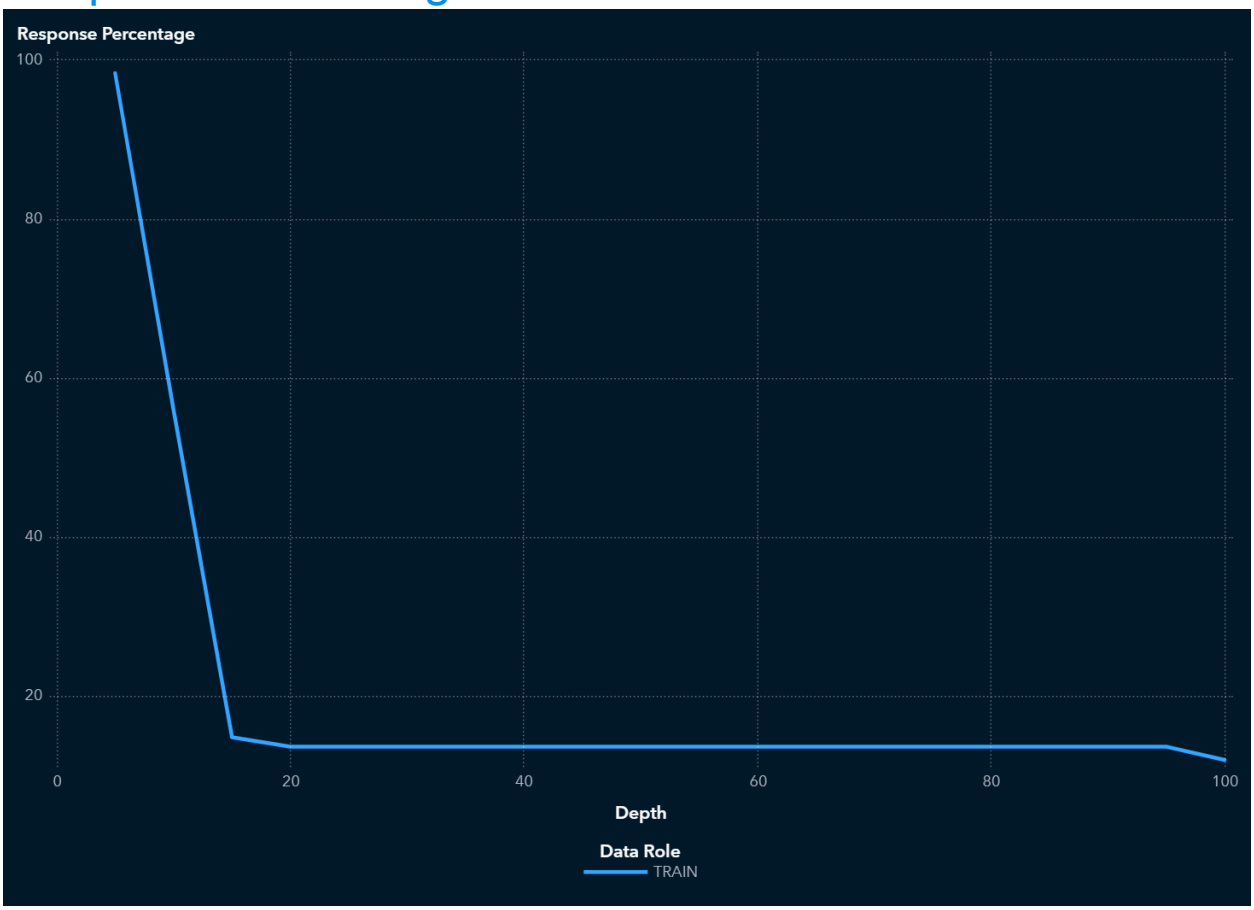
Cumulative Captured Response Percentage



In the top 10% of the data (depth 10), the TRAIN partition has a Cumulative captured response percentage of 38.7 (compared to the expected value of 10 for no model). The best possible value of Cumulative captured response percentage for this partition at depth 10 is 50.13.

Cumulative captured response percentage is calculated by sorting each partition in descending order by the predicted probability of the target event P_BAD1, which represents the predicted probability of the event "1" for the target BAD. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. The cumulative captured response percentage for a particular quantile is the percentage of the total number of events that are in the quantiles up to and including the current quantile. With no model, it is expected that 5% of the events are in each quantile, so the cumulative captured response percentage at depth 10 would be 10%.

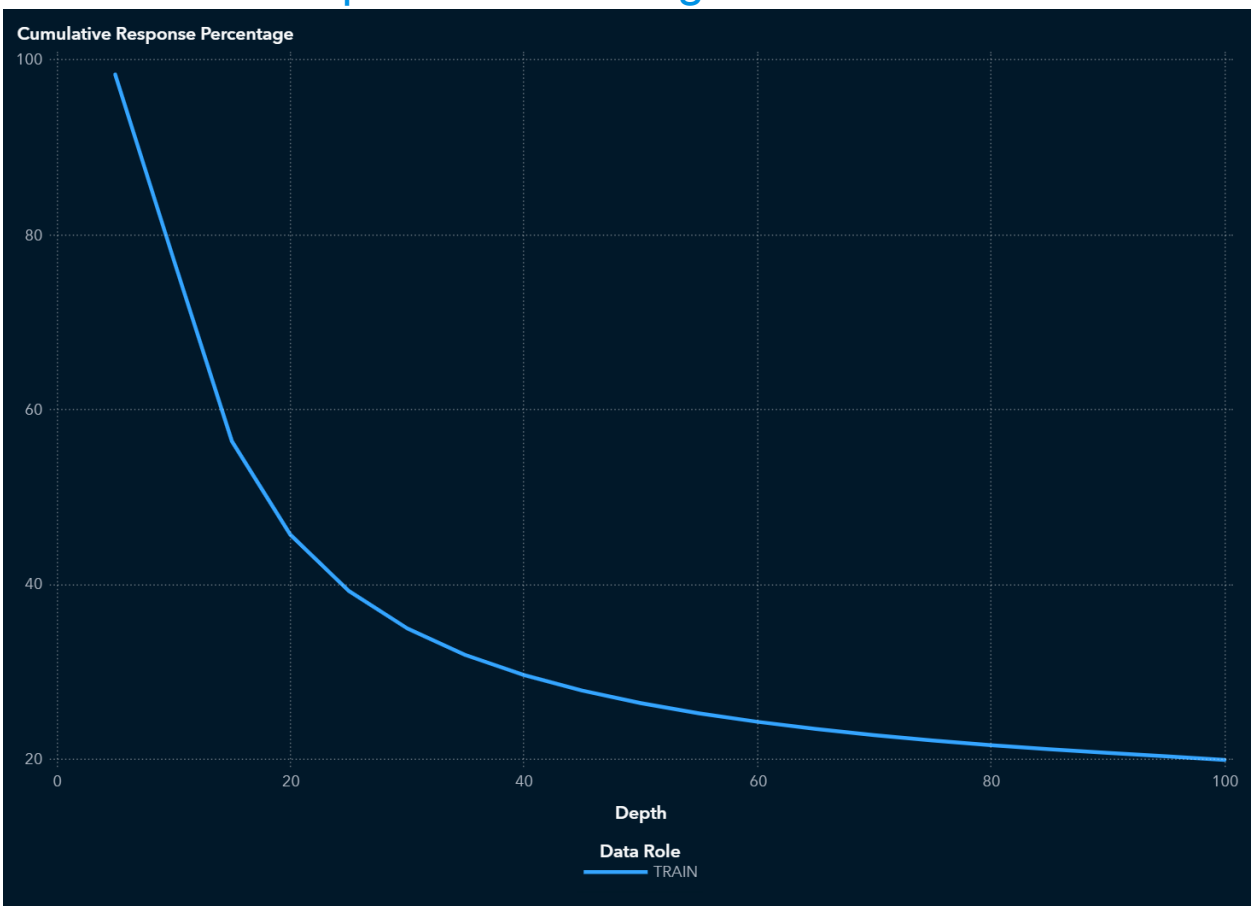
Response Percentage



At the 5% quantile (depth of 5), the TRAIN partition has a Response percentage of 98.4. The best possible value of Response percentage for this partition at depth 5 is 100.

Response percentage is calculated by sorting each partition in descending order by the predicted probability of the target event P_BAD1, which represents the predicted probability of the event "1" for the target BAD. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. Response percentage is the percentage of observations that are events in that quantile. With no model, it is expected that the response percentage is constant across quantiles, 100*overall-event-rate. This is also called the baseline response percentage.

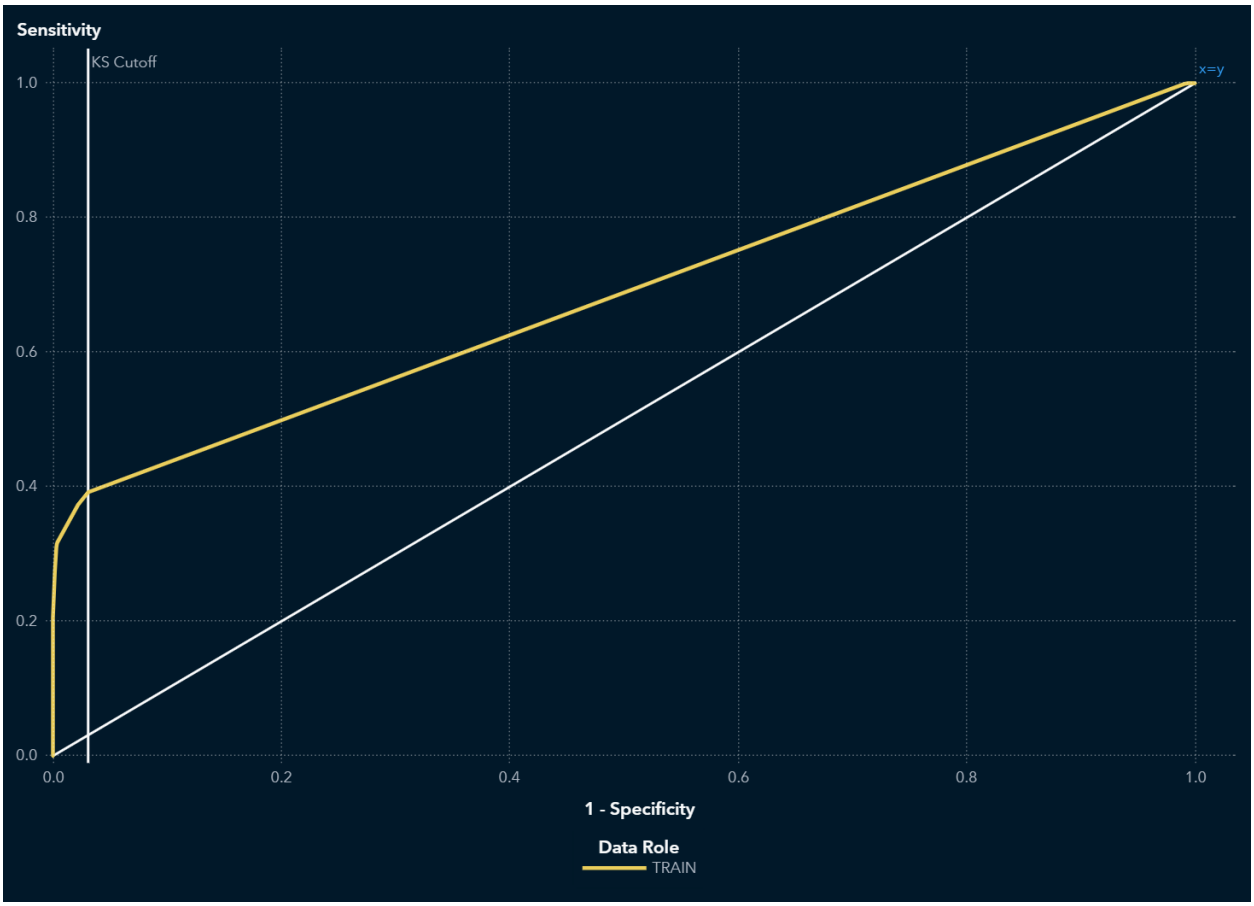
Cumulative Response Percentage



In the top 10% of the data (depth 10), the TRAIN partition has a Cumulative response percentage of 77.2. The best possible value of Cumulative response percentage for this partition at depth 10 is 100.

Cumulative response percentage is calculated by sorting in descending order each partition of the data by the predicted probability of the target event P_BAD1, which represents the predicted probability of the event "1" for the target BAD. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. The cumulative response percentage for a particular quantile is the percentage of observations that are events in the quantiles up to and including the current quantile. With no model, it is expected that the response percentage is constant across quantiles, 100*overall-event-rate. This is also called the baseline response percentage.

ROC



The ROC curve is a plot of sensitivity (the true positive rate) against 1-specificity (the false positive rate), which are both measures of classification based on the confusion matrix. These measures are calculated at various cutoff values. To help identify the best cutoff to use when scoring your data, the KS Cutoff reference line is drawn at the value of 1-specificity where the greatest difference between sensitivity and 1-specificity is observed for the TRAIN partition. The KS Cutoff line is drawn at the cutoff value 0.14, where the 1-specificity value is 0.031 and the sensitivity value is 0.392.

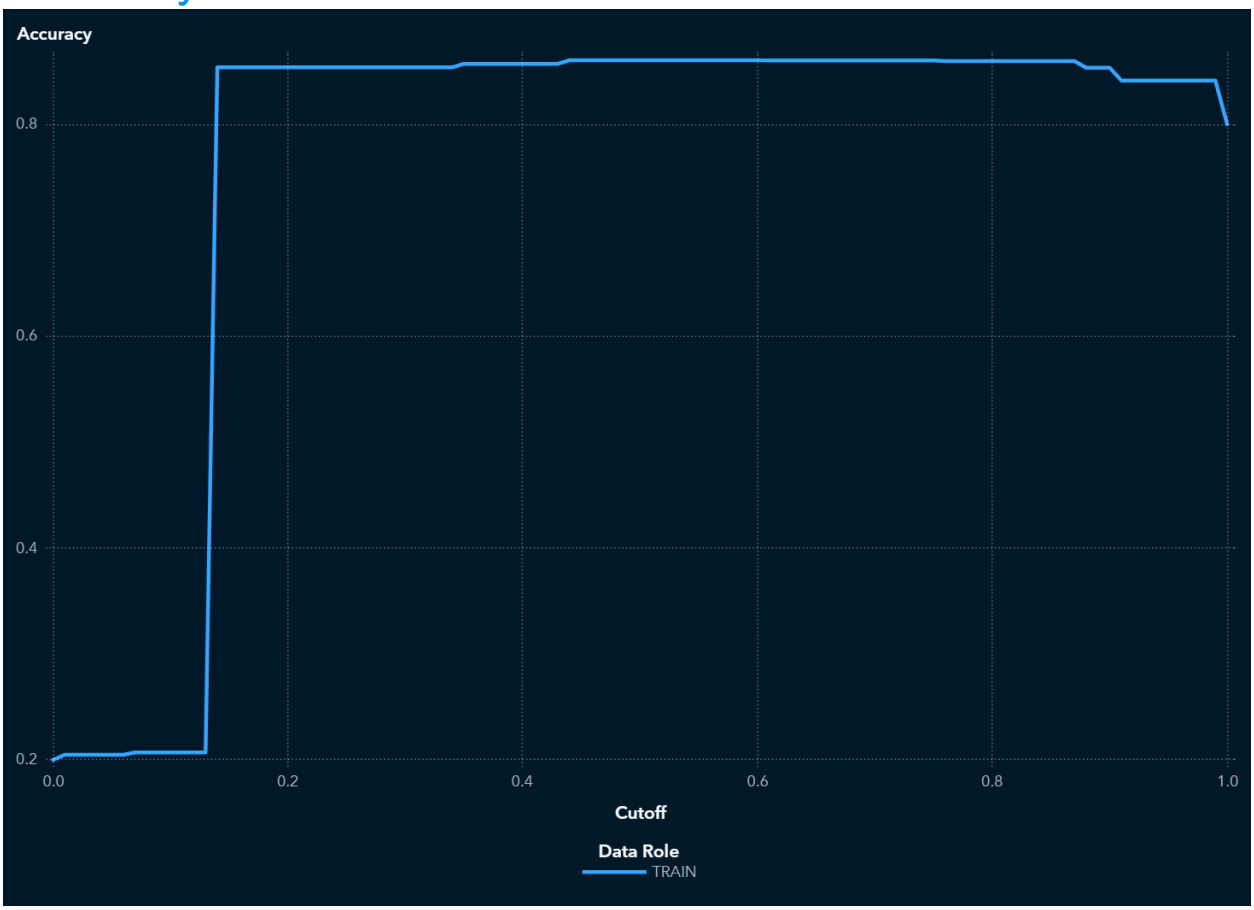
Cutoff values range from 0 to 1, inclusive, in increments of 0.01. At each cutoff value, the predicted target classification is determined by whether P_BAD1, which is the predicted probability of the event "1" for the target BAD, is greater than or equal to the cutoff value. When P_BAD1 is greater than or equal to the cutoff value, then the predicted classification is the event, otherwise it is a non-event.

The confusion matrix for each cutoff value contains four cells that display the true positives for events that are correctly classified (TP), false positives for non-events that are classified as events (FP), false negatives for events that are classified as non-

events (FN), and true negatives for non-events that are classified as non-events (TN). True negatives include non-event classifications that specify a different non-event. Sensitivity is calculated as TP / (TP + FN). Specificity, the true negative rate, is calculated as TN / (TN + FP), so 1-specificity is FP / (TN + FP). The values of sensitivity and 1-specificity are plotted at each cutoff value.

A ROC curve that rapidly approaches the upper-left corner of the graph, where the difference between sensitivity and 1-specificity is the greatest, indicates a more accurate model. A diagonal line where sensitivity = 1-specificity indicates a random model.

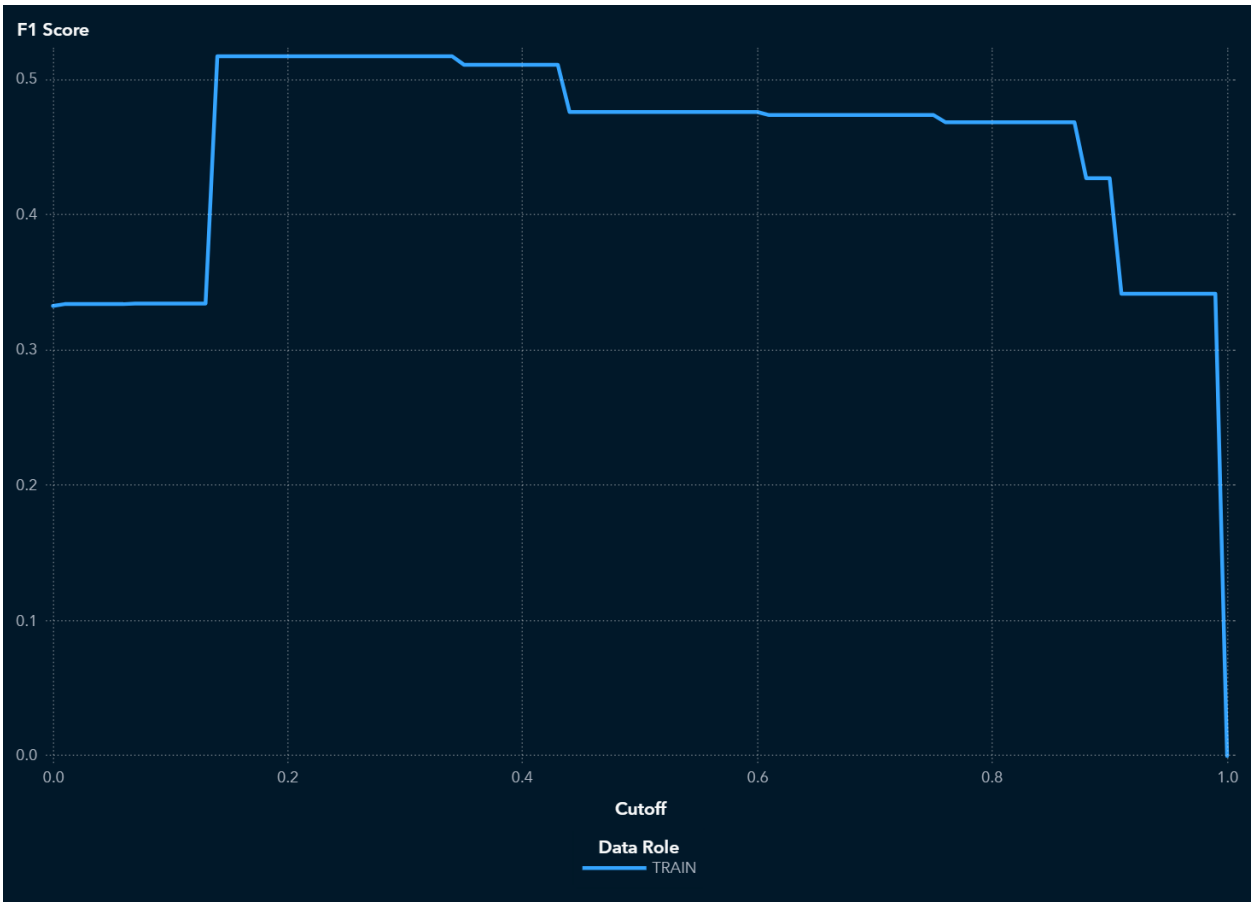
Accuracy



For this model, the accuracy in the TRAIN partition at the cutoff of 0.5 is 0.861.

Accuracy is the proportion of observations that are correctly classified as either an event or non-event, calculated at various cutoff values. Cutoff values range from 0 to 1, inclusive, in increments of 0.01. At each cutoff value, the predicted target classification is determined by whether P_BAD1, which is the predicted probability of the event "1" for the target BAD, is greater than or equal to the cutoff value. When P_BAD1 is greater than or equal to the cutoff value, then the predicted classification is the event, otherwise it is a non-event. When the predicted classification and the actual classification are both events (true positives) or both non-events (true negatives), the observation is correctly classified. If the predicted classification and actual classification disagree, then the observation is incorrectly classified. Accuracy is calculated as (true positives + true negatives) / (total observations).

F1 Score



For this model, the F1 score in the TRAIN partition at the cutoff of 0.5 is 0.476.

The F1 score combines the measures of precision and recall (or sensitivity), which are measures of classification based on the confusion matrix that are calculated at various cutoff values. Cutoff values range from 0 to 1, inclusive, in increments of 0.01. At each cutoff value, the predicted target classification is determined by whether P_BAD1, which is the predicted probability of the event "1" for the target BAD, is greater than or equal to the cutoff value. When P_BAD1 is greater than or equal to the cutoff value, then the predicted classification is the event, otherwise it is a non-event.

The confusion matrix for each cutoff value contains four cells that display the true positives for events that are correctly classified (TP), false positives for non-events that are classified as events (FP), false negatives for events that are classified as non-events (FN), and true negatives for non-events that are classified as non-events (TN). True negatives include non-event classifications that specify a different non-event.

Precision is calculated as TP / (TP + FP), and recall (or sensitivity) is calculated as TP /

(TP + FN). The F1 score is calculated as 2*Precision*Recall / (Precision + Recall), which is the harmonic mean of Precision and Recall. Larger F1 scores indicate a more accurate model.

Fit Statistics

Target Name	Data Role	Number of Observations	Average Squared Error
BAD	TRAIN	5,960	0.1161

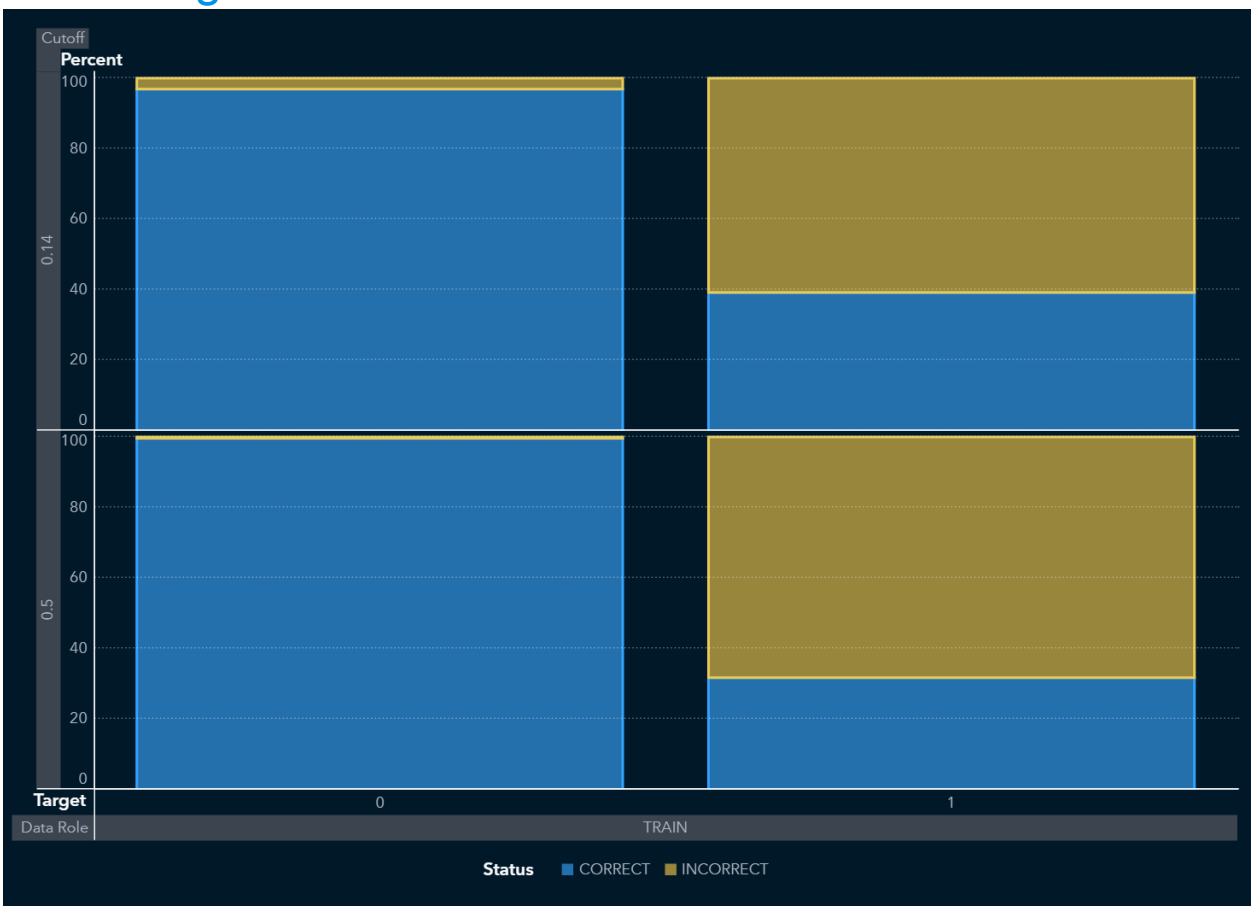
Divisor for ASE	Root Average Squared Error	Misclassification Rate	Multi-Class Log Loss
5,960	0.3408	0.1393	0.3881

KS (Youden)	Area Under ROC	Gini Coefficient	Gamma
0.3611	0.6876	0.3751	0.9026

Tau	KS Cutoff	KS at User-Specified Cutoff	Misclassification Rate at KS Cutoff (Event)
0.1198	0.1400	0.3133	0.1460

Misclassification Rate (Event)
0.1393

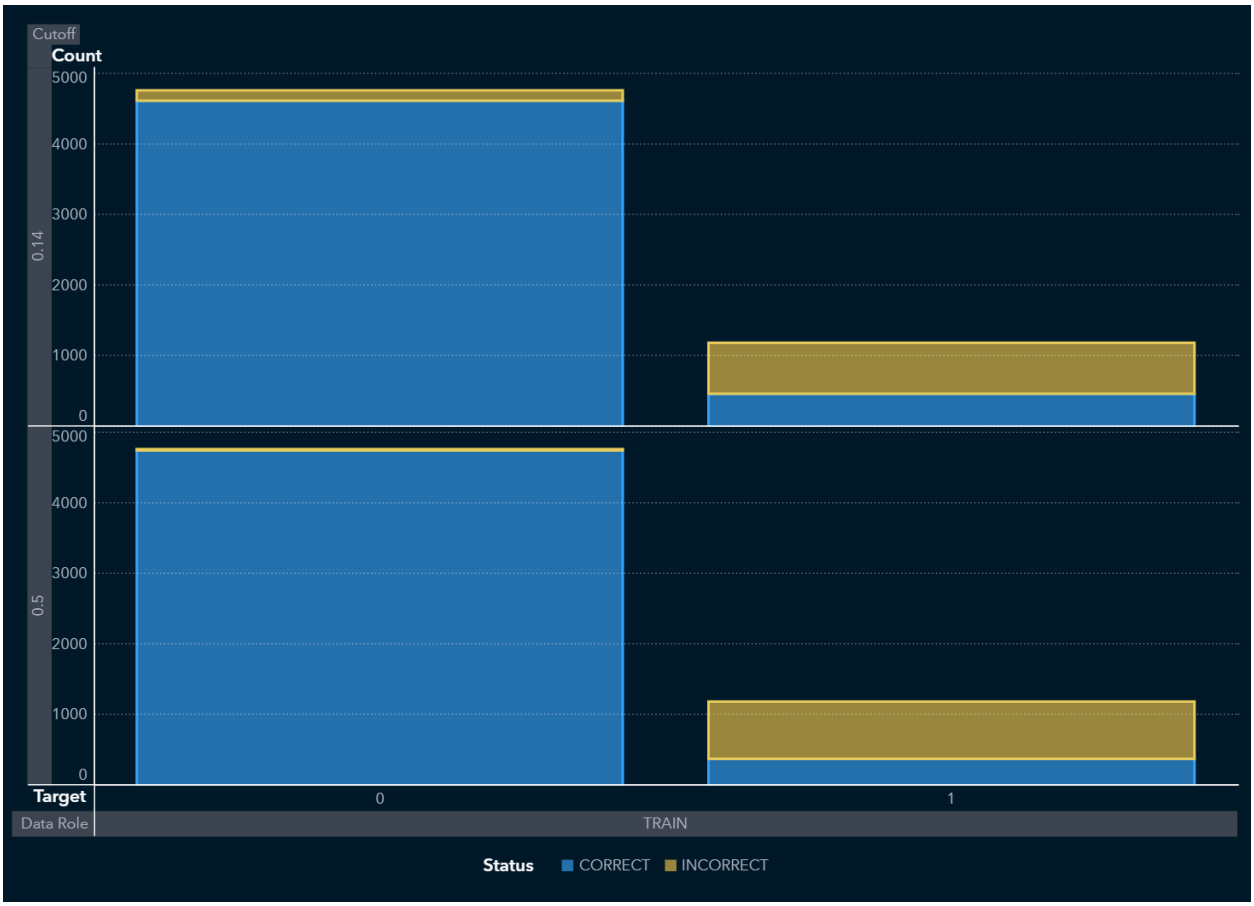
Percentage Plot



The Event Classification report is a visual representation of the confusion matrix at various cutoff values for each partition. The classification cutoffs used in the plot are the default (0.5) and the KS cutoff value 0.14 for the TRAIN partition.

For this data, for the bar corresponding to the event level of BAD, " 1", the segment of the bar colored as "CORRECT" corresponds to true positives.

Count Plot



The Event Classification report is a visual representation of the confusion matrix at various cutoff values for each partition. The classification cutoffs used in the plot are the default (0.5) and the KS cutoff value 0.14 for the TRAIN partition.

For this data, for the bar corresponding to the event level of BAD, " 1", the segment of the bar colored as "CORRECT" corresponds to true positives.

Table

Cutoff	Cutoff Source	Target Name	Response
0.1400	KS	BAD	CORRECT
0.1400	KS	BAD	INCORRECT
0.1400	KS	BAD	CORRECT
0.1400	KS	BAD	INCORRECT
0.5000	Default	BAD	CORRECT
0.5000	Default	BAD	INCORRECT
0.5000	Default	BAD	CORRECT
0.5000	Default	BAD	INCORRECT

Event	Value	Training Frequency	Validation Frequency
1	True Positive	466	
1	False Negative	723	
0	True Negative	4,624	
0	False Positive	147	
1	True Positive	377	
1	False Negative	812	
0	True Negative	4,753	
0	False Positive	18	

Test Frequency	Training Percentage	Validation Percentage	Test Percentage
	39.1926		
	60.8074		
	96.9189		
	3.0811		
	31.7073		
	68.2927		
	99.6227		

Test Frequency	Training Percentage	Validation Percentage	Test Percentage
	0.3773		

Properties

Property Name	Property Value
alpha	0.2000
atAppendLookup	false
atCreateHistory	false
atHistoryLibUri	
atHistoryTblName	
atLeaveAutotuneOn	false
atLookupTableUri	
atMaxBayes	100
atMaxEval	50
atMaxIter	5
atMaxTime	60
atObjectiveInt	ASE
atObjectiveNom	KS
atPopSize	10
atSampleSize	50
atSearchMethod	GA
atTrainProp	0.7000
atUpdateProperties	false
atUseLookup	false
atValidFold	5
atValidMethod	PARTITION
atValidProp	0.3000
atgrowcrit	true
atgrowcritValsi	VARIANCE FTEST CHAID
atgrowcritValsn	ENTROPY CHAID IGR GINI CHISQUARE

Property Name	Property Value
atleafSize	false
atleafSizeInit	5
atleafSizeLB	1
atleafSizeUB	100
atmaxdepth	true
atmaxdepthInit	10
atmaxdepthLB	1
atmaxdepthUB	19
atnumbin	true
atnumbinInit	50
atnumbinLB	20
atnumbinUB	200
autotune_enabled	false
binaryProbCutoff	0.5000
bonferroni	false
ccAlpha	0
codeLocation	mlearning
confidence	0.2500
criterionMethod	IGR
cvccFolds	10
dataMiningVersion	V2024.03
embeddedBarChart	true
exactPctlLift	true
explainFidelity	false
explainInfo	false
fullDatasetReconstitution	false
hLeafSize	5
iCriterionMethod	VARIANCE

Property Name	Property Value
icePlots	false
inodeColor	AVERAGE
intBinMethod	QUANTILE
intervalBins	50
maxBranch	2
maxCategories	128
maxDepth	10
maxNumShapVars	20
minUseinsearch	1
missingValue	USEINSEARCH
nBins	50
nPLeaves	1
nodeColor	PROBEVENT
pdNumImportantInputs	5
pdObsSamples	1,000
pdPlots	false
performKernelShap	false
performLime	false
performVI	false
pruningMethod	COSTCOMPLEXITY
rapidGrowth	false
reportingOnly	false
seRule	false
seed	12,345
seedId	12,345
selMethod	AUTOMATIC
specifyRows	RANDOM
templateRevision	4

Property Name	Property Value
train	true
truncateLl	5
truncateUl	95
useVarOnce	false
userProbCutoff	false

Output

The SAS System

The TREESPLIT Procedure

Model Information	
Split Criterion	IGR
Pruning Method	Cost Complexity
Max Branches per Node	2
Max Tree Depth	10
Tree Depth Before Pruning	10
Tree Depth After Pruning	10
Number of Leaves Before Pruning	30
Number of Leaves After Pruning	21

	Training
Number of Observations Read	5960
Number of Observations Used	5960

The SAS System

The TREESPLIT Procedure

10-Fold Cross Validation Assessment of Pruning Parameter

			Misclassification Rate			
N Leaves	Pruning Parameter		Min	Avg	Standard Error	Max
21	9.8E-11	*	0.1087	0.1474	0.0221	0.1820
20	0.000237	.	0.1138	0.1479	0.0215	0.1820
16	0.000628	.	0.1154	0.1483	0.0218	0.1839
15	0.00126	.	0.1154	0.1488	0.0216	0.1839
14	0.00142	.	0.1239	0.1503	0.0205	0.1839
12	0.00174	.	0.1290	0.1542	0.0191	0.1876
11	0.00210	.	0.1290	0.1555	0.0206	0.1951
9	0.00234	.	0.1307	0.1577	0.0207	0.1989
8	0.00294	.	0.1324	0.1599	0.0211	0.1989
6	0.00395	.	0.1358	0.1657	0.0217	0.2008
3	0.00735	.	0.1511	0.1764	0.0167	0.1994
1	1.1913	.	0.1698	0.1996	0.0150	0.2201

* Selected pruning parameter

The SAS System

The TREESPLIT Procedure

Fit Statistics for Selected Tree		
	Number of Leaves	Misclassification Rate
Training	21	0.1393

Variable Importance			
	Training		
Variable	Importance	Relative Importance	Count
DELINQ	133.12	1.0000	3
VALUE	132.04	0.9919	3
IM_DEBTINC	93.9868	0.7060	2
DEROG	67.6619	0.5083	4
LOAN	37.4862	0.2816	2
IM_MORTDUE	21.1258	0.1587	2
IM_CLAGE	17.4319	0.1309	1
JOB	8.2839	0.0622	2
NINQ	8.1380	0.0611	1

The SAS System

The TREESPLIT Procedure

Predicted Probability Variables	
BAD	Variable
1	P_BAD1
0	P_BAD0

Predicted Target Variable	
Level Index	Variable
	I_BAD