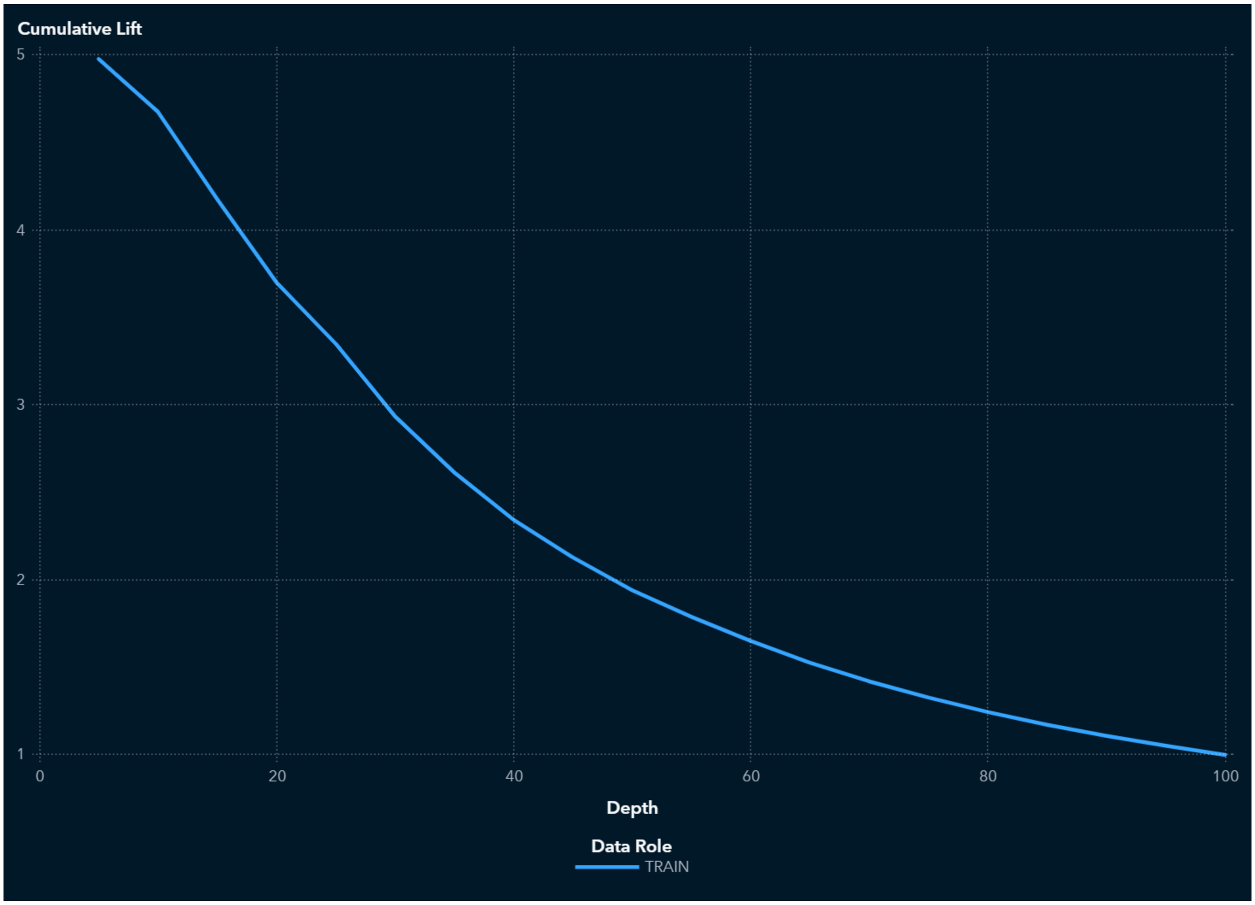
Machine Learning Analytic

"Ensemble" Results

by: jbae7@ncsu.edu

Contents

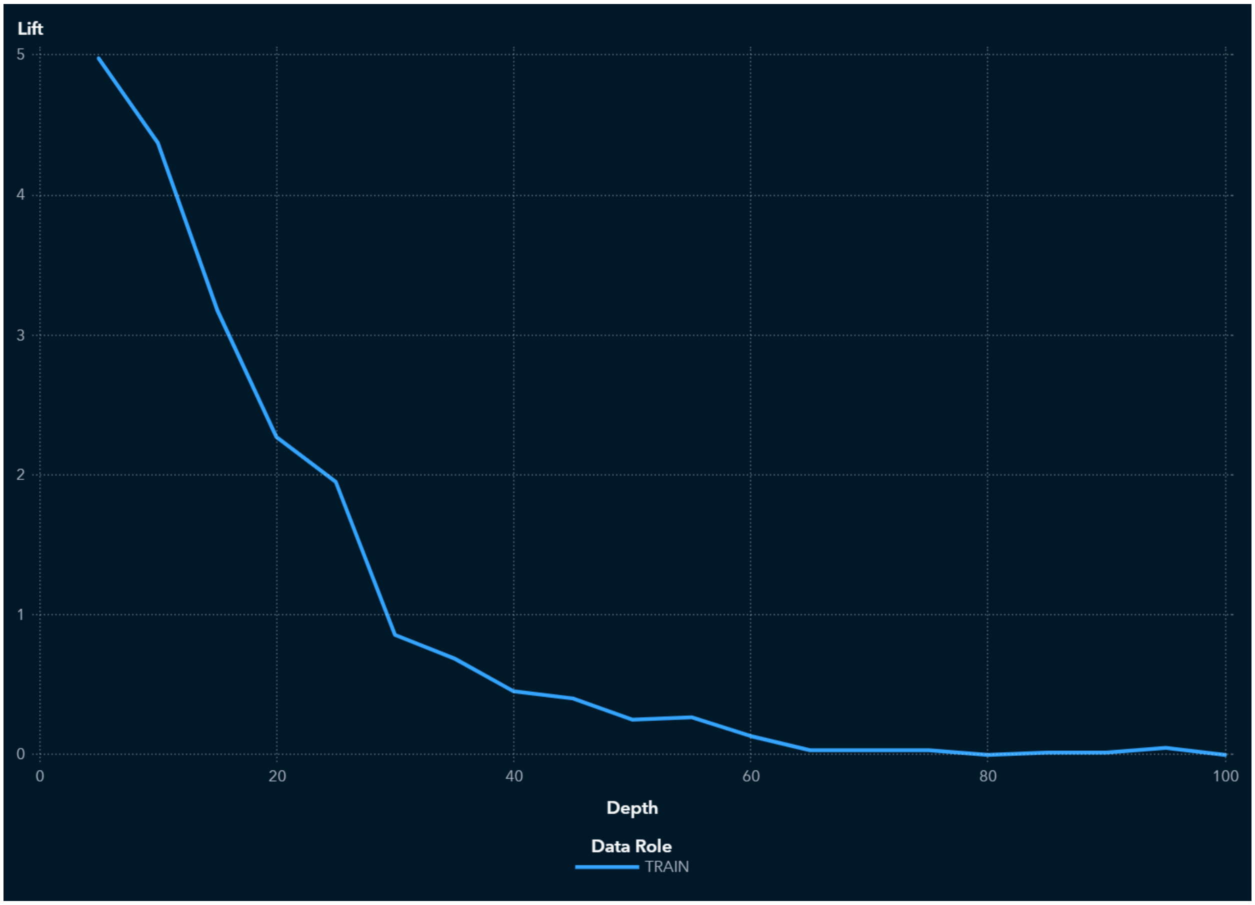
Cumulative Lift



The TRAIN partition has a Cumulative Lift of 4.68 in the 10% quantile (depth of 10) meaning there are 4.68 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.

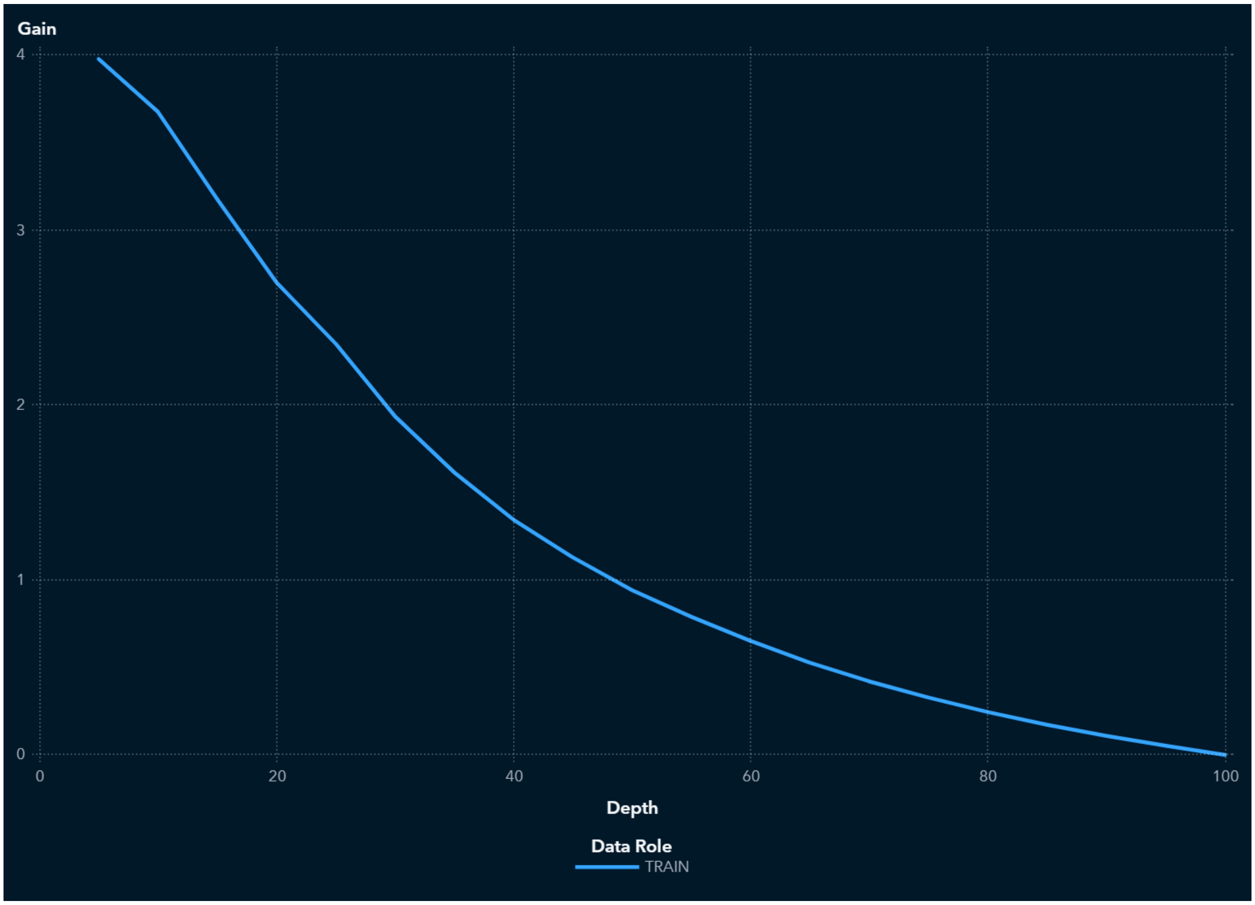
Lift



The TRAIN partition has a Lift of 4.98 in the 5% quantile (depth of 5) meaning there are 4.98 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.

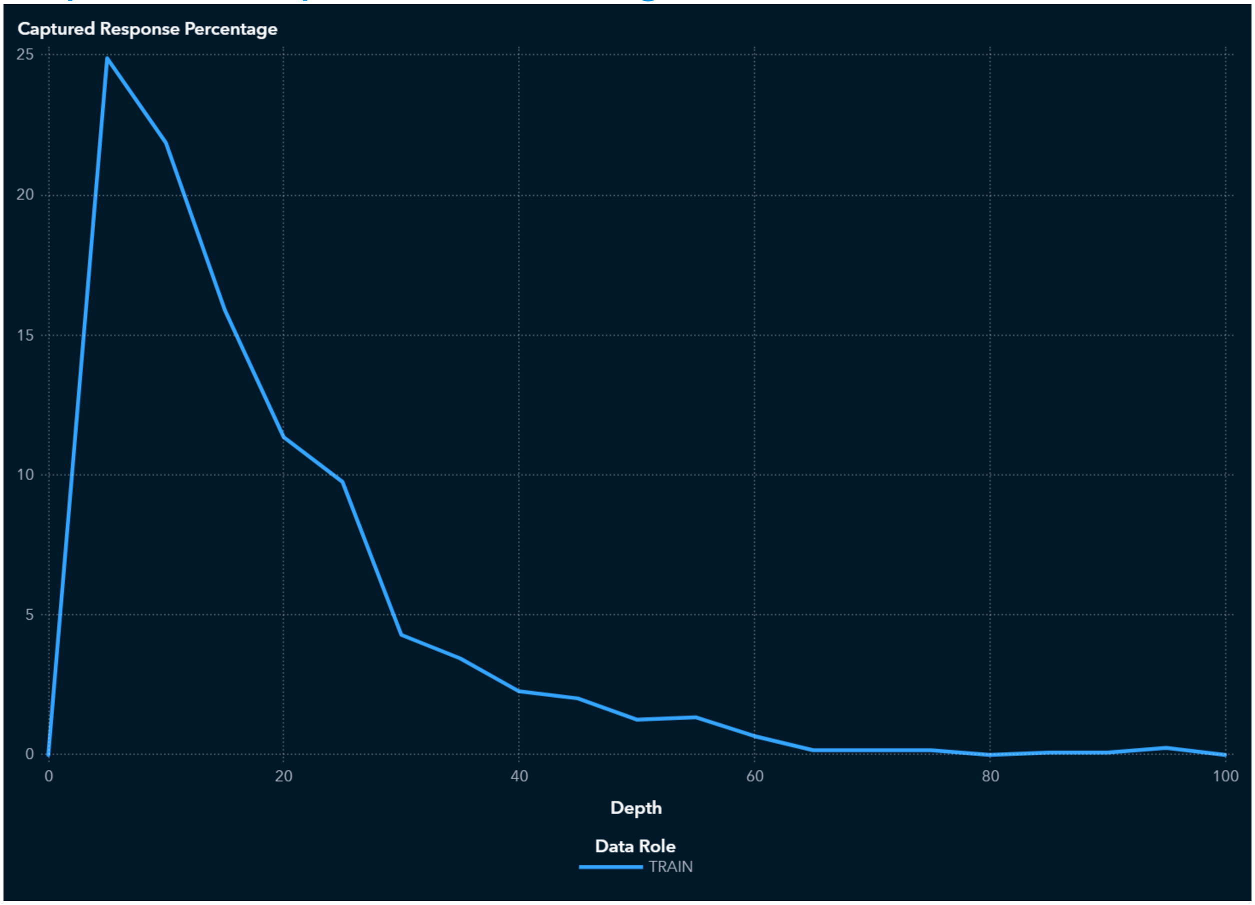
Gain



The TRAIN partition has a Gain of 3.7 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.

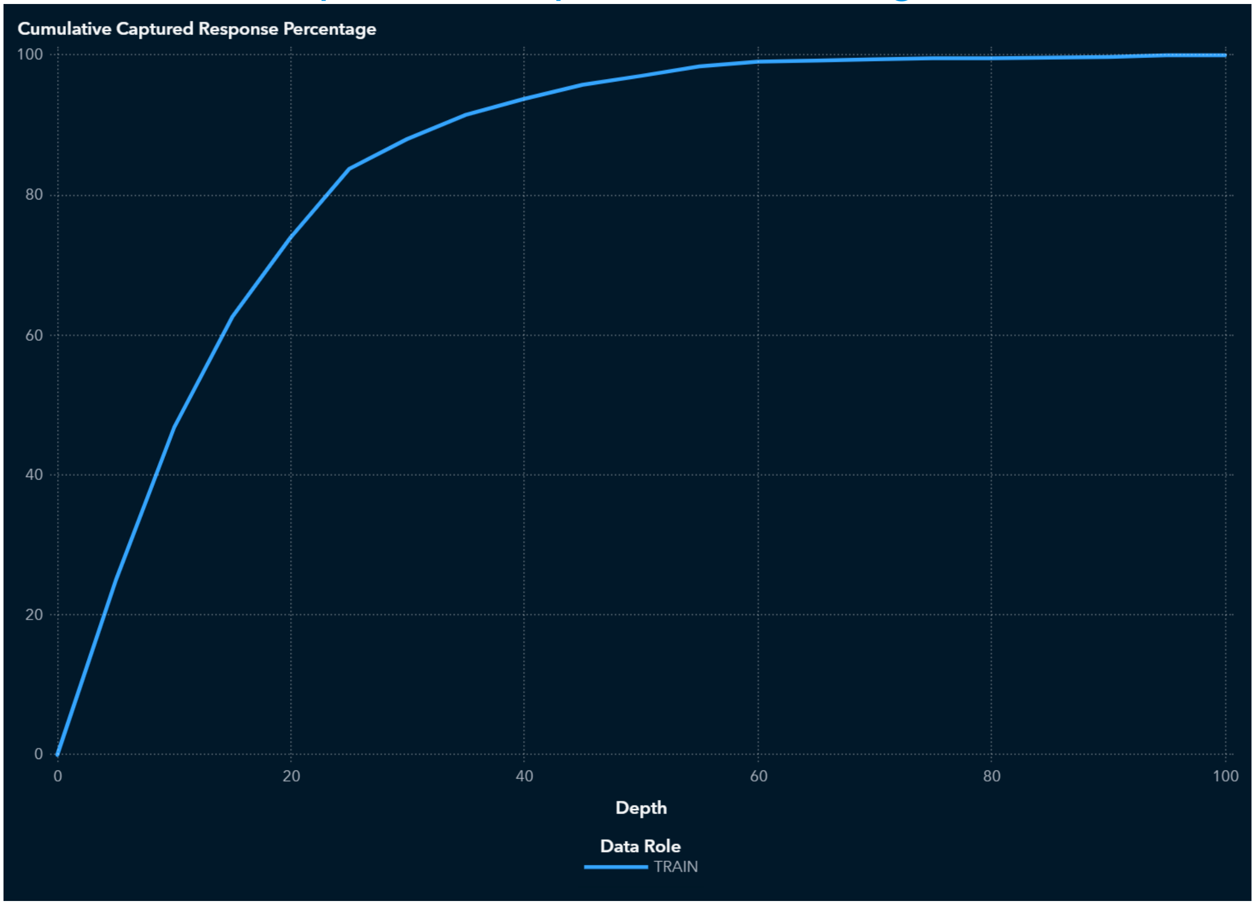
Captured Response Percentage



At the 5% quantile (depth of 5), the TRAIN partition has a Captured response percentage of 24.9 (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.

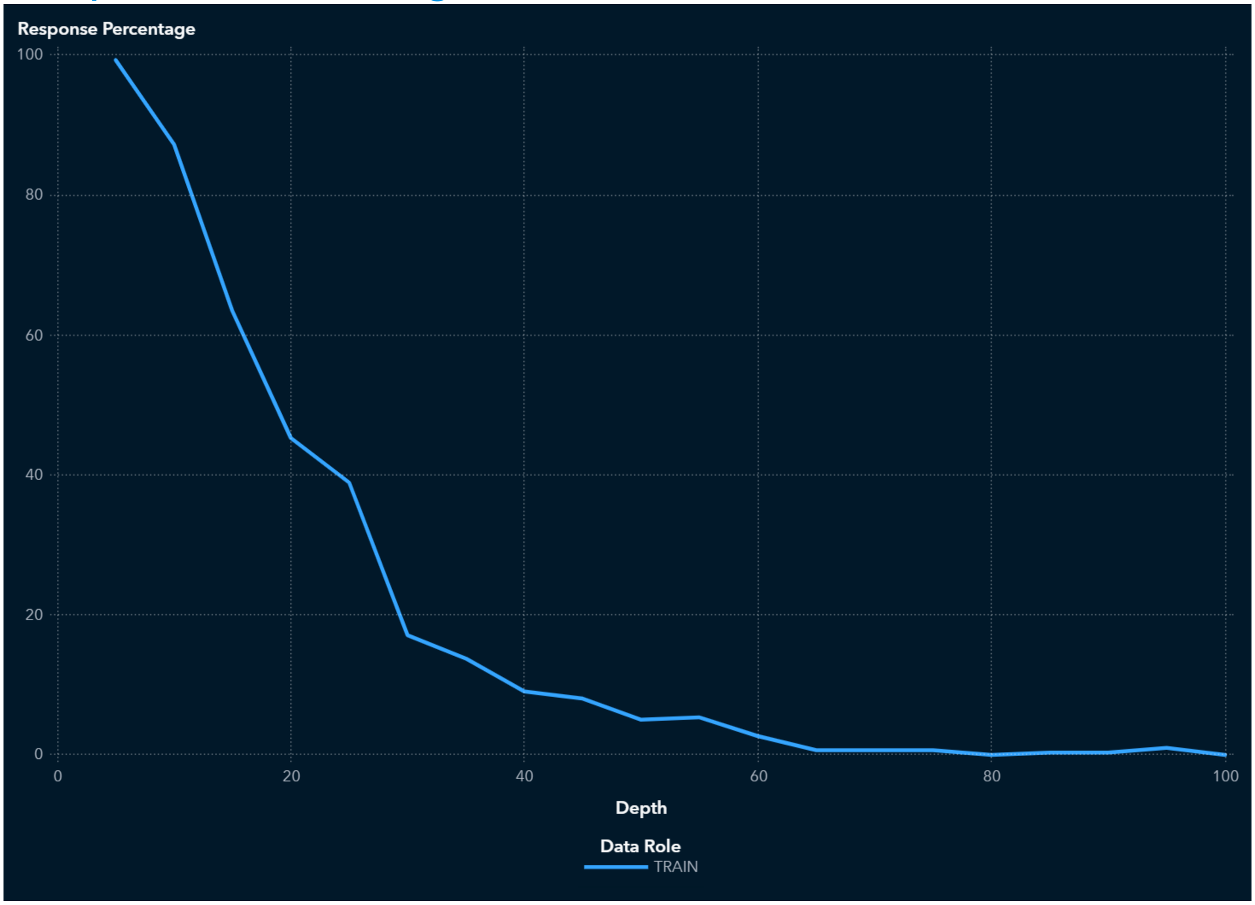
Cumulative Captured Response Percentage



In the top 10% of the data (depth 10), the TRAIN partition has a Cumulative captured response percentage of 46.8 (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%.

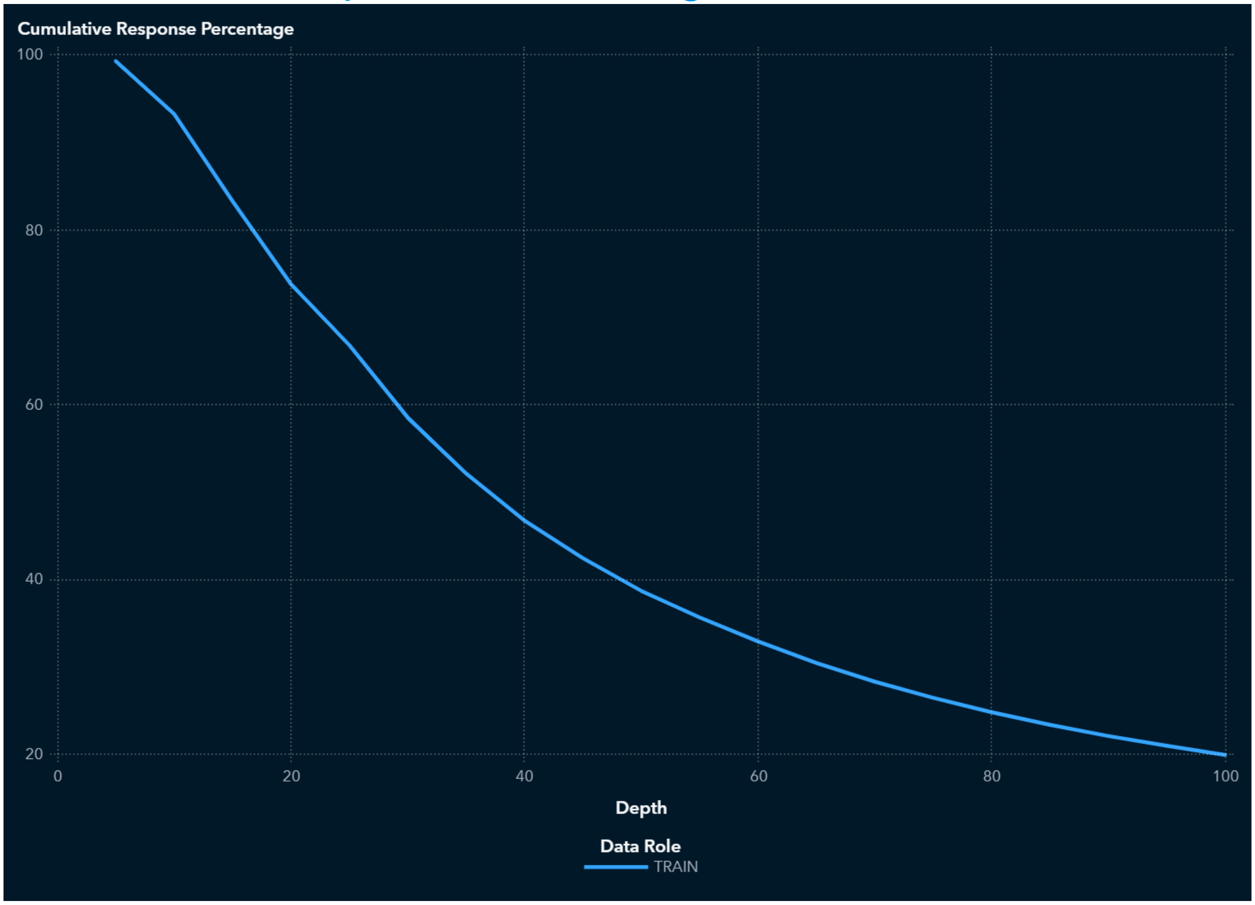
Response Percentage



At the 5% quantile (depth of 5), the TRAIN partition has a Response percentage of 99.3. 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.

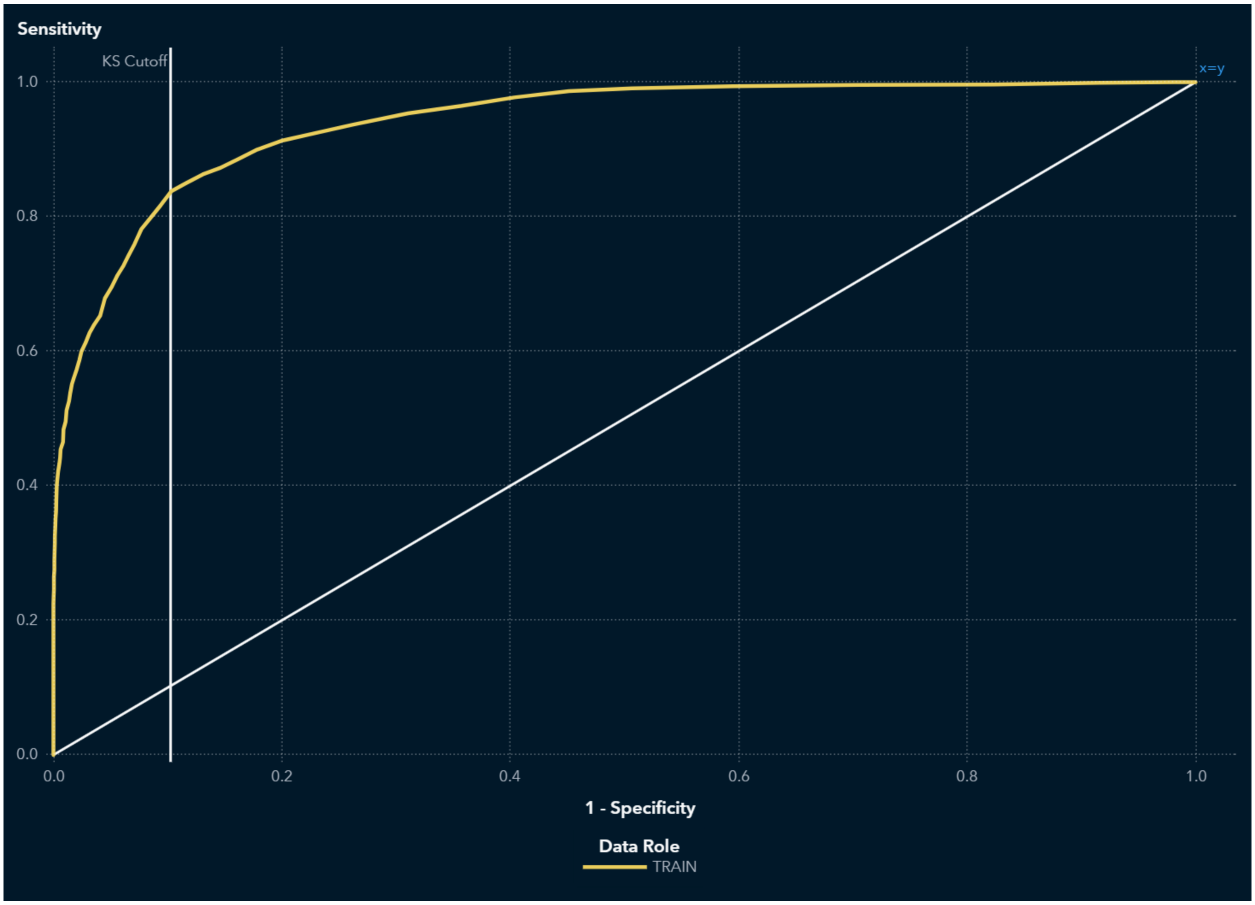
Cumulative Response Percentage



In the top 10% of the data (depth 10), the TRAIN partition has a Cumulative response percentage of 93.3. 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



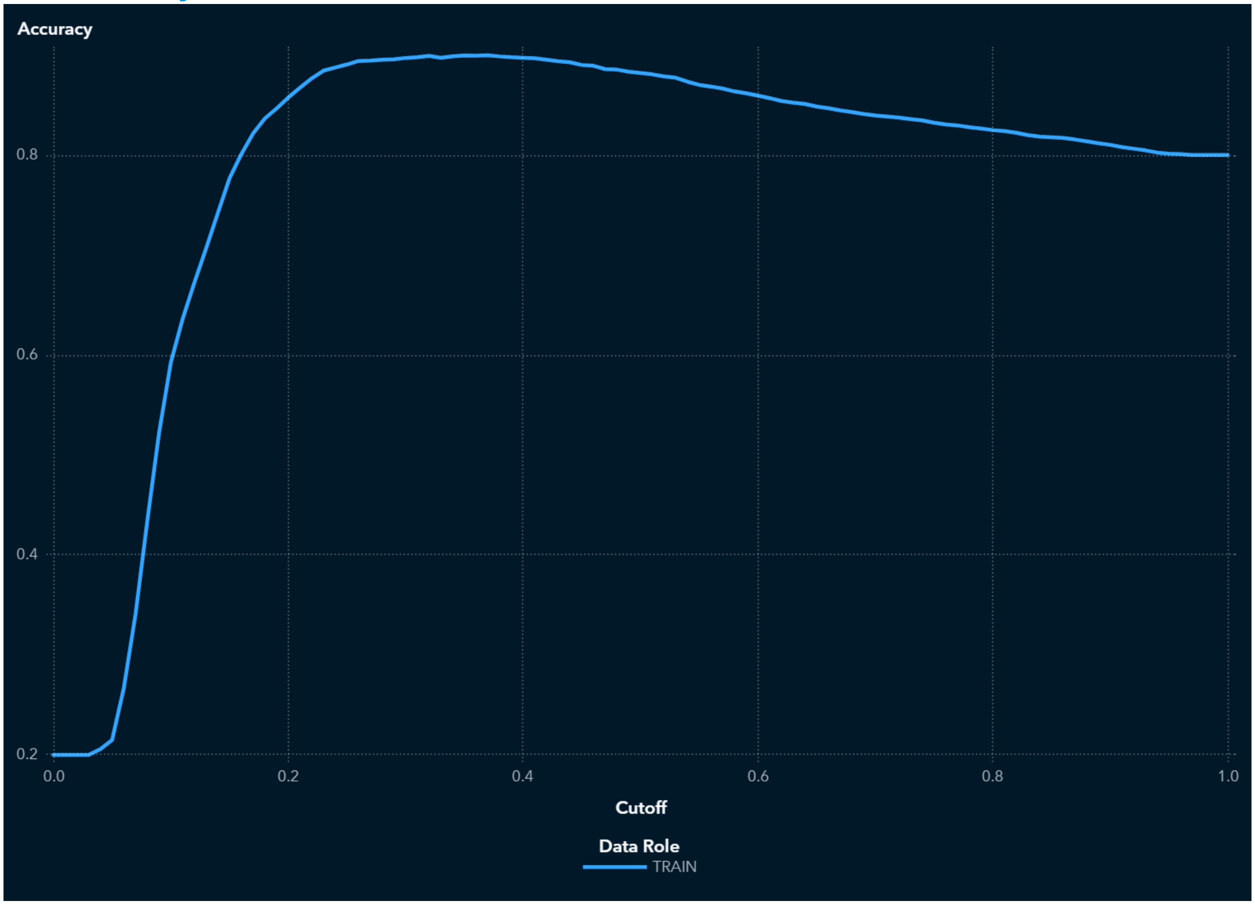
The ROC curve is a plot of sensitivity (the true positive rate) against 1-specificity (the false positive rate), which are both measures of classification based on the confusion matrix. These measures are calculated at various cutoff values. To help identify the best cutoff to use when scoring your data, 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 0.23, where the 1-specificity value is 0.103 and the sensitivity value is 0.838.

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.

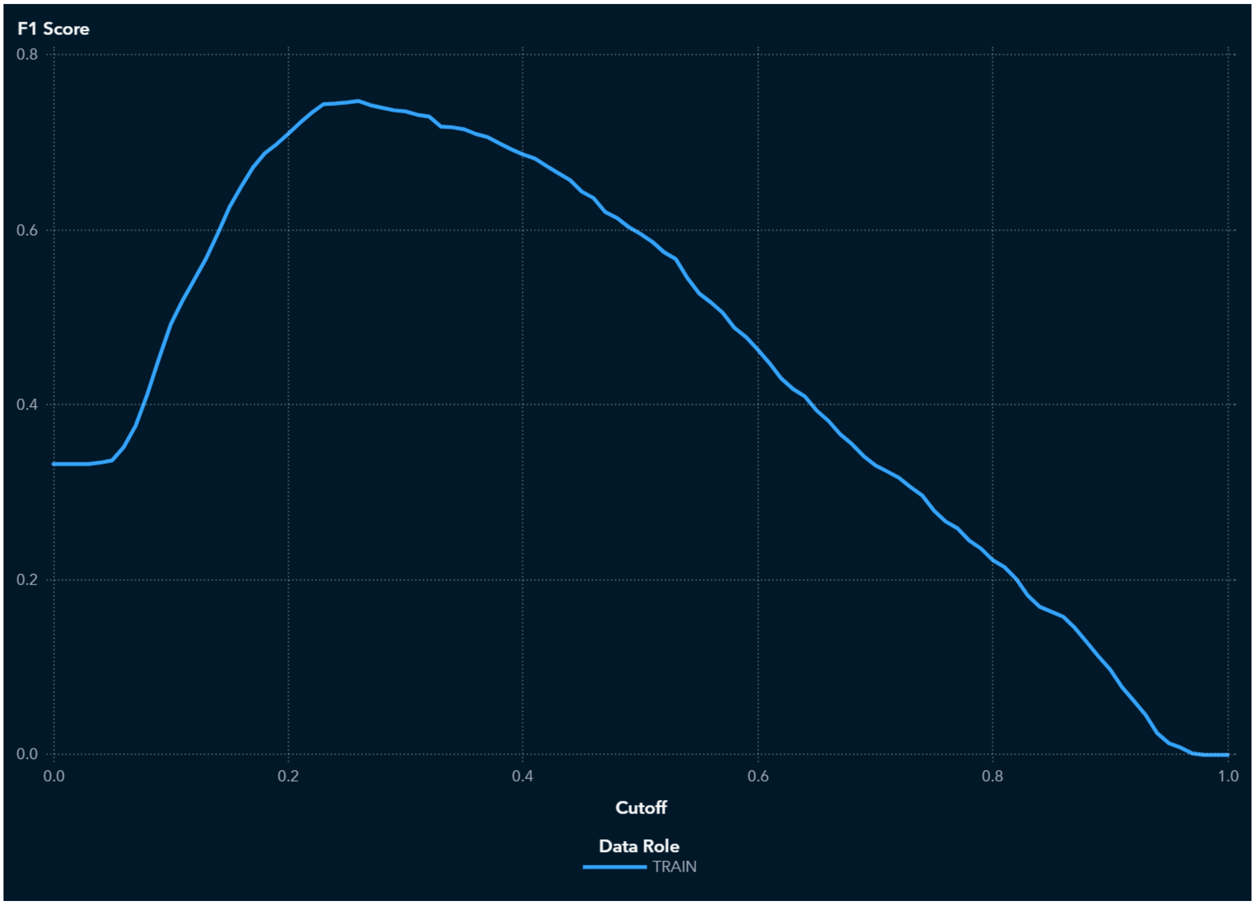
Accuracy



For this model, the accuracy in the TRAIN partition at the cutoff of 0.5 is 0.883.

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 0.596.

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	0.0853

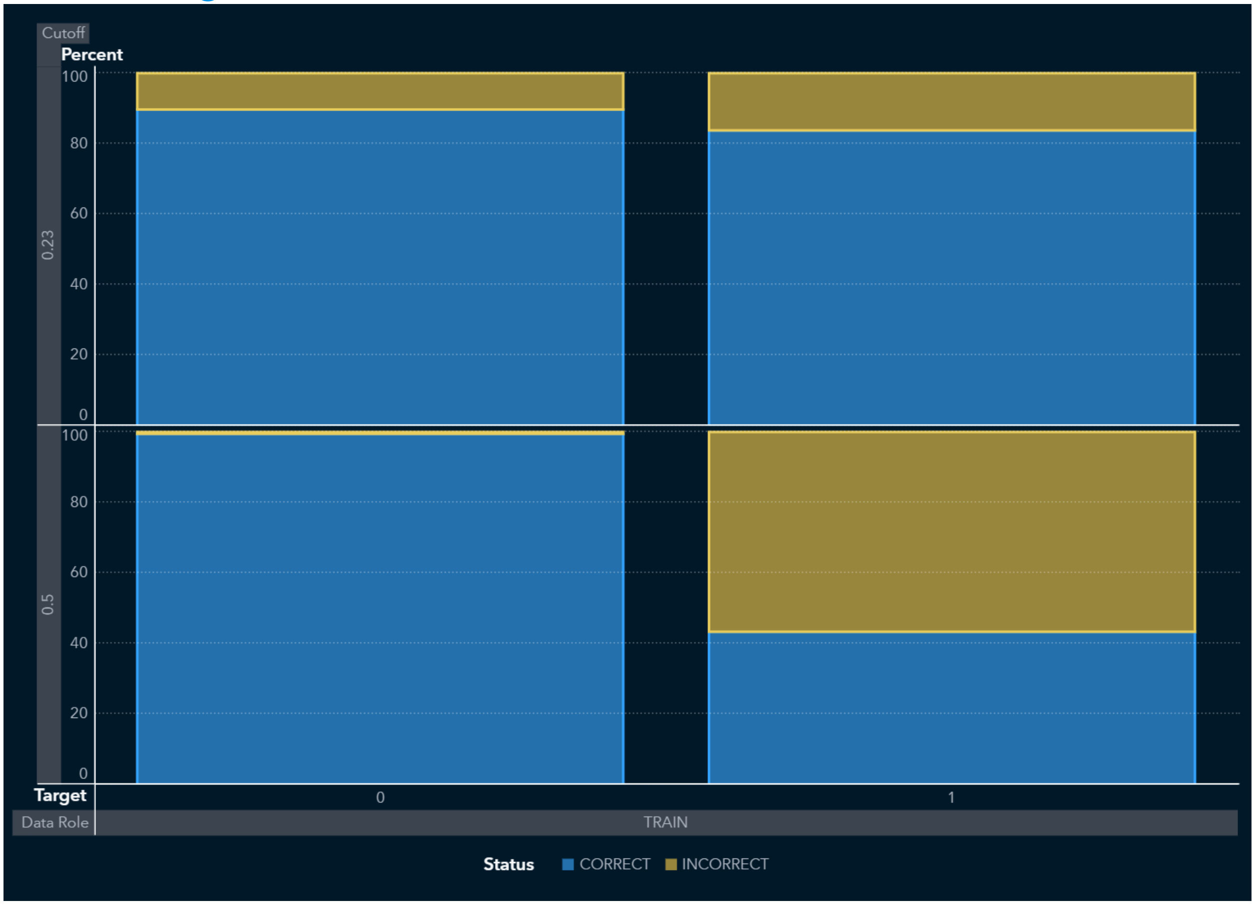
Divisor for ASE	Root Average Squared Error	Misclassification Rate	Multi-Class Log Loss
5,960	0.2921	0.1173	0.2939

KS (Youden)	Area Under ROC	Gini Coefficient	Gamma
0.7346	0.9407	0.8813	0.8876

Tau	KS Cutoff	KS at User-Specified Cutoff	Misclassification Rate at KS Cutoff (Event)
0.2815	0.2300	0.4279	0.1149

Misclassification Rate (Event)
0.1173

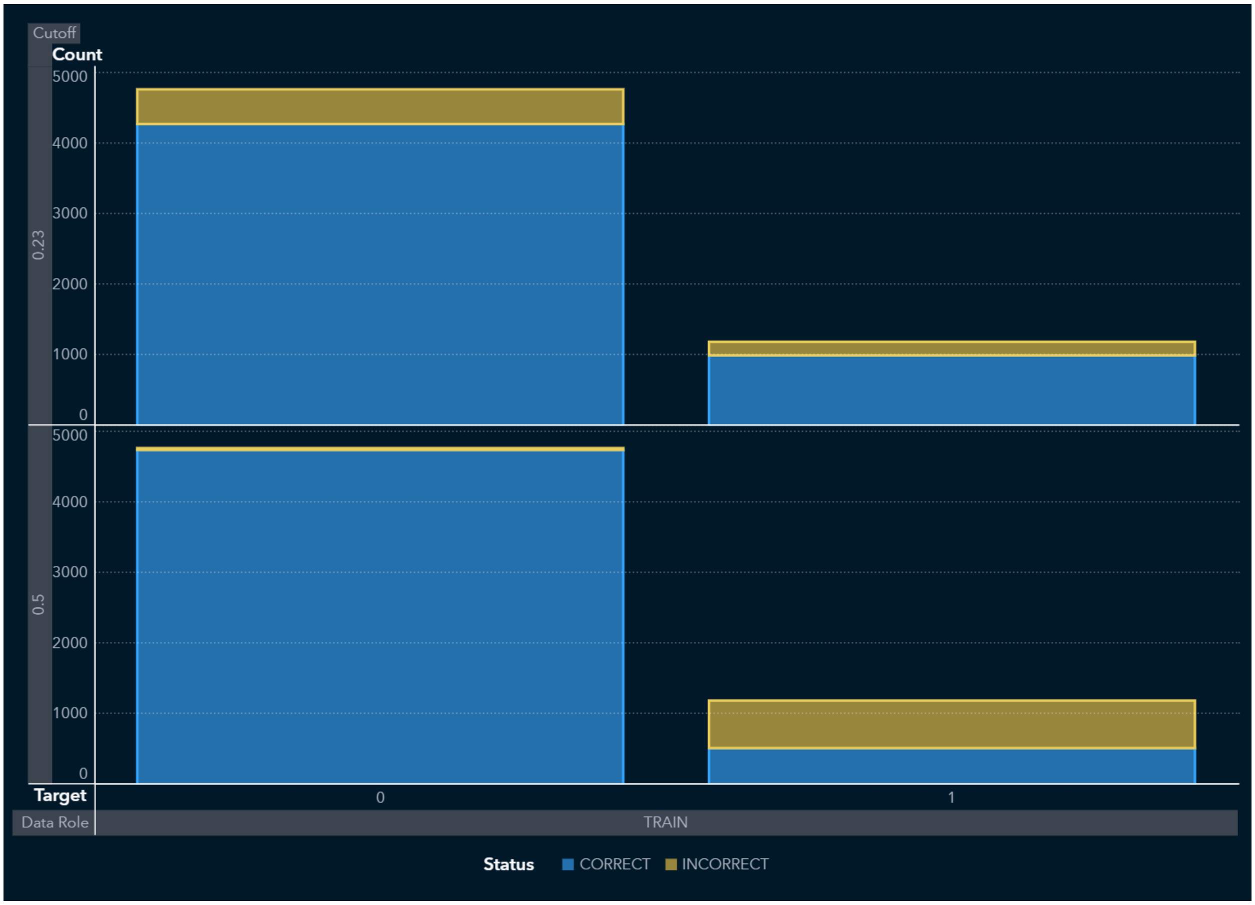
Percentage Plot



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 0.23 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 0.23 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
0.2300	KS	BAD	CORRECT
0.2300	KS	BAD	INCORRECT
0.2300	KS	BAD	CORRECT
0.2300	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	996	
1	False Negative	193	
0	True Negative	4,279	
0	False Positive	492	
1	True Positive	515	
1	False Negative	674	
0	True Negative	4,746	
0	False Positive	25	

Test Frequency	Training Percentage	Validation Percentage	Test Percentage
	83.7679		
	16.2321		
	89.6877		
	10.3123		
	43.3137		
	56.6863		
	99.4760		

Test Frequency	Training Percentage	Validation Percentage	Test Percentage
	0.5240		

Score Inputs

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

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

Score Outputs

Name	Variable Label	Role	Type
EM_CLASSIFICATION	Predicted for BAD	CLASSIFICATION	C
EM_EVENTPROBABILITY	Probability for BAD=1	PREDICT	N
EM_PROBABILITY	Probability of Classification	PREDICT	N
I_BAD	Into: BAD	CLASSIFICATION	C
P_BAD0	Predicted: BAD=0	PREDICT	N
P_BAD1	Predicted: BAD=1	PREDICT	N
WARN	Warnings	ASSESS	C
IMP_DELINQ	Imputed DELINQ	INPUT	N
IMP_DEROG	Imputed DEROG	INPUT	N
IMP_JOB	Imputed JOB	INPUT	C
IMP_NINQ	Imputed NINQ	INPUT	N
IMP_REASON	Imputed REASON	INPUT	C
IMP_VALUE	Imputed VALUE	REJECTED	N

Variable Type	Variable Format	Variable Length	Creator
char		12	forest
double		8	forest
double		8	forest
char		12	forest
double		8	forest
double		8	forest
char		4	forest
double		8	impute
double		8	impute
char		7	impute
double		8	impute

Variable Type	Variable Format	Variable Length	Creator
char		7	impute
double		8	impute

Function	Creator GUID
CLASSIFICATION	6de51889-3c65-40c9-b5da-a333df38325f
PREDICT	6de51889-3c65-40c9-b5da-a333df38325f
PREDICT	6de51889-3c65-40c9-b5da-a333df38325f
CLASSIFICATION	6de51889-3c65-40c9-b5da-a333df38325f
PREDICT	6de51889-3c65-40c9-b5da-a333df38325f
PREDICT	6de51889-3c65-40c9-b5da-a333df38325f
ASSESS	6de51889-3c65-40c9-b5da-a333df38325f
TRANSFORM	b727c11b-1371-4c1d-8746-51b94c237069
TRANSFORM	b727c11b-1371-4c1d-8746-51b94c237069
TRANSFORM	b727c11b-1371-4c1d-8746-51b94c237069

Function	Creator GUID
TRANSFORM	b727c11b-1371-4c1d-8746-51b94c237069
TRANSFORM	b727c11b-1371-4c1d-8746-51b94c237069
TRANSFORM	b727c11b-1371-4c1d-8746-51b94c237069

Properties

Property Name	Property Value
accumulatedCode	true
binaryProbCutoff	0.5000
codeLocation	mlearning
dataMiningVersion	V2024.03
exactPctlLift	true
explainFidelity	false
explainInfo	false
fullDatasetReconstitution	false
geometricConstant	1
icePlots	false
inputs	immediatePredecessors(@currentNode)
maxNumShapVars	20
nBins	50
pdNumImportantInputs	5
pdObsSamples	1,000
pdPlots	false
performKernelShap	false
performLime	false
performVI	false
posteriorFunction	AVERAGE
predictedFunction	AVERAGE
reportingOnly	false
seedId	12,345
specifyRows	RANDOM
templateRevision	1

Property Name	Property Value
truncateLI	5
truncateUI	95
userProbCutoff	false
votingFunction	AVERAGE

Models

Name
Forest
Stepwise Logistic Regression
Neural Network
Gradient Boosting
Decision Tree
Forward Logistic Regression