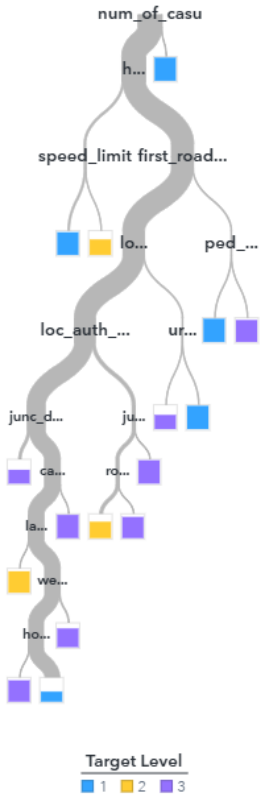
Accident_Fatality_Predict...

"Decision Tree" Results

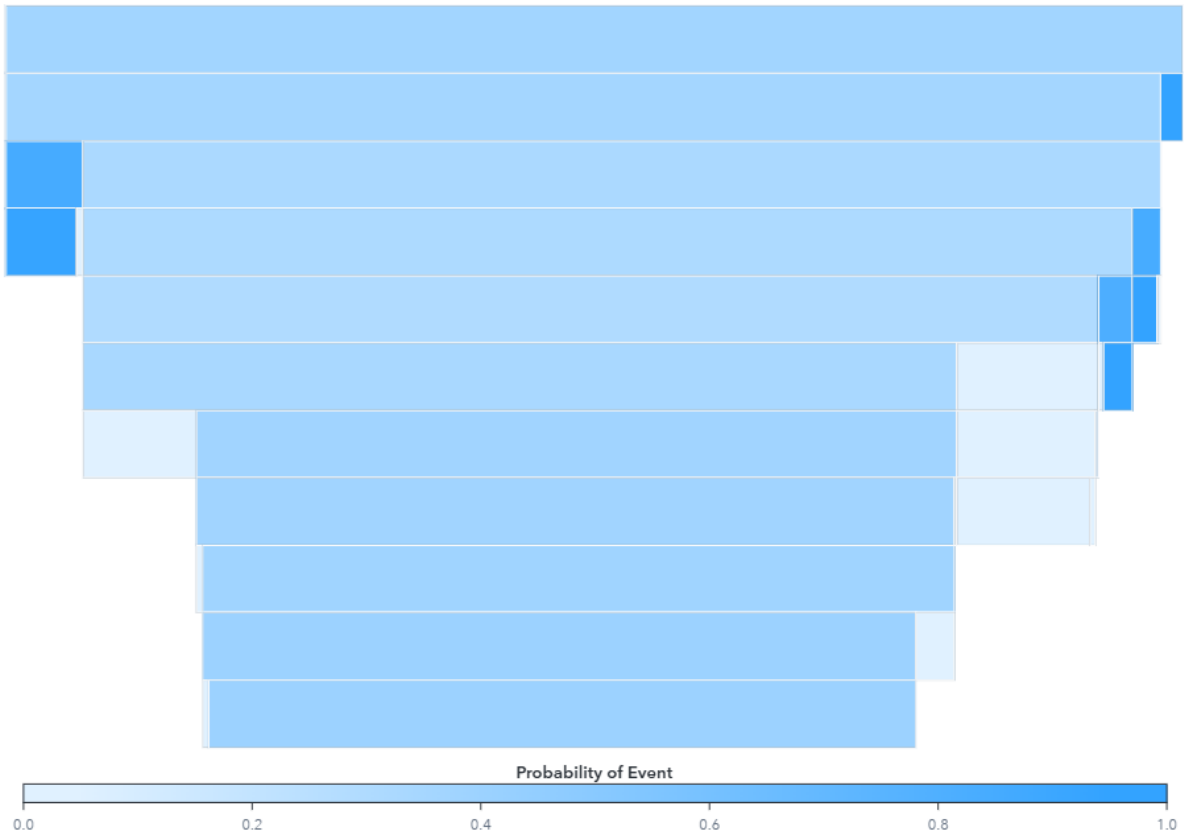
by: ta01468@surrey.ac.uk

Contents

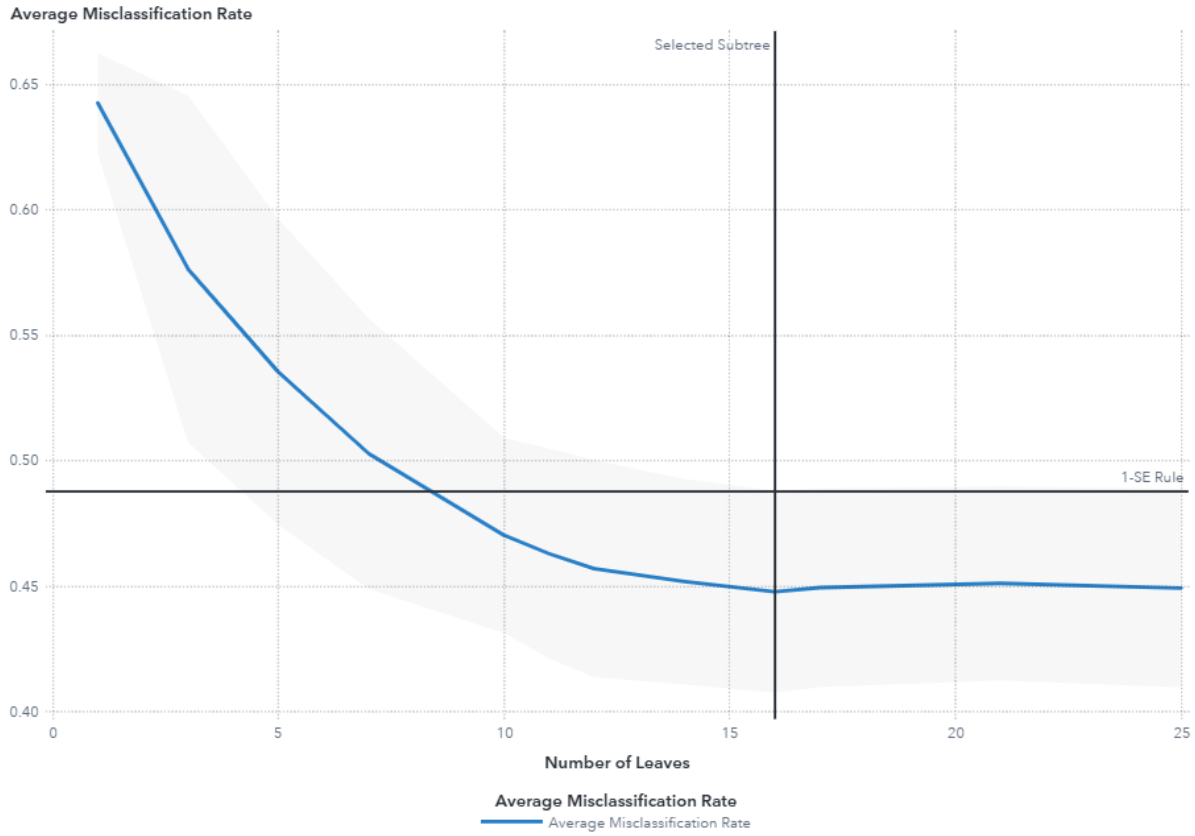
Tree Diagram



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 0.488, 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 16 leaves with an average misclassification rate across folds of 0.448.

Variable Importance

Variable Name	Training Relative Importance	Count	Training Importance
hour_of_day	1	2	85.4301
junc_detail	0.5337	2	45.5964
loc_auth_ons_distr	0.5146	1	43.9653
longitude	0.4049	1	34.5935
first_road_num	0.3553	1	30.3521
num_of_casu	0.3323	1	28.3856
weath_con	0.3105	1	26.5254
speed_limit	0.2329	1	19.9007
urb_or_rur_area	0.1641	1	14.0177
ped_cross_hum_con	0.1267	1	10.8281
latitude	0.0854	1	7.2924
road_type	0.0825	1	7.0448
carri_haz	0.0408	1	3.4867

Score Inputs

Name	Role	Variable Level	Type
carri_haz	INPUT	NOMINAL	N
first_road_num	INPUT	INTERVAL	N
hour_of_day	INPUT	INTERVAL	N
junc_detail	INPUT	NOMINAL	N
latitude	INPUT	INTERVAL	N
loc_auth_ons_distr	INPUT	NOMINAL	C
longitude	INPUT	INTERVAL	N
num_of_casu	INPUT	NOMINAL	N
ped_cross_hum_con	INPUT	NOMINAL	N
road_type	INPUT	NOMINAL	N
speed_limit	INPUT	NOMINAL	N
urb_or_rur_area	INPUT	BINARY	N
weath_con	INPUT	NOMINAL	N

Variable Type	Variable Label	Variable Format	Variable Length
double			8
double			8
double			8
double			8
double			8
varchar			9
double			8
double			8
double			8
double			8
double			8
double			8
double			8

Score Outputs

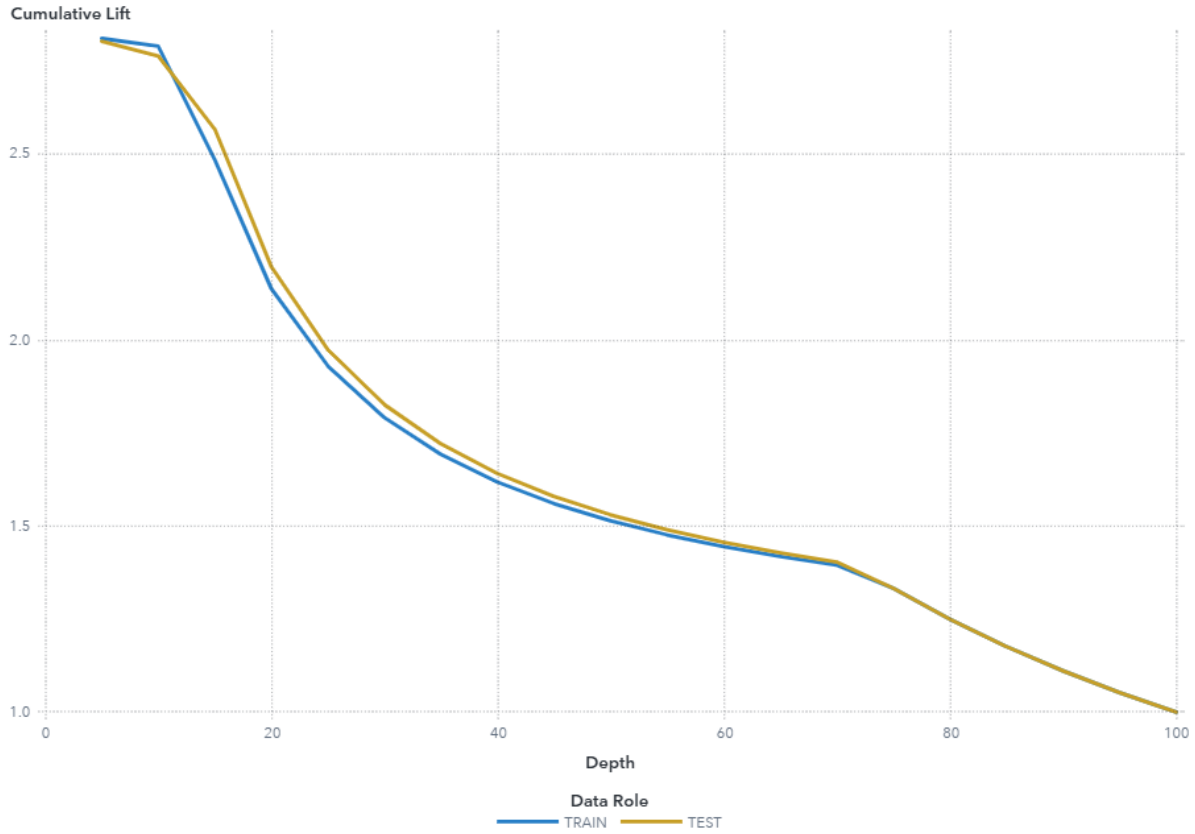
Name	Role	Type	Variable Type
EM_CLASSIFICATION	CLASSIFICATION	C	char
EM_EVENTPROBABILITY	PREDICT	N	double
EM_PROBABILITY	PREDICT	N	double
I_acci_severity	CLASSIFICATION	C	char
P_acci_severity1	PREDICT	N	double
P_acci_severity2	PREDICT	N	double
P_acci_severity3	PREDICT	N	double
WARN	ASSESS	C	char

Variable Label	Variable Format	Variable Length	Creator
Predicted for acci_severity		12	tree
Probability for acci_severity=1		8	tree
Probability of Classification		8	tree
Into: acci_severity		32	tree
Predicted: acci_severity=1		8	tree
Predicted: acci_severity=2		8	tree
Predicted: acci_severity=3		8	tree
Warnings		4	tree

Function	Creator GUID
CLASSIFICATION	8c30316c-3ed3-4883-8c3a-f5ffe1867aa3
PREDICT	8c30316c-3ed3-4883-8c3a-f5ffe1867aa3

Function	Creator GUID
PREDICT	8c30316c-3ed3-4883-8c3a-f5ffe1867aa3
CLASSIFICATION	8c30316c-3ed3-4883-8c3a-f5ffe1867aa3
PREDICT	8c30316c-3ed3-4883-8c3a-f5ffe1867aa3
PREDICT	8c30316c-3ed3-4883-8c3a-f5ffe1867aa3
PREDICT	8c30316c-3ed3-4883-8c3a-f5ffe1867aa3
ASSESS	8c30316c-3ed3-4883-8c3a-f5ffe1867aa3

Cumulative Lift



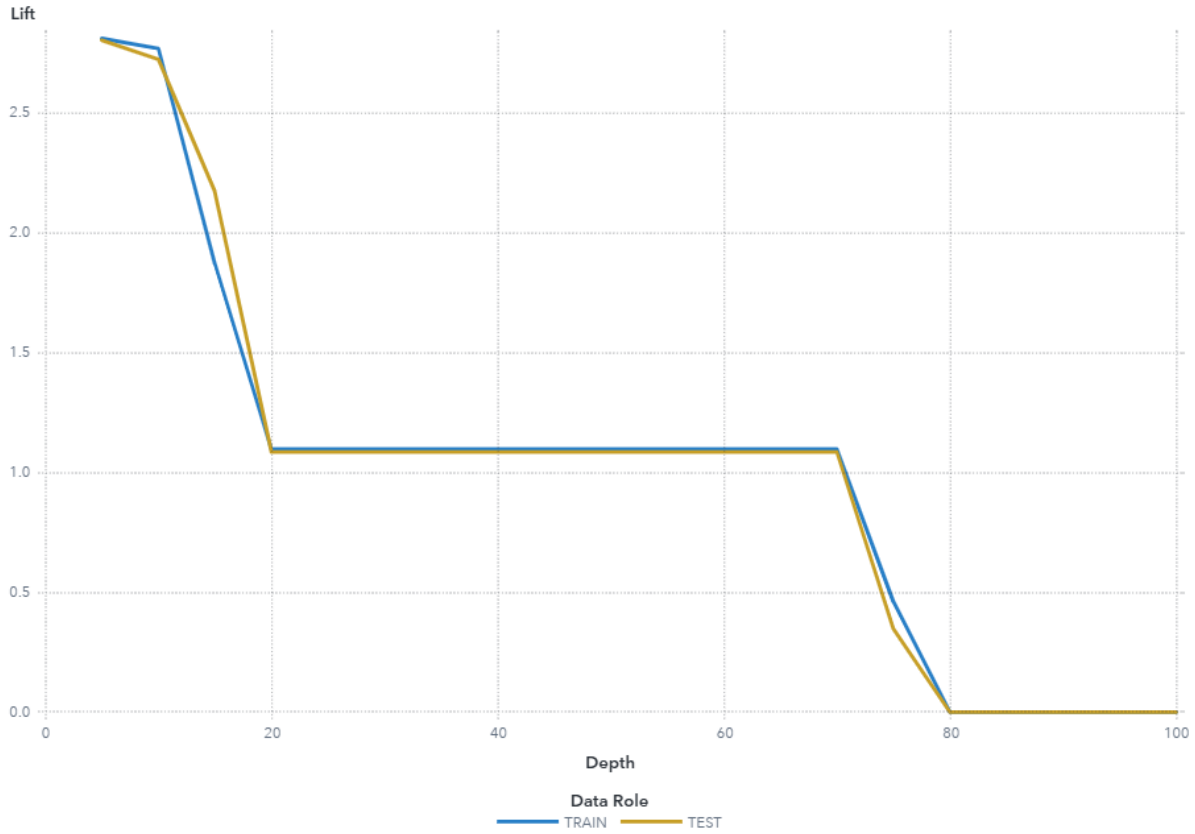
The TRAIN partition has a Cumulative Lift of 2.79 in the 10% quantile (depth of 10) meaning there are 2.79 times more events in the first two quantiles than expected by random (10% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition.

The TEST partition has a Cumulative Lift of 2.76 in the 10% quantile (depth of 10) meaning there are 2.76 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_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. 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

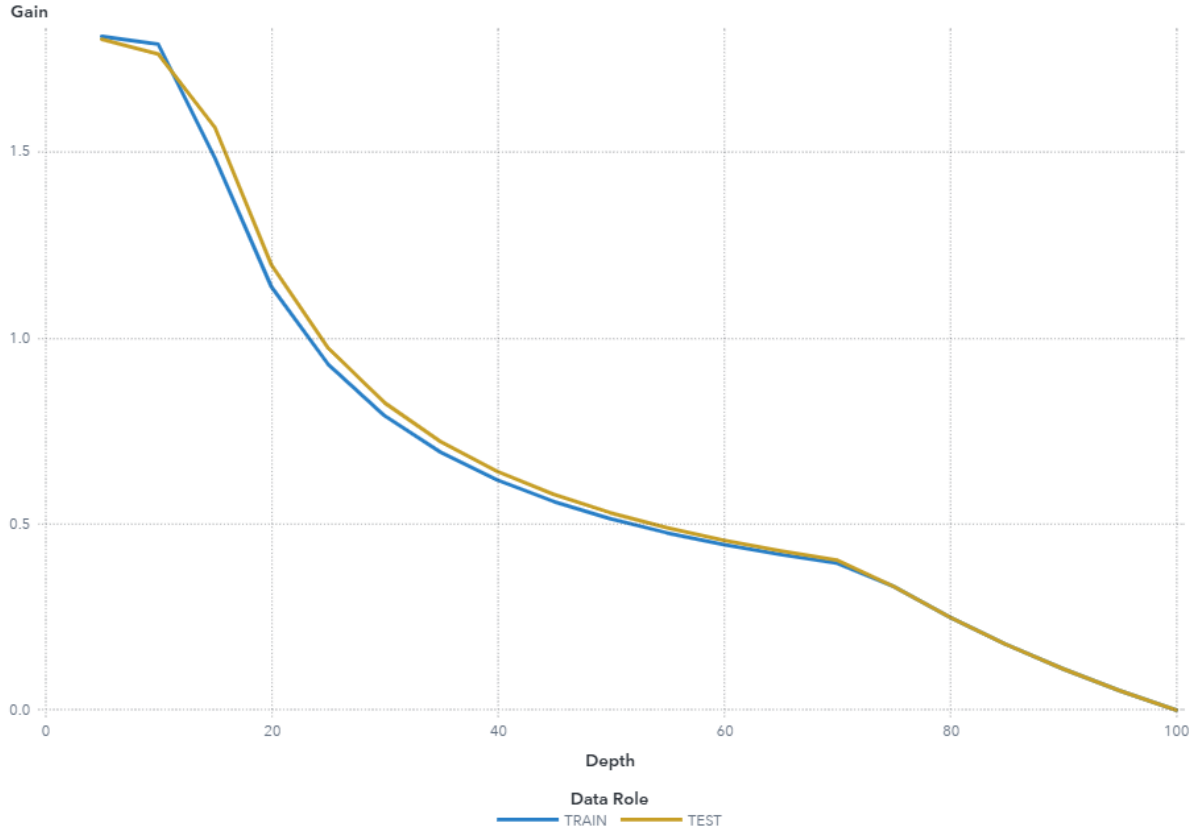


The TRAIN partition has a Lift of 2.81 in the 5% quantile (depth of 5) meaning there are 2.81 times more events in that quantile than expected by random (5% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition.

The TEST partition has a Lift of 2.8 in the 5% quantile (depth of 5) meaning there are 2.8 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_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. 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

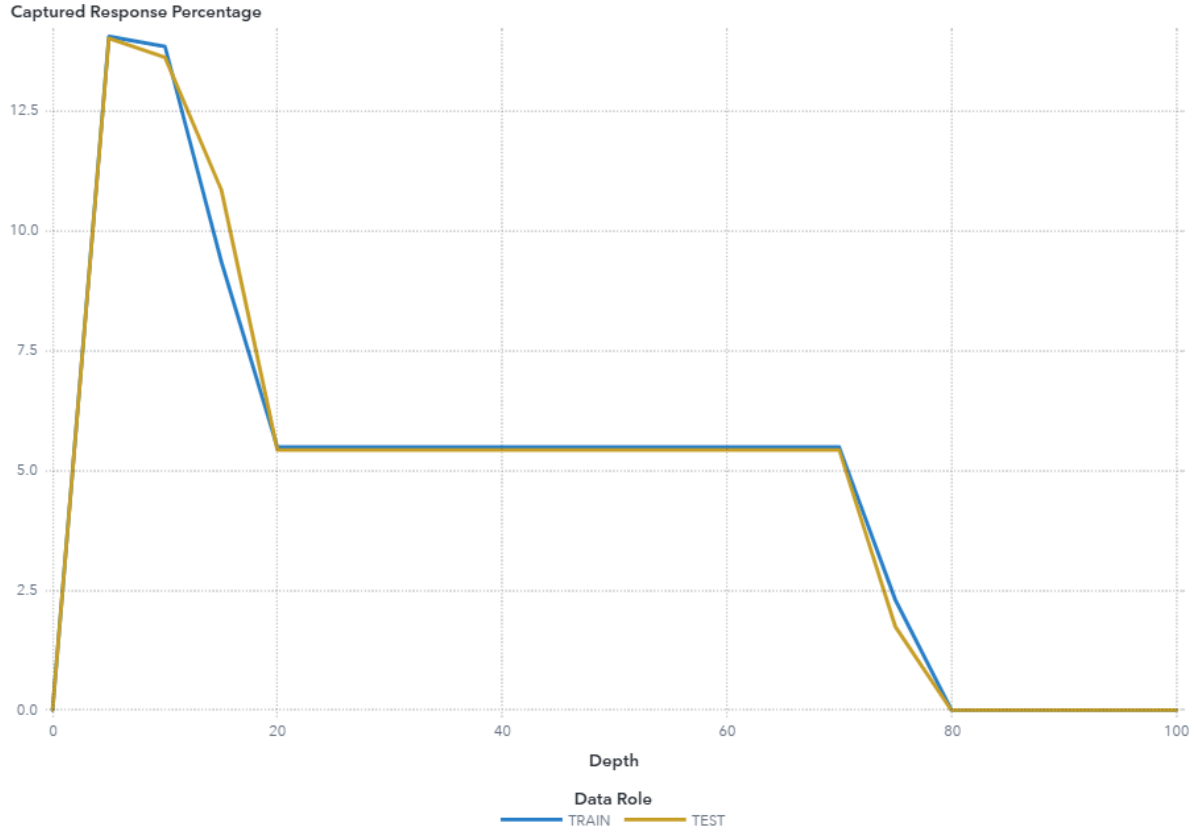


The TRAIN partition has a Gain of 1.8 at the 10% quantile (depth of 10). Because this value is greater than 0, it is better to use your model to identify responders than no model, based on the selected partition. The best possible value of Gain for this partition at depth 10 is 1.81.

The TEST partition has a Gain of 1.8 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 1.8.

Gain is calculated by sorting each partition in descending order by the predicted probability of the target event P_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. 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

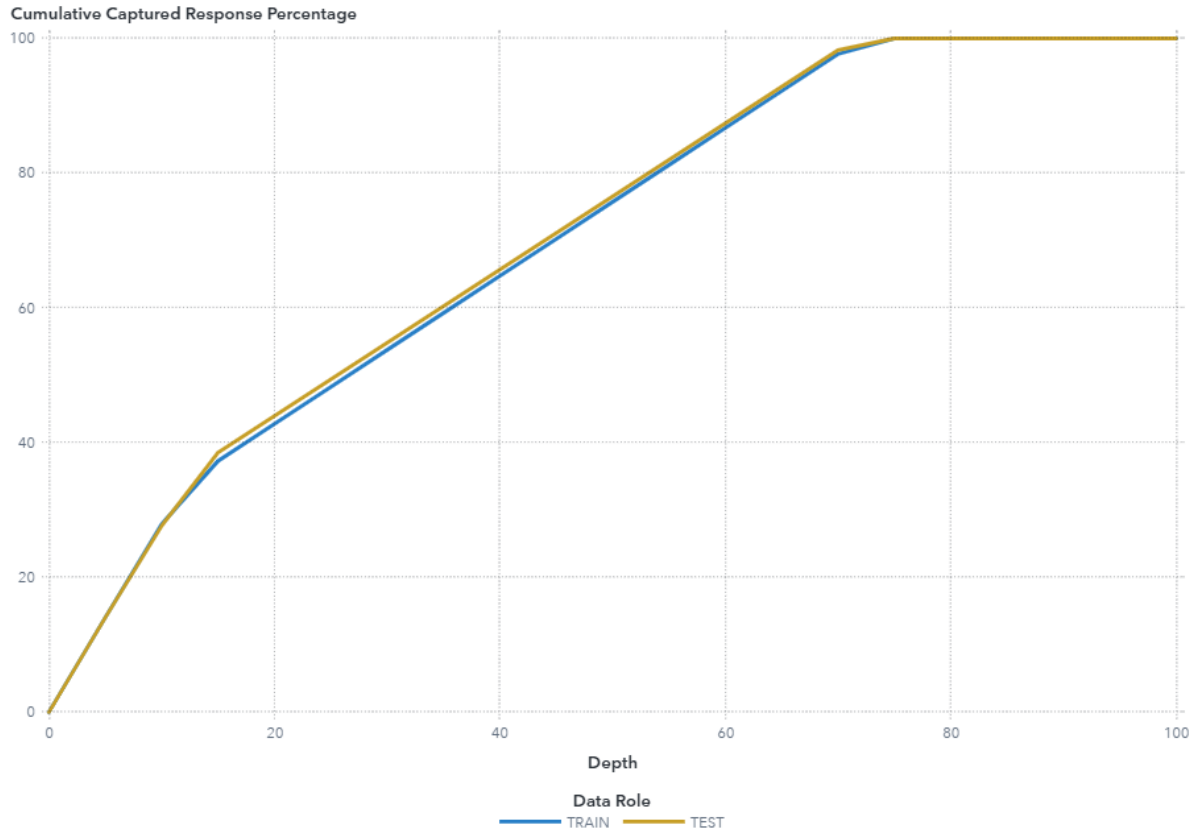


At the 5% quantile (depth of 5), the TRAIN partition has a Captured response percentage of 14.1 (compared to the expected value of 5 for no model). The best possible value of Captured response percentage for this partition at depth 5 is 14.06.

At the 5% quantile (depth of 5), the TEST partition has a Captured response percentage of 14 (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 14.02.

Captured response percentage is calculated by sorting each partition in descending order by the predicted probability of the target event P_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. 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

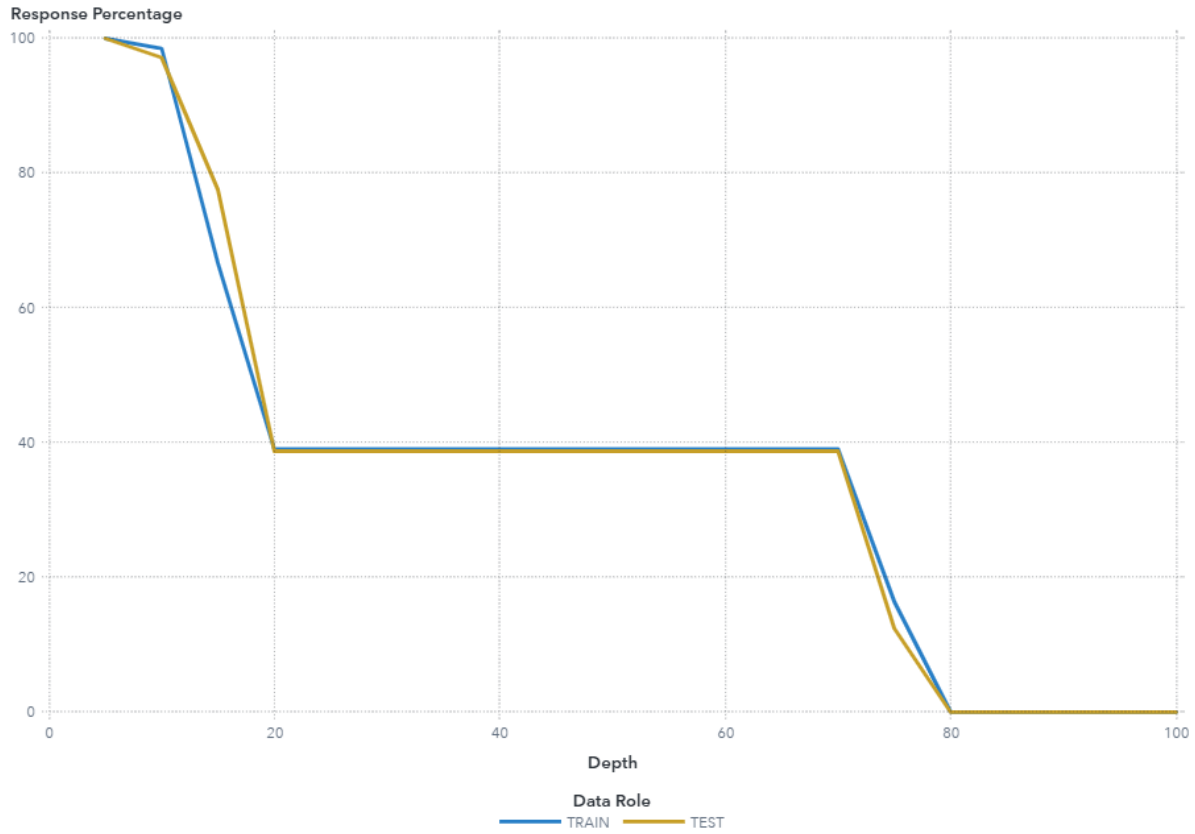


In the top 10% of the data (depth 10), the TRAIN partition has a Cumulative captured response percentage of 27.9 (compared to the expected value of 10 for no model). The best possible value of Cumulative captured response percentage for this partition at depth 10 is 28.12.

In the top 10% of the data (depth 10), the TEST partition has a Cumulative captured response percentage of 27.6 (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 28.04.

Cumulative captured response percentage is calculated by sorting each partition in descending order by the predicted probability of the target event P_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. 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

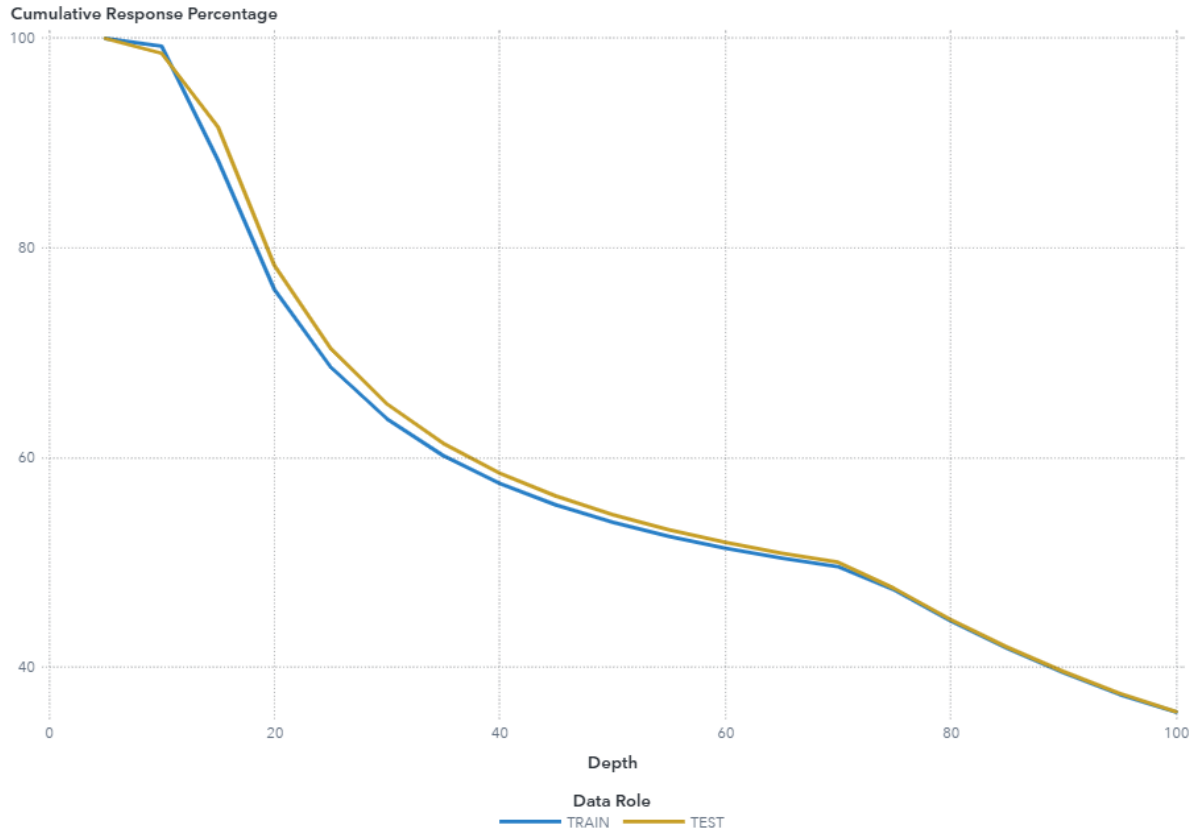


At the 5% quantile (depth of 5), the TRAIN partition has a Response percentage of 100. The best possible value of Response percentage for this partition at depth 5 is 100.

At the 5% quantile (depth of 5), the TEST partition has a Response percentage of 100. 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_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. 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

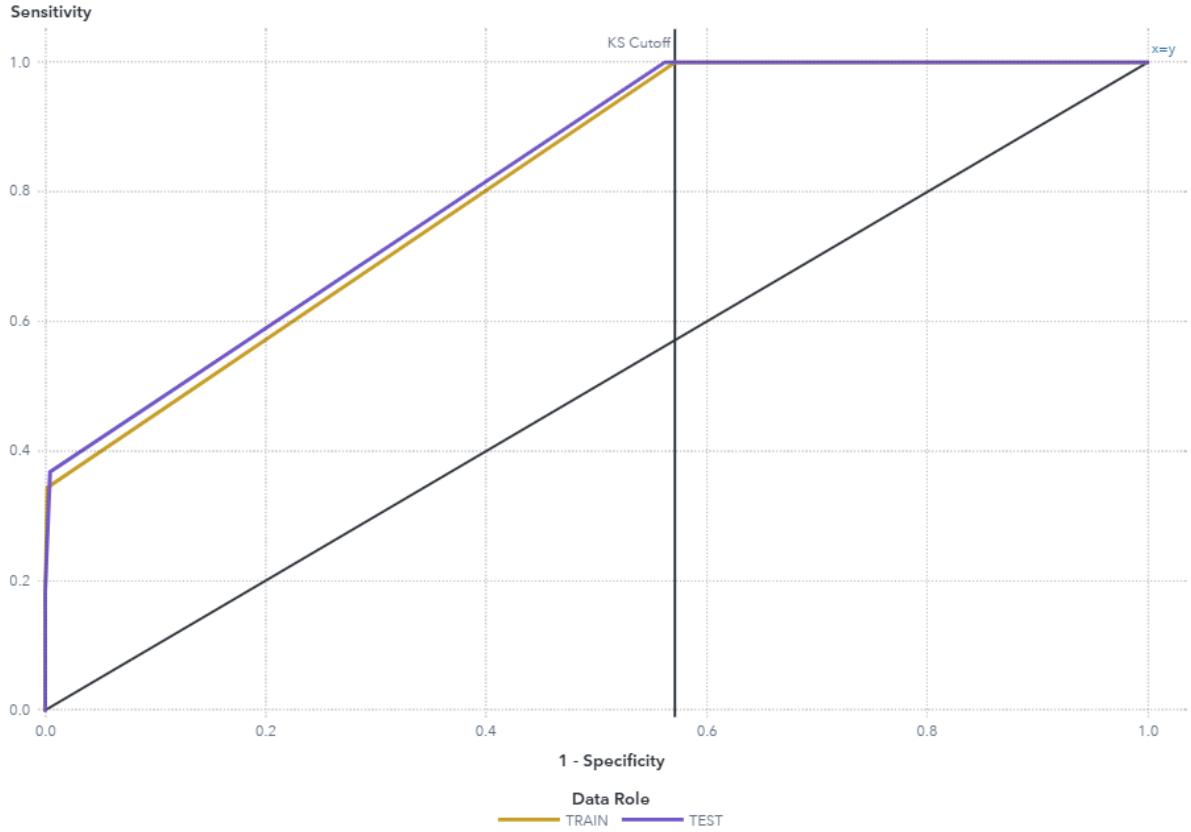


In the top 10% of the data (depth 10), the TRAIN partition has a Cumulative response percentage of 99.3. The best possible value of Cumulative response percentage for this partition at depth 10 is 100.

In the top 10% of the data (depth 10), the TEST partition has a Cumulative response percentage of 98.6. 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_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. 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 0.01, where the 1-specificity value is 0.572 and the sensitivity value is 1.

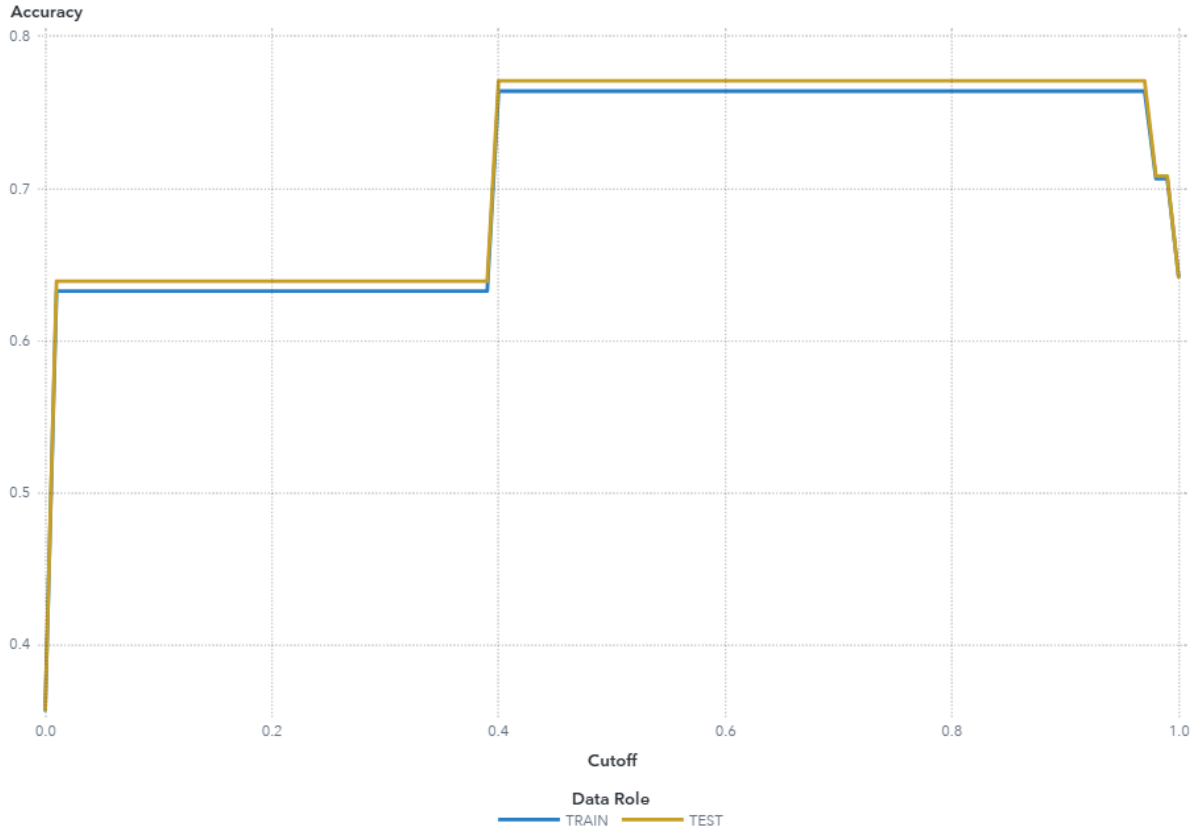
Cutoff values range from 0 to 1, inclusive, in increments of 0.01. At each cutoff value, the predicted target classification is determined by whether P_acci_severity1, which is the predicted probability of the event "1" for the target acci_severity, is greater than or equal to the cutoff value. When P_acci_severity1 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

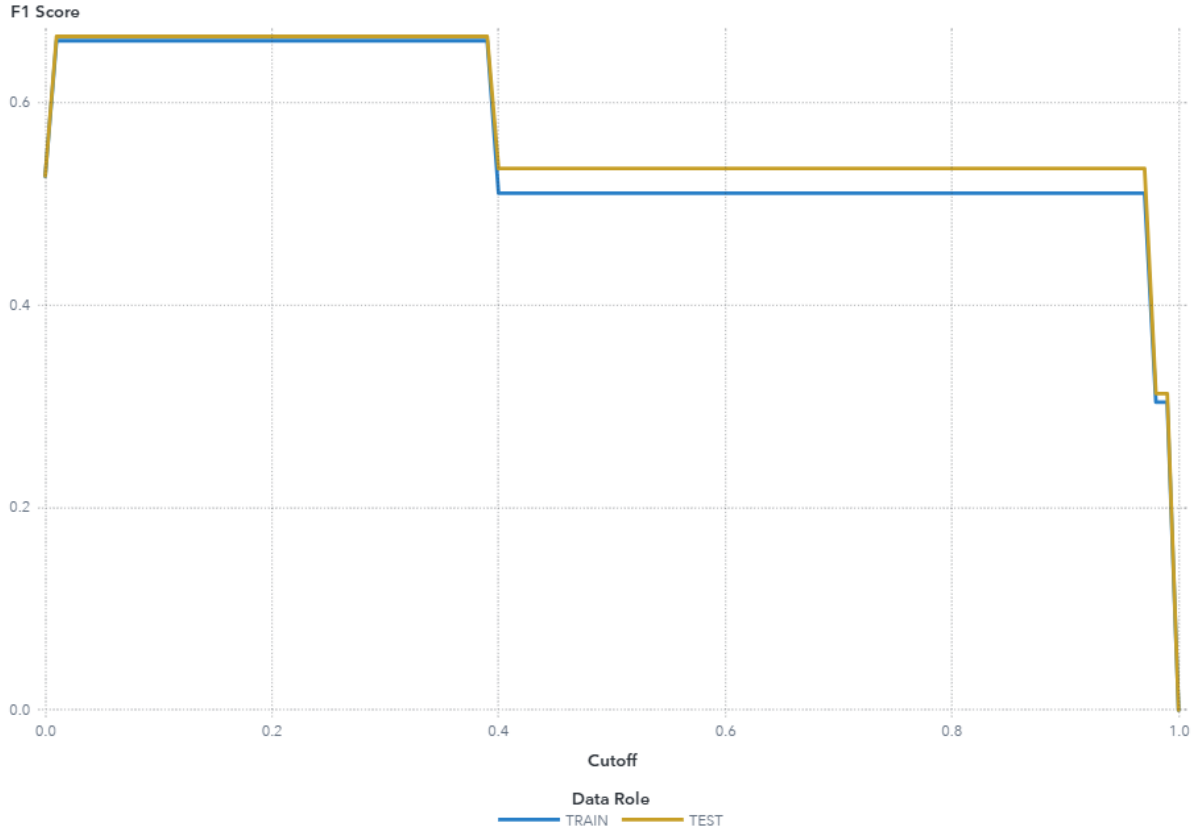


For this model, the accuracy in the TEST partition at the cutoff of 0.5 is 0.771.

For this model, the accuracy in the TRAIN partition at the cutoff of 0.5 is 0.764.

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_acci_severity1, which is the predicted probability of the event "1" for the target acci_severity, is greater than or equal to the cutoff value. When P_acci_severity1 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 TEST partition at the cutoff of 0.5 is 0.535.

For this model, the F1 score in the TRAIN partition at the cutoff of 0.5 is 0.51.

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_acci_severity1, which is the predicted probability of the event "1" for the target acci_severity, is greater than or equal to the cutoff value. When P_acci_severity1 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	Partition Indicator	Formatted Partition
acci_severity	TEST	2	2
acci_severity	TRAIN	1	1

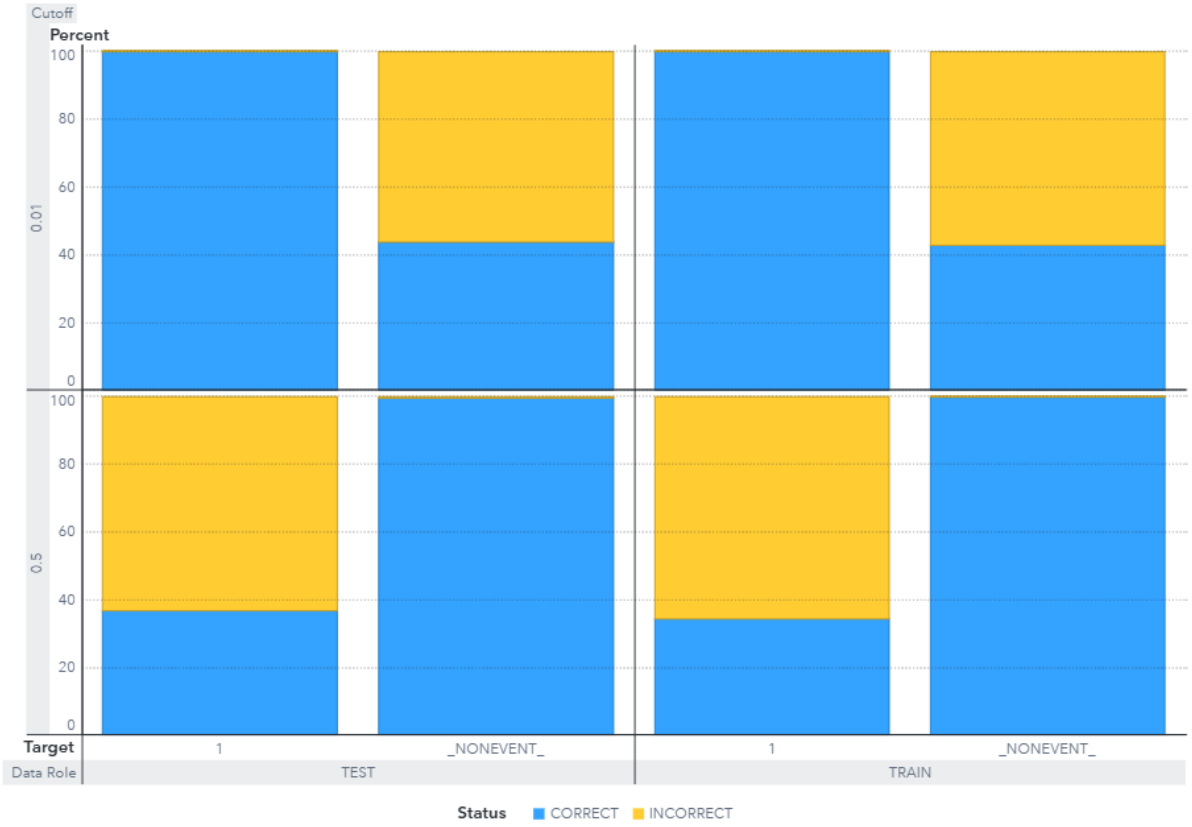
Number of Observations	Average Squared Error	Divisor for ASE	Root Average Squared Error
1,056	0.1768	1,056	0.4205
2,467	0.1736	2,467	0.4167

Misclassification Rate	Multi-Class Log Loss	KS (Youden)	Area Under ROC
0.4782	0.8457	0.4381	0.8205
0.4682	0.8342	0.4284	0.8116

Gini Coefficient	Gamma	Tau	KS Cutoff
0.6411	0.9913	0.2950	0.0100
0.6232	0.9960	0.2864	0.0100

KS at Default Cutoff	Misclassification Rate at KS Cutoff (Event)	Misclassification Rate (Event)
0.3633	0.3608	0.2292
0.3416	0.3672	0.2359

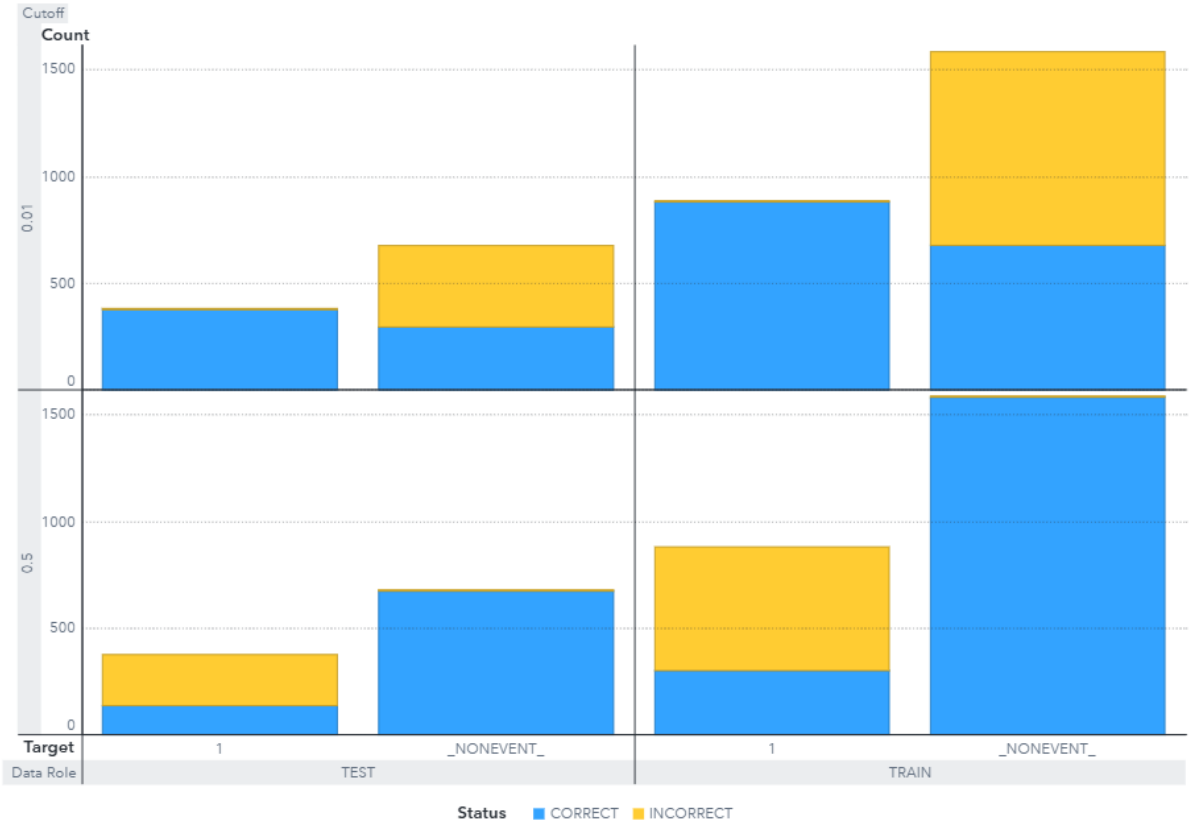
Percentage Plot



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 these KS cutoff values for existing partitions: 0.01 (TRAIN), 0.01 (TEST).

For this data, for the bar corresponding to the event level of acci_severity, " 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 these KS cutoff values for existing partitions: 0.01 (TRAIN), 0.01 (TEST).

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

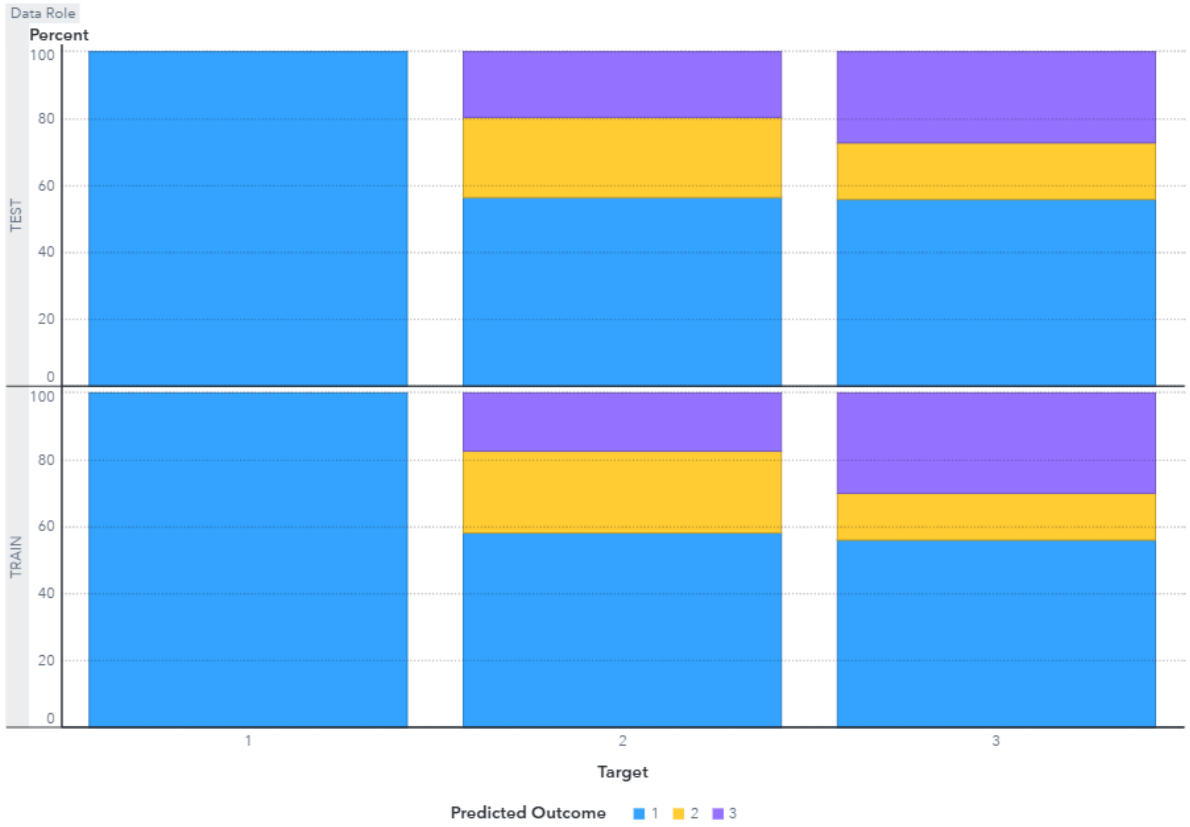
Table

Cutoff	Cutoff Source	Target Name	Response
0.0100	KS	acci_severity	CORRECT
0.0100	KS	acci_severity	INCORRECT
0.0100	KS	acci_severity	CORRECT
0.0100	KS	acci_severity	INCORRECT
0.5000	Default	acci_severity	CORRECT
0.5000	Default	acci_severity	INCORRECT
0.5000	Default	acci_severity	CORRECT
0.5000	Default	acci_severity	INCORRECT

Event	Value	Training Frequency	Validation Frequency
1	True Positive	882	
1	False Negative	0	
NONEVENT	True Negative	679	
NONEVENT	False Positive	906	
1	True Positive	303	
1	False Negative	579	
NONEVENT	True Negative	1,582	
NONEVENT	False Positive	3	

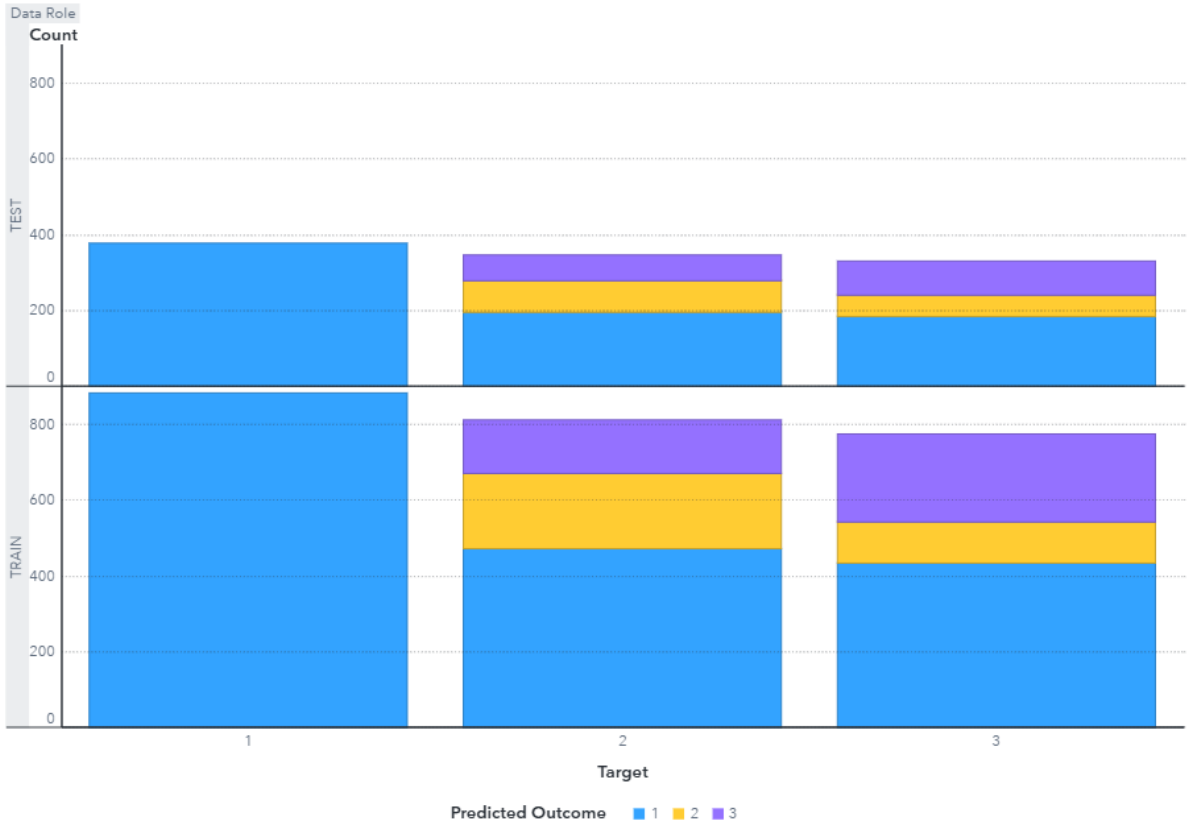
Test Frequency	Training Percentage	Validation Percentage	Test Percentage
378	100		100
0	0		0
297	42.8391		43.8053
381	57.1609		56.1947
139	34.3537		36.7725
239	65.6463		63.2275
675	99.8107		99.5575
3	0.1893		0.4425

Percentage Plot



The Nominal Classification report displays either the percentage of or the number of observations predicting each target level. The plot is segmented by target level and partition level. The target level with the greatest predicted probability is the predicted outcome. A greater number of observations where the target and predicted outcome are the same indicates a better model.

Count Plot



The Nominal Classification report displays either the percentage of or the number of observations predicting each target level. The plot is segmented by target level and partition level. The target level with the greatest predicted probability is the predicted outcome. A greater number of observations where the target and predicted outcome are the same indicates a better model.

Table

Target Name	Data Role	Target	Unformatted Target
acci_severity	TRAIN	1	1
acci_severity	TRAIN	2	2
acci_severity	TRAIN	2	2
acci_severity	TRAIN	2	2
acci_severity	TRAIN	3	3
acci_severity	TRAIN	3	3
acci_severity	TRAIN	3	3
acci_severity	TEST	1	1
acci_severity	TEST	2	2
acci_severity	TEST	2	2
acci_severity	TEST	2	2
acci_severity	TEST	3	3
acci_severity	TEST	3	3
acci_severity	TEST	3	3

Predicted Outcome	Count	Percent	Status
1	882	100	CORRECT
1	472	58.1998	INCORRECT
2	198	24.4143	CORRECT
3	141	17.3859	INCORRECT
1	434	56.0724	INCORRECT
2	108	13.9535	INCORRECT
3	232	29.9742	CORRECT
1	378	100	CORRECT
1	196	56.4841	INCORRECT
2	83	23.9193	CORRECT
3	68	19.5965	INCORRECT
1	185	55.8912	INCORRECT
2	56	16.9184	INCORRECT

Predicted Outcome	Count	Percent	Status
3	90	27.1903	CORRECT

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
atleafSize	false

Property Name	Property Value
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
editedInteractively	false
embeddedBarChart	true
exactPctlLift	true
explainFidelity	false
explainInfo	false
fullDatasetReconstitution	false
hLeafSize	5
iCriterionMethod	VARIANCE
icePlots	false
inodeColor	AVERAGE

Property Name	Property Value
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
train	true
truncateLl	5
truncateUl	95

Property Name	Property Value
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	16

	Training	Test	Total
Number of Observations Read	2467	1056	3523
Number of Observations Used	2467	1056	3523

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
25	2E-11	.	0.3745	0.4494	0.0395	0.5176
21	0.00116	.	0.3786	0.4513	0.0385	0.5176
17	0.00157	.	0.3786	0.4496	0.0395	0.5098
16	0.00222	*	0.3786	0.4479	0.0401	0.5098
14	0.00281	.	0.3868	0.4520	0.0408	0.5137
12	0.00344	.	0.3868	0.4571	0.0432	0.5333
11	0.00421	.	0.4033	0.4632	0.0417	0.5333
10	0.00711	.	0.4074	0.4705	0.0388	0.5373
7	0.0136	.	0.4115	0.5030	0.0538	0.6130
5	0.0197	.	0.4115	0.5355	0.0608	0.6130
3	0.0219	.	0.4115	0.5765	0.0689	0.6453
1	2.2092	.	0.6211	0.6426	0.0198	0.6872

* Selected pruning parameter

The SAS System

The TREESPLIT Procedure

Fit Statistics for Selected Tree		
	Number of Leaves	Misclassification Rate
Training	16	0.4682
Test	16	0.4782

Variable Importance			
	Training		
Variable	Importance	Relative Importance	Count
hour_of_day	85.4301	1.0000	2
junc_detail	45.5964	0.5337	2
loc_auth_ons_distr	43.9653	0.5146	1
longitude	34.5935	0.4049	1
first_road_num	30.3521	0.3553	1
num_of_casu	28.3856	0.3323	1
weath_con	26.5254	0.3105	1
speed_limit	19.9007	0.2329	1
urb_or_rur_area	14.0177	0.1641	1
ped_cross_hum_con	10.8281	0.1267	1
latitude	7.2924	0.0854	1
road_type	7.0448	0.0825	1
carri_haz	3.4867	0.0408	1

The SAS System

The TREESPLIT Procedure

Predicted Probability Variables	
acci_severity	Variable
1	P_acci_severity1
3	P_acci_severity3
2	P_acci_severity2

Predicted Target Variable	
Level Index	Variable
	I_acci_severity