Gain Summary

The Gain Summary summarizes a tree by displaying descriptive statistics for each terminal node. This allows users to recognize the relative contribution of each terminal node and identify the subsets of terminal nodes that are most useful. This document can be used for all tree growing algorithms CART, CHAID, exhaustive CHAID and QUEST.

Note that case weight is not involved in gain summary calculations though it is involved in tree growing process and class assignment.

Types of Gain Summaries

Depending on the type of dependent variable, different statistics are given in the gain summary.

Average Oriented Gain Summary (Y continuous)

Statistics related to the node mean of Y are given. Through this summary, users may identify the terminal nodes that give the largest (or smallest) average of the dependent variable.

Target Class Gain Summary (Y categorical)

Statistics related to an interested dependent variable class (target class) are given. Users may identify the terminal nodes that have a large relative contribution to the target class.

Average Profit Value Gain Summary (Y categorical)

Statistics related to average profits are given. Users may be interested in identifying the terminal nodes that have relatively large average profit values.

Node-by-Node, Cumulative, Percentile Gain Summary

To assist users in identifying the interesting terminal nodes and in understanding the result of a tree, three different ways (node-by-node, cumulative and percentile) of looking at the gain summaries mentioned above are provided.

Notations

Y	The dependent variable, or target variable. It can be either categorical (nominal or ordinal) or continuous. If Y is categorical with J classes, its class takes values in C = {1, …, J}.
D	Data set used to calculate gain statistics. It can be either learning sample data set or test sample data set.
D(t)	Cases in D fallen in node t.

y_n	The dependent variable value for case n.
f_n	The frequency weight associated with case n. Non-integral positive value is rounded to its nearest integer.
N_f	The number of cases in *D*, $N_f = \sum_{n \in D} f_n$
$N_f(t)$	The number of cases in *D*(t), $N_f(t) = \sum_{n \in D(t)} f_n$
$N_{f,j}$	The number of class j cases in *D*, $N_{f,j} = \sum_{n \in D} f_n I(y_n = j)$
$N_{f,j}(t)$	The number of class j cases in *D*(t), $N_{f,j}(t) = \sum_{n \in D(t)} f_n I(y_n = j)$
$\bar{y}(t)$	The mean of dependent variable in D(t), $\bar{y}(t) = \frac{1}{N_f(t)} \sum_{n \in D(t)} f_n y_n$
j''	Target class of interest, it is any value in {1, …, J}. Target class j'' is user-specified. If not, the default target class is j'' = 1.
$r(j)$, $e(j)$	They are respectively the revenue and expense associated with class j.
$pv(j)$	The profit value associated with class j, $pv(j) = r(j) - e(j)$.
$j^*(\tilde{t})$	Class assignment given by terminal node $\tilde{t}$.
$\pi(j)$	Prior probability of class j Y = j, j = 1, …, J.
M1	For categorical Y, denote empirical prior situation. CHAID and exhaustive CHAID always considered as having empirical prior.
M2	For categorical Y, denote non-empirical prior situation.

Gain Summary: Node by Node

The node-by-node gain summary includes statistics for each node that are defined below.

Terminal Node

The identity of a terminal node. It is denoted by $\tilde{t}$.

Size: n

Total number of cases in the terminal node. It is denoted by $N_f(\tilde{t})$.

Size: %

Percentage of cases in the node. It is denoted by $p_f(\tilde{t})100\%$, where $p_f(\tilde{t})$ is given by

$$p_f(\tilde{t}) = \begin{cases} \dfrac{N_f(\tilde{t})}{N_f} & \text{M1, or, Y continuous} \\ \displaystyle\sum_j \frac{\pi(j) N_{f,j}(\tilde{t})}{N_{f,j}} & \text{M2} \end{cases}$$

Gain: n (for target class gain summary only)

Total number of target class j'' cases in the node, $N_{f,j''}(\tilde{t})$.

Gain: % (for target class gain summary only)

Percentage of target class j'' cases in the sample that belong to the node. It is denoted by $p_f(\tilde{t} \mid j'')$ 100%, where $p_f(\tilde{t} \mid j'')$ is

$$p_f(\tilde{t} \mid j'') = \frac{N_{f,j''}(\tilde{t})}{N_{f,j''}}.$$

Score

Depending on the type of gain summary, the score is defined and named differently. But they are all denoted by $s(\tilde{t})$.

Response: % (for target class gain summary only)

The ratio of target class j'' cases in the node to the total number of cases in the node.

$$s(\tilde{t}) = \begin{cases} \dfrac{N_{f,j''}(\tilde{t})}{N_f(\tilde{t})} & \text{M1} \\ \dfrac{1}{p_f(\tilde{t})} \cdot \dfrac{\pi(j'') N_{f,j''}(\tilde{t})}{N_{f,j''}} & \text{M2} \end{cases}.$$

Average Profit (for average profit value gain summary only)

The average profit value for the node.

$$s(\tilde{t}) = \begin{cases} \dfrac{\displaystyle\sum_j N_{f,j}(\tilde{t}) \cdot pv(j)}{N_f(\tilde{t})} & \text{M1} \\ \dfrac{1}{p_f(\tilde{t})} \cdot \displaystyle\sum_j \frac{\pi(j) N_{f,j}(\tilde{t}) \cdot pv(j)}{N_{f,j}} & \text{M2} \end{cases}.$$

Mean (for average oriented gain summary only)

The respective mean $\bar{y}(\tilde{t})$ of the continuous dependent variable *Y* at the node.

$$s(\tilde{t}) = \bar{y}(\tilde{t}).$$

ROI (Return on Investment, for average profit value gain summary only)

ROI for a node is calculated as average profit divided by average expense.

$$ROI(\tilde{t}) = \frac{s(\tilde{t})}{s_0(\tilde{t})}.$$

Where $s_0(\tilde{t})$ is the average expense for node $\tilde{t}$ and is calculated using equation for $s(\tilde{t})$ with *pv(j)* replaced by *e(j)*.

Index (%)

For target class gain summary, it is the ratio (in %) of score for the node to the proportion of class j'' cases in the sample. It is denoted by *is*($\tilde{t}$)100%, where *is*($\tilde{t}$) is

$$is(\tilde{t}) = \begin{cases} \dfrac{s(\tilde{t})}{N_{f,j''}/N_f} & \text{M1} \\ \dfrac{s(\tilde{t})}{\pi(j'')} & \text{M2} \end{cases}.$$

For average profit value gain summary, it is the ratio (in %) of score for the node to the average profit value for the sample.

$$is(\tilde{t}) = \begin{cases} \dfrac{s(\tilde{t})}{\sum_j N_{f,j} pv(j)/N_f} & \text{M1} \\ \dfrac{s(\tilde{t})}{\sum_j \pi(j) pv(j)} & \text{M2} \end{cases}.$$

For average oriented gain summary, it is the ratio (in %) of the gain score for the node to the gain score $s(t = 1)$ for root node $t = 1$.

$$is(\tilde{t}) = \frac{s(\tilde{t})}{s(t=1)}.$$

Notice that if the denominator is 0, the index is not available.

Gain Summary: Cumulative

In the cumulative gain summary, all nodes are first sorted with respect to the values of score $s(\tilde{t})$. To simplify the formulas, we assume that nodes in the collection $\{\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_{|\tilde{T}|}\}$ are already sorted either in descending or ascending order according to user-request.

Terminal Node

The identity of a terminal node. It is denoted by $\tilde{t}_s$.

Cumulative Size: n, Cumulative Size: %, Cumulative gain: n, Cumulative gain: %

These statistics are simply defined as the cumulative sum of corresponding node-by-node items up to the terminal node of interest. Let $a(\tilde{t}_i)$ be the node-by-node statistics, then its cumulative count part up to node $\tilde{t}_s$ is $\oplus a(\tilde{t}_s) = \sum_{i=1}^{s} a(\tilde{t}_i)$. These four cumulative statistics are denoted respectively by $\oplus N_f(\tilde{t}_s)$, $\oplus p_f(\tilde{t}_s)$, $\oplus N_{f,j''}(\tilde{t}_s)$ and $\oplus p_f(\tilde{t}_s \mid j'')$.

Cumulative Score

For Cumulative response, it is the ratio of target class j'' cases up to the node to the total number of cases up to the node. For cumulative average profit, it is the average profit value up to the node. For cumulative mean, it is the mean of all y_n's up to the nodes $\tilde{t}_s$. In all cases, the same formula is used. However, readers should use the appropriate formulas for $s(\tilde{t})$ and $p_f(\tilde{t})$ in the calculations. This cumulative score is denoted by $\oplus s(\tilde{t}_s)$.

$$\oplus s(\tilde{t}_s) = \begin{cases} \dfrac{\sum_{i=1}^{s} s(\tilde{t}_i) \cdot N_f(\tilde{t}_i)}{\sum_{i=1}^{s} N_f(\tilde{t}_i)} & \text{M1, or, Y continuous} \\ \dfrac{\sum_{i=1}^{s} s(\tilde{t}_i) \cdot p_f(\tilde{t}_i)}{\sum_{i=1}^{s} p_f(\tilde{t}_i)} & \text{M2} \end{cases} .$$

Cumulative ROI (for average profit value gain summary only)

Cumulative ROI up to a node is

$$\oplus ROI(\tilde{t}_s) = \frac{\oplus s(\tilde{t}_s)}{\oplus s_0(\tilde{t}_s)}.$$

Where $\oplus s_0(\tilde{t}_s)$ is the cumulative expense and calculated by equation for $\oplus s(\tilde{t}_s)$ with $pv(\tilde{t})$ replaced by $e(\tilde{t})$.

Cumulative Index (%)

For target class cumulative gain summary, it is the ratio (in %) of cumulative gain score for the node to the proportion of class j'' cases in the sample. It is denoted by $\oplus is(\tilde{t}_s)100\%$, where $\oplus is(\tilde{t}_s)$ is

$$\oplus is(\tilde{t}_s) = \begin{cases} \dfrac{\oplus s(\tilde{t}_s)}{N_{f,j''}/N_f} & \text{M1} \\ \dfrac{\oplus s(\tilde{t}_s)}{\pi(j'')} & \text{M2} \end{cases}.$$

For average profit value cumulative gain summary, it is the ratio (in %) of cumulative gain score for the node to the average profit value for the sample.

$$\oplus is(\tilde{t}_s) = \begin{cases} \dfrac{\oplus s(\tilde{t}_s)}{\sum_j N_{f,j} \cdot pv(j)/N_f} & \text{M1} \\ \dfrac{\oplus s(\tilde{t}_s)}{\sum_j \pi(j) \cdot pv(j)} & \text{M2} \end{cases}.$$

For average oriented cumulative gain summary, it is the ratio (in %) of cumulative score for the node to the score $s(t = 1)$ for root node $t = 1$.

$$\oplus is(\tilde{t}_s) = \frac{\oplus s(\tilde{t}_s)}{s(t=1)} = \sum_{i=1}^{s} is(\tilde{t}_i).$$

Notice that if the denominator is 0, the index is not available.

Percentile Gain Summary

Like cumulative gain summary, all nodes are first sorted with respect to the values of their scores. To simplify the formulas, we assume that nodes in the collection $\{\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_{|\tilde{T}|}\}$ are already sorted in either descending or ascending order. Let q be any positive integer divisible to 100. The value of q will be used as the percentage increment for percentiles, and is user-

specified (default q = 10). For fixed q, the number of percentiles to be studied is 100/q. The p-th percentile to be studied is the pq%-tile, and its size is $N_{f\cdot pq} = N_f \cdot pq\%$, p = 1, …, 100/q. For any pq%-tile, let s_p and s'_p be the two smallest integers in {1, …, $|\tilde{T}|$} such that

$$N_{f\cdot pq} \in \left(\oplus N_f(\tilde{t}_{s_p-1}), \oplus N_f(\tilde{t}_{s_p})\right], \quad N_{f\cdot pq} \in \left[\oplus N_f(\tilde{t}_{s'_p-1}), \oplus N_f(\tilde{t}_{s'_p})\right)$$

where $\oplus N_f(\tilde{t}_0) \equiv 0$.

Terminal Nodes

The identity of all terminal nodes that belong to the p[th] increment. Node $\tilde{t}$ belongs to the p[th] increment if $\tilde{t} \in [s'_{p-1}, s_p]$.

Percentile (%)

Percentile being studied. The p-th percentile is the pq%-tile.

Percentile: n

Total number of cases in the percentile, $N_{f\cdot pq}$ = [$N_f \cdot pq$%], where [x] denotes the nearest integer of x.

Gains: n (for target class percentile gain summary only)

Total number of class j'' cases in the pq%-tile. It is denoted by $\Diamond N_{f,j''}(p)$.

$$\Diamond N_{f,j''}(p) = \oplus N_{fj''}(\tilde{t}_{s_p\text{-}1}) + \frac{N_{f\cdot pq} - \oplus N_f(\tilde{t}_{s_p\text{-}1})}{N_f(t_{s_p})} N_{f,j''}(\tilde{t}_{s_p})$$

where $\oplus N_{f,j''}(\tilde{t}_0)$ is defined to be 0.

Gains: % (for target class percentile gain summary only)

Percentage of class j'' cases in the sample that belong to the pq%-tile. It is denoted by $\Diamond p_{f,j''}(p)$ 100%, where $\Diamond p_{f,j''}(p)$ is

$$\Diamond p_{f,j''}(p) = \frac{\Diamond N_{f,j''}(p)}{N_{f,j''}}.$$

Percentile score

For target class percentile gain summary, it is an estimate of ratio of the number of class j'' cases in the *pq*%-tile to the total number of cases in the percentile. For average profit value percentile gain summary, it is an estimate of the average profit value in the *pq*%-tile. For average oriented percentile gain summary, it is an estimate of the ratio of gain score for all nodes in the percentile. In all charts, the same formula is used.

$$\Diamond s(p) = \begin{cases} \dfrac{\oplus N_f(\tilde{t}_{s_p-1}) \cdot \oplus s(\tilde{t}_{s_p-1}) + \{N_{f.pq} - \oplus N_f(\tilde{t}_{s_p-1})\} \cdot s(\tilde{t}_{s_p})}{N_{f.pq}} & \text{M1} \\ \dfrac{\oplus p_f(\tilde{t}_{s_p-1}) \cdot \oplus s(\tilde{t}_{s_p-1}) + \{p_{f.pq} - \oplus p_f(\tilde{t}_{s_p-1})\} \cdot s(\tilde{t}_{s_p})}{p_{f.pq}} & \text{M2} \end{cases}$$

where

$$p_{f.pq} = \oplus p_f(\tilde{t}_{s_p-1}) + \frac{N_{f.pq} - \oplus N_f(\tilde{t}_{s_p-1})}{N_f(\tilde{t}_{s_p})} p_f(\tilde{t}_{s_p}).$$

Percentile ROI (for average profit value gain summary only)

The definition of percentile ROI is

$$\Diamond ROI(p) = \frac{\Diamond s(p)}{\Diamond s_0(p)}.$$

Where $\Diamond s_0(p)$ is the percentile expense and calculated through equation $\Diamond s(p)$ with $pv(\tilde{t})$ replaced by $e(\tilde{t})$.

Percentile Index (in %)

For target class percentile gain summary, it is the ratio (in %) of percentile gain score for the *pq*%-tile to the proportion of class j'' cases in the sample. It is denoted by $\Diamond is(p)100\%$, where $\Diamond is(p)$ is

$$\Diamond is(p) = \begin{cases} \dfrac{\Diamond s(p)}{N_{f,j''}/N_f} & \text{M1} \\ \dfrac{\Diamond s(p)}{\pi(j'')} & \text{M2} \end{cases}.$$

For average profit value percentile gain summary, it is the ratio (in %) of percentile gain score for the *pq*%-tile to the average of the profit values for the sample.

$$\Diamond is(p) = \begin{cases} \dfrac{\Diamond s(p)}{\sum_j N_{f,j} \cdot pv(j) / N_f} & \text{M1} \\ \dfrac{\Diamond s(p)}{\sum_j \pi(j) \cdot pv(j)} & \text{M2} \end{cases}.$$

For average oriented cumulative gain summary, it is the ratio (in %) of percentile gain score in the *pq*%-tile to the gain score $s(t = 1)$ for root node $t = 1$.

$$\Diamond is(p) = \frac{\Diamond s(p)}{s(t=1)}.$$

Notice that if the denominator, which is the average or the median of y_n's in the sample, is 0, it may happen for the learning sample or the test sample, the index is not available.