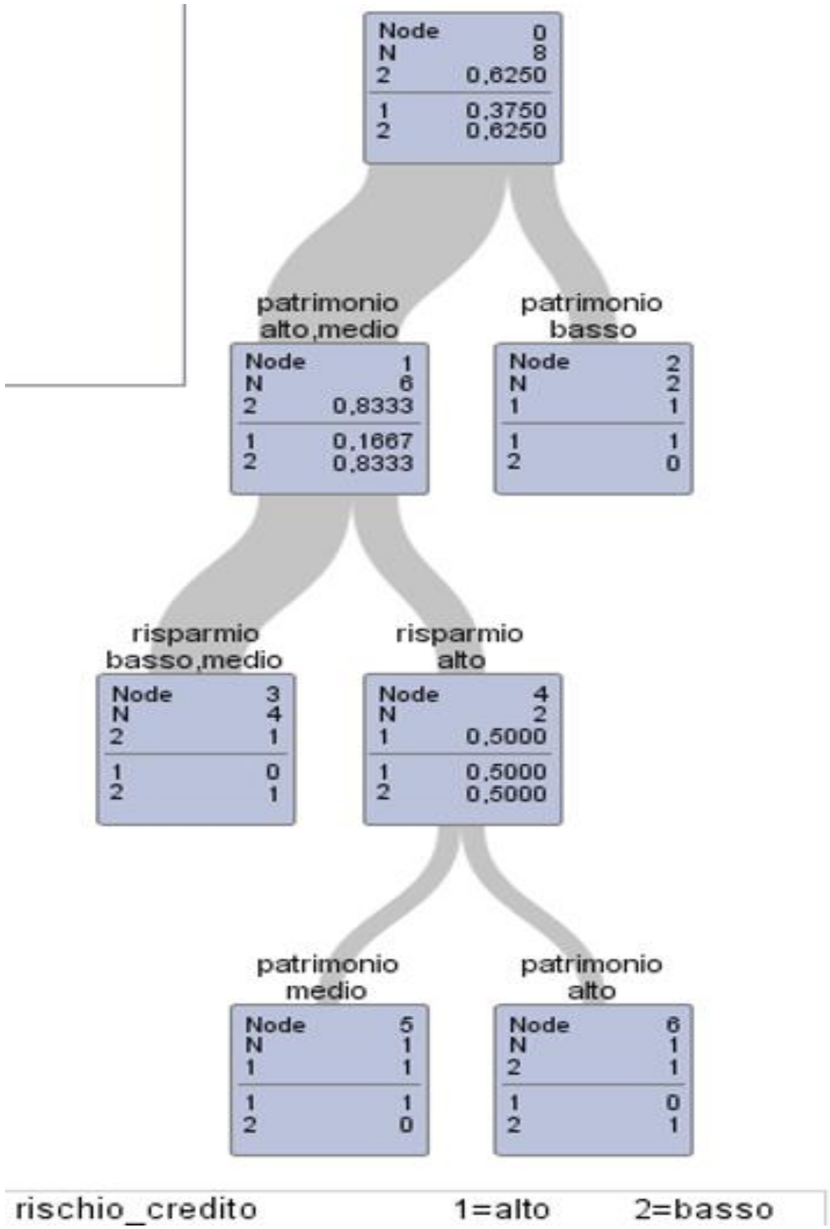
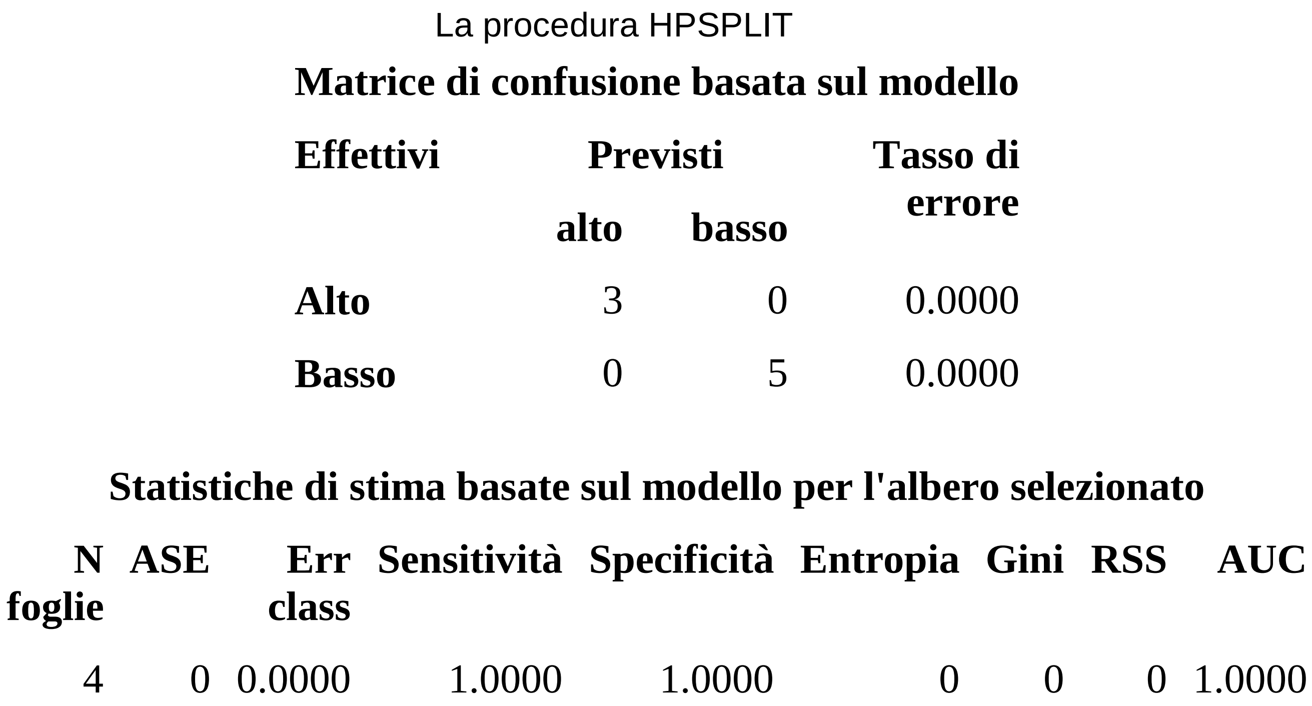
Alberi decisionali: Osservazioni Operative

Alberi di classificazione



Alberi di classificazione



La procedura HPSPLIT

Matrice di confusione basata sul modello

Effettivi	Previsti		Tasso di errore
	alto	basso	
Alto	3	0	0.0000
Basso	0	5	0.0000

Statistiche di stima basate sul modello per l'albero selezionato

N foglie	ASE	Err class	Sensitività	Specificità	Entropia	Gini	RSS	AUC
4	0	0.0000	1.0000	1.0000	0	0	0	1.0000

Alberi di regressione

La procedura HPSPLIT

Matrice di confusione basata sul modello

Effettivi	Previsti		Tasso di errore
	alto	basso	
Alto	3	0	0.0000
Basso	0	5	0.0000

Statistiche di stima basate sul modello per l'albero selezionato

N foglie	ASE	Err class	Sensitività	Specificità	Entropia	Gini	RSS	AUC
4	0	0.0000	1.0000	1.0000	0	0	0	1.0000

- GROW criterio
- var. categoriali e quantitative
  - CHAID
- var. categoriali
  - CHISQUARE
  - ENTROPY (default)
  - FASTCHAID
  - BONFERRONI
  - GINI
  - IGR
- var. quantitative
  - FTEST
  - RSS (default)
  - VARIANCE

In generale, l’indice di eterogeneità di Gini è così definito:

$$G = 1 - \sum_{j=1}^{k} f^2_j$$

dove f_j è la frequenza relativa dell’j-esima modalità di una variabile qualitativa con k modalità. G assume tutti i valori compresi tra zero (massima omogeneità) e (k-1)/k (massima eterogeneità).

Indice relativo
$G' = G/G_{max}$

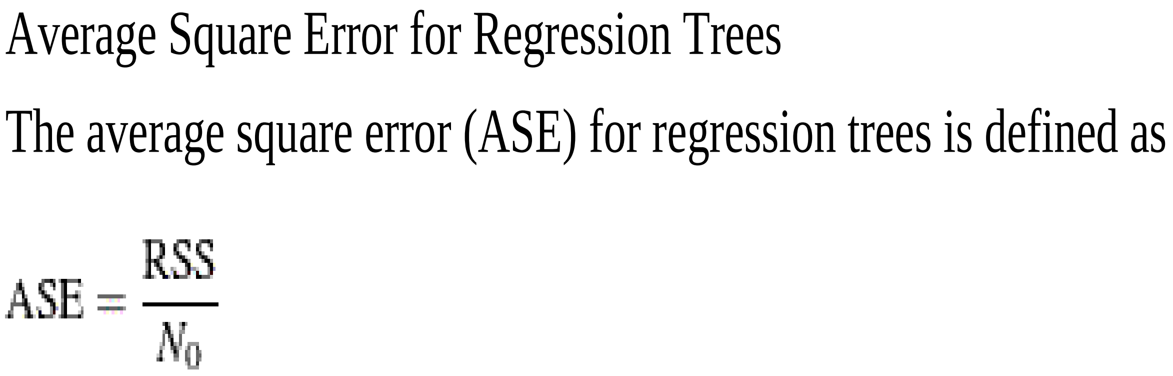
In generale, l'entropia di Shannon è così definita

$$H = \sum_{j=1}^{k} f_j log\ 1/f_j = -\sum_{j=1}^{k} f_j log\ f_j$$

dove f_j è la frequenza relativa dell'j-esima modalità di una variabile qualitativa con k modalità. H assume tutti i valori compresi tra zero (massima omogeneità) e log k (massima eterogeneità).

Indice relativo
$H' = H/H_{max}$

- ASE (SAS Guide)

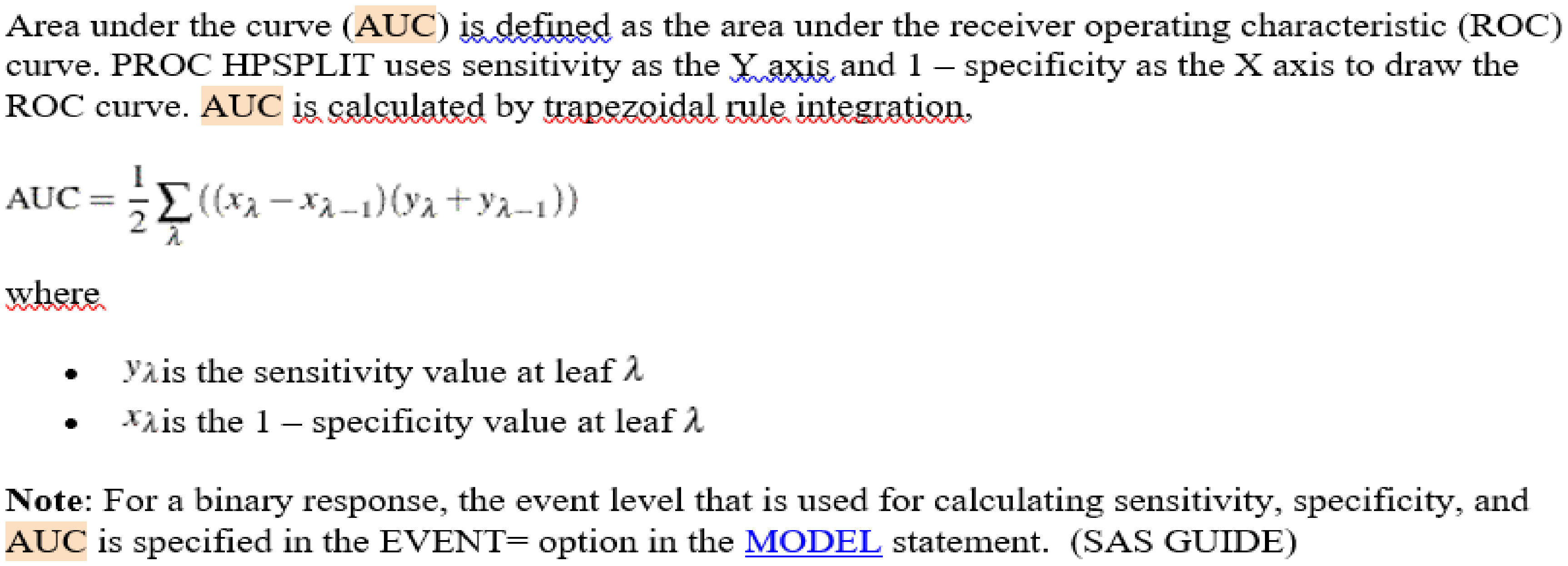


Average Square Error for Regression Trees

The average square error (ASE) for regression trees is defined as

$$\text{ASE} = \frac{\text{RSS}}{N_0}$$

Area sotto la curva AUC



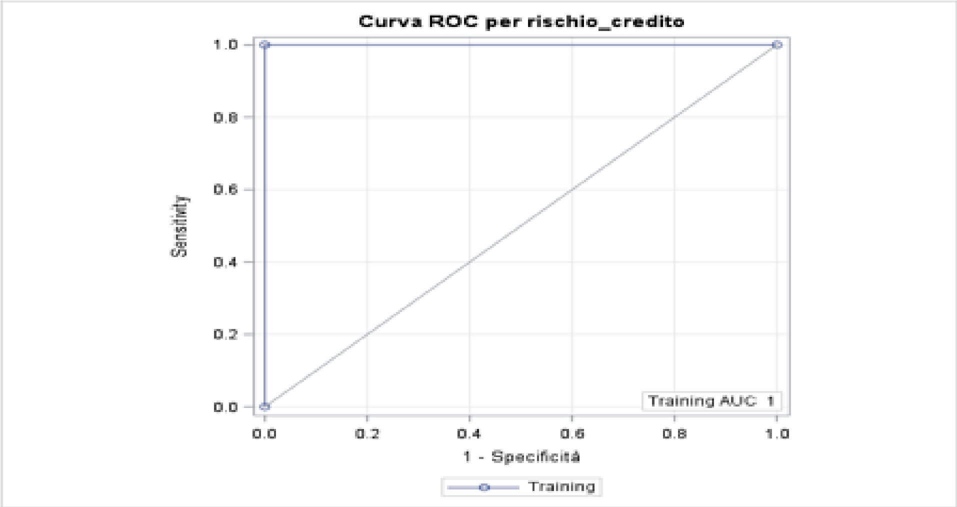
Area under the curve (AUC) is defined as the area under the receiver operating characteristic (ROC) curve. PROC HPSPLIT uses sensitivity as the Y axis and 1 – specificity as the X axis to draw the ROC curve. AUC is calculated by trapezoidal rule integration.

$$\mathrm{AUC} = \frac{1}{2}\sum_{\lambda}((x_{\lambda} - x_{\lambda-1})(y_{\lambda} + y_{\lambda-1}))$$

where

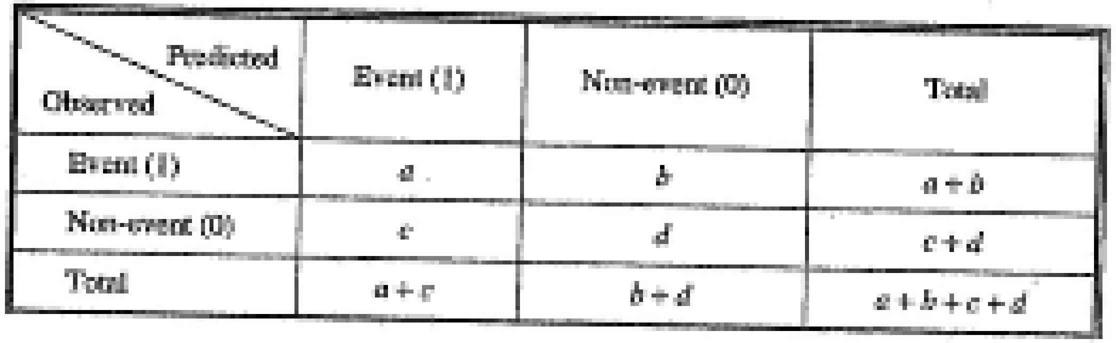
- $y_{\lambda}$is the sensitivity value at leaf λ
- $x_{\lambda}$is the 1 – specificity value at leaf λ

Note: For a binary response, the event level that is used for calculating sensitivity, specificity, and AUC is specified in the EVENT= option in the MODEL statement. (SAS GUIDE)

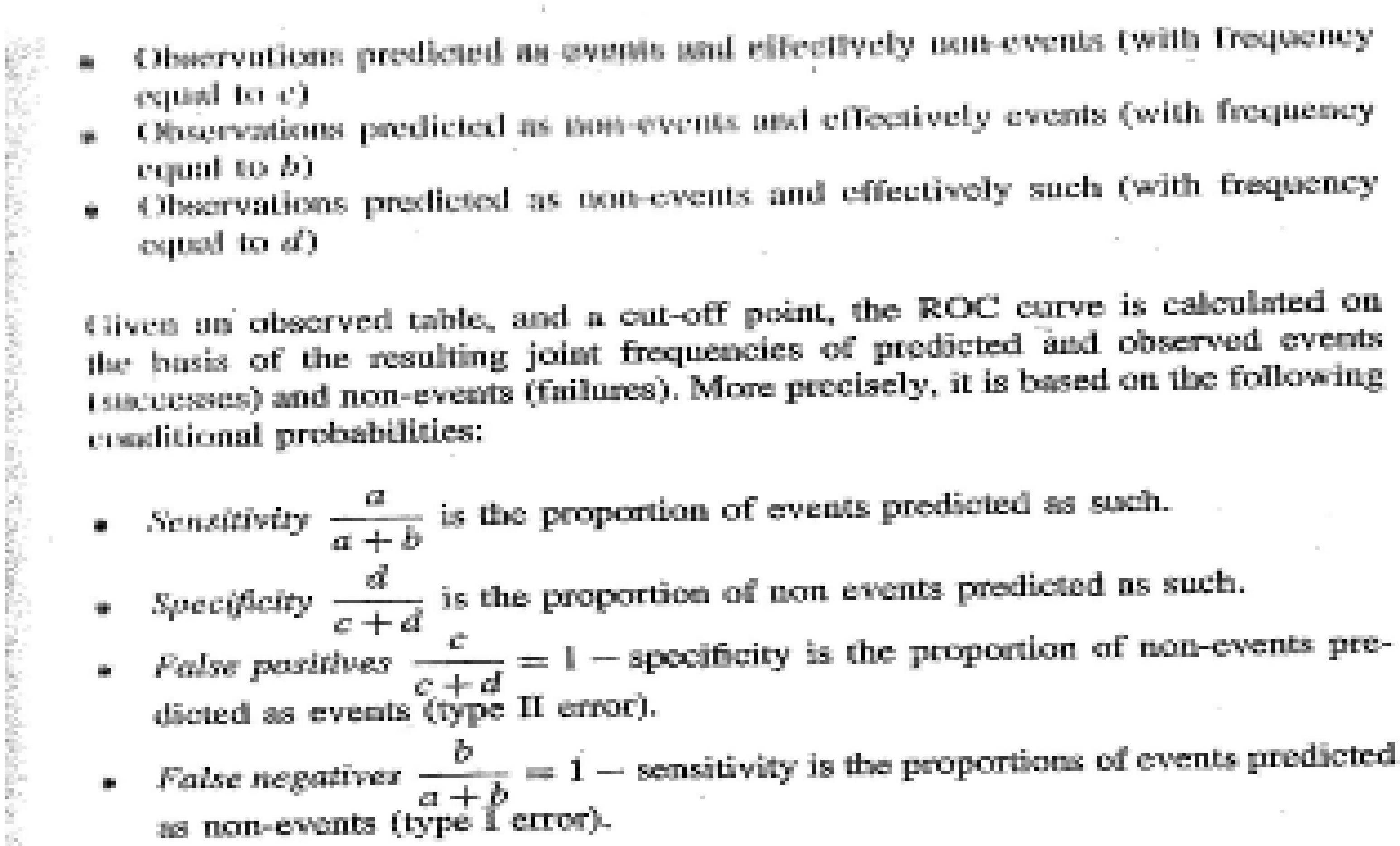


La curva di ROC si basa sulla matrice di confusione

ROC curve



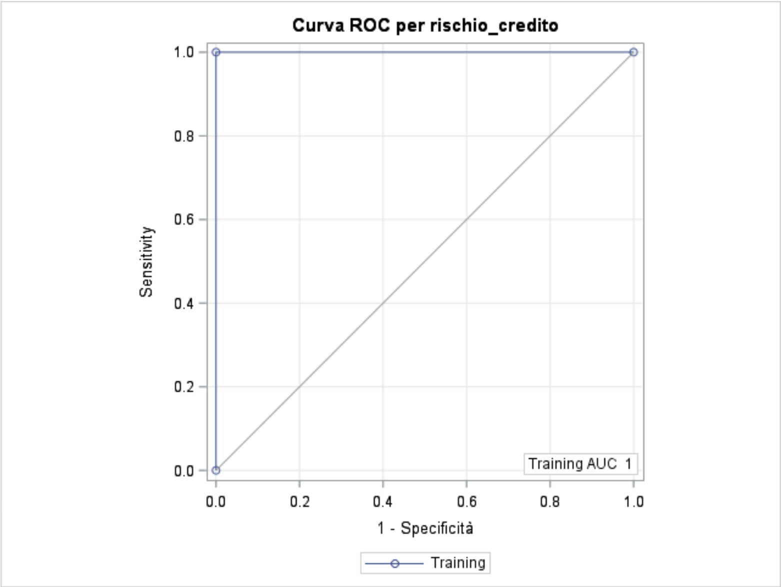
Observed \ Predicted	Event (1)	Non-event (0)	Total
Event (1)	a	b	$a+b$
Non-event (0)	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$



- Observations predicted as events and effectively non-events (with frequency equal to c)
- Observations predicted as non-events and effectively events (with frequency equal to b)
- Observations predicted as non-events and effectively such (with frequency equal to d)

Given an observed table, and a cut-off point, the ROC curve is calculated on the basis of the resulting joint frequencies of predicted and observed events (successes) and non-events (failures). More precisely, it is based on the following conditional probabilities:

- *Sensitivity* $\frac{a}{a+b}$ is the proportion of events predicted as such.
- *Specificity* $\frac{d}{c+d}$ is the proportion of non events predicted as such.
- *False positives* $\frac{c}{c+d} = 1 -$ specificity is the proportion of non-events predicted as events (type II error).
- *False negatives* $\frac{b}{a+b} = 1 -$ sensitivity is the proportions of events predicted as non-events (type I error).



- SPLITTING CRITERIA
  - Criteri basati sull’impurità: (classification trees)

    GINI

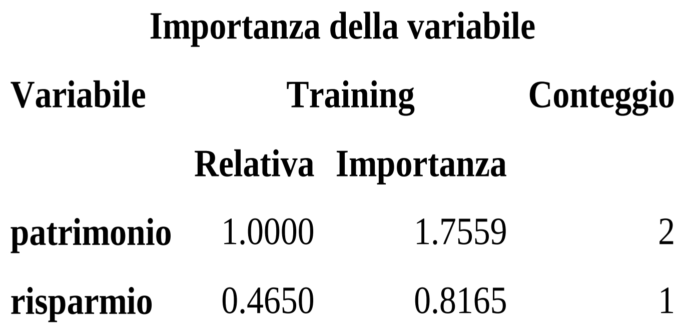
    Entropia (default) etc..
  - Criteri basati sull’impurità: (regression trees)
  - RSS (default) etc..

- Criteri basati su test Statistici
  - CHI-SQUARE criterion (categorical var.)
  - F-test criterion (continuous var.)
  - CHAID criterion (categorical and continuous var.)

Importanza della variabile

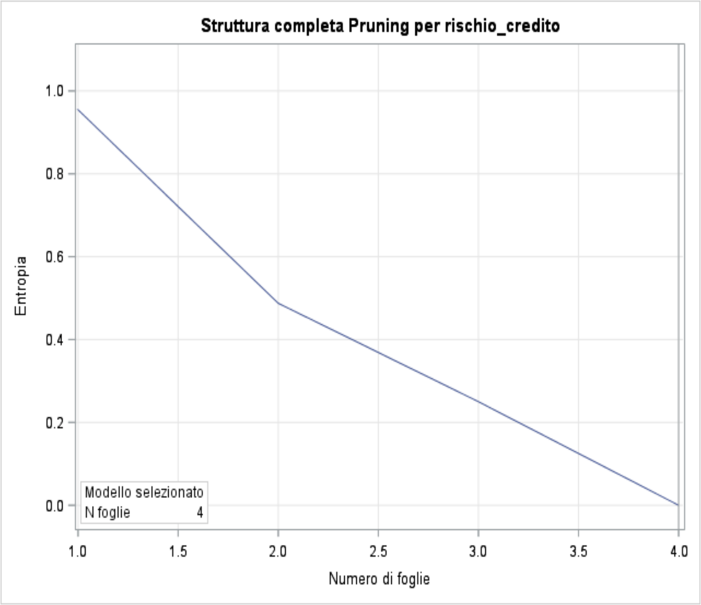
Variable Importance

A training data set can contain a large number of predictors. Some predictors are useful for predicting the response variable, and others are not. You can use the HPSPLIT procedure to select the most useful predictors based on variable importance. Variable importance is an indication of which predictors are most useful for predicting the response variable. Various measures of variable importance have been proposed in the data mining literature (SAS Guide)



Importanza della variabile

Variabile	Training		Conteggio
	Relativa	Importanza	
patrimonio	1.0000	1.7559	2
risparmio	0.4650	0.8165	1



Importanza della variabile

The most important variables might not be the ones near the top of the tree. PROC HPSPLIT measures variable importance based on the following metrics:

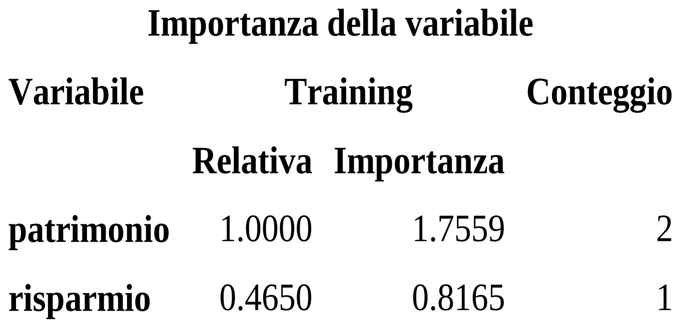
count,

surrogate count,

RSS

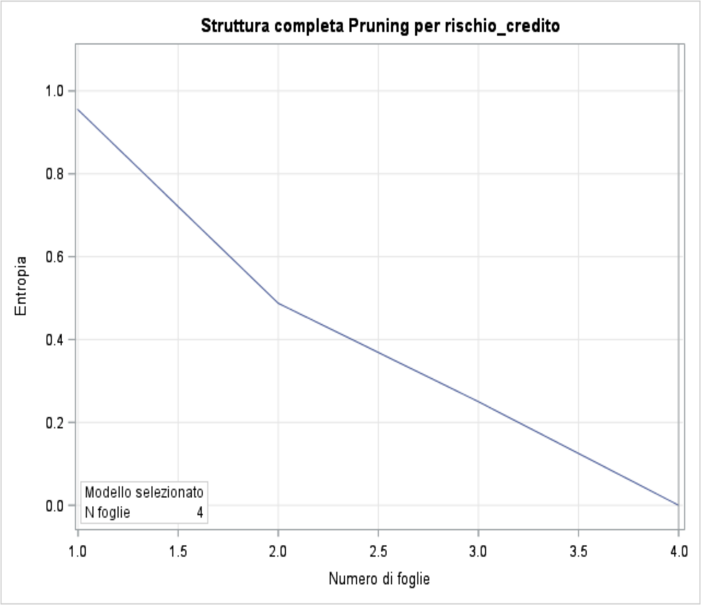
relative importance.

The count-based variable importance simply counts the number of times in the tree that a particular variable is used in a split. Similarly, the surrogate count tallies the number of times that a variable is used in a surrogate splitting rule. (SAS Guide)



Importanza della variabile

Variabile	Training		Conteggio
	Relativa	Importanza	
patrimonio	1.0000	1.7559	2
risparmio	0.4650	0.8165	1

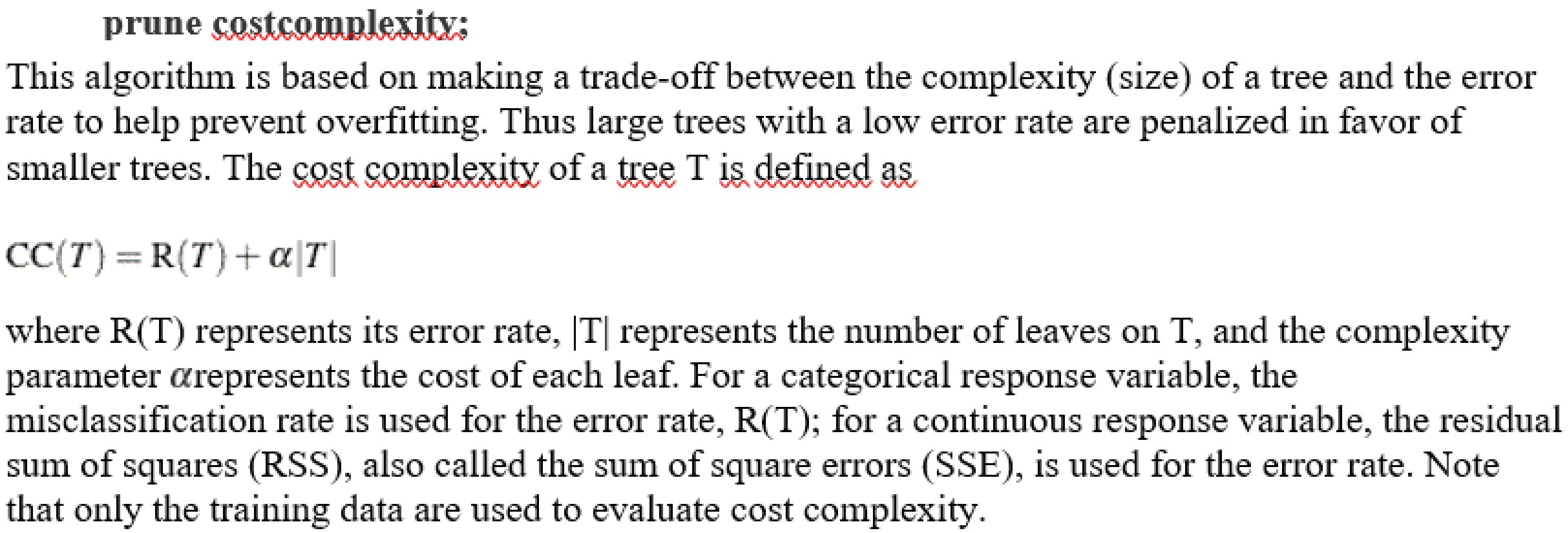


PRUNE statement

PRUNE statement (SAS Guide)

- C45 (classification trees)

- COSTCOMPLEXITY



prune costcomplexity;

This algorithm is based on making a trade-off between the complexity (size) of a tree and the error rate to help prevent overfitting. Thus large trees with a low error rate are penalized in favor of smaller trees. The cost complexity of a tree T is defined as

$$\mathrm{CC}(T) = \mathrm{R}(T) + \alpha|T|$$

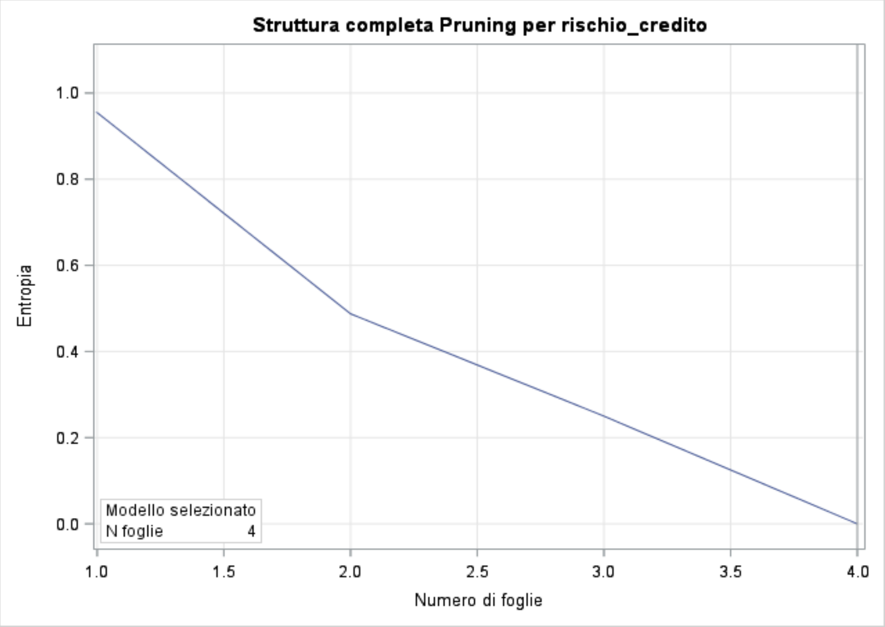
where R(T) represents its error rate, |T| represents the number of leaves on T, and the complexity parameter αrepresents the cost of each leaf. For a categorical response variable, the misclassification rate is used for the error rate, R(T); for a continuous response variable, the residual sum of squares (RSS), also called the sum of square errors (SSE), is used for the error rate. Note that only the training data are used to evaluate cost complexity.

- CC (classification e regression trees)

- CHI-SQUARE …..

Importanza della variabile

Variabile	Training		Conteggio
	Relativa	Importanza	
patrimonio	1.0000	1.7559	2
risparmio	0.4650	0.8165	1



- OUTPUT statement
  - OUTPUT out=dataset_sas
- PARTITION statement
- CODE statement
- PERFORMANCE statement
- PRUNE statement
- Etc…