

# *The Hybrid CART-Logit Model in Classification and Data Mining*

**Dan Steinberg**  
**N. Scott Cardell**



<http://www.salford-systems.com>

# *Data mining*

- **Attempt to discover possibly very complex structure in huge databases (large number of records and large number of variables)**
- **Problems include classification, regression, clustering, association (Market basket analysis)**
- **Need tools to partially or fully automate the discovery process**
- **Large databases support search for rare but important patterns**
  - **can use methods which would not work at all on small databases (very slow asymptotic convergence)**
- **Conclusion of a number of studies is that no single method is best for all problems**
  - **thus a collection of tools is required**

# ***CART & LOGIT: Two of the essential tools***

- **CART - nontraditional decision tree methodology, style of new data mining tools**
  - high degree of automation
  - communication via pictures
  - ability to handle arbitrarily complex data structures
  - relatively easy to use and understand even by non-statisticians
  - demonstrates remarkable accuracy in a broad range of contexts
- **LOGIT - traditional methodology relying on classical statistical principles**
  - requires expert to develop hand-crafted models
  - can often be understood only via simulation
  - ability to handle linear and smooth curvilinear data structures
  - demonstrates remarkable accuracy in a broad range of contexts

# *Core CART features*

- **Automatic separation of relevant from irrelevant predictors (variable selection)**
- **Does not require a transform such as log, square root (model specification)**
- **Automatic interaction detection (model specification)**
- **Impervious to outliers (can handle dirty data)**
- **Unaffected by missing values (does not require list-wise deletion or missing value imputation)**
- **Requires only moderate supervision by the analyst**
- **First time model is often as good as a neural net developed by an expert**
- **None of these features shared by LOGIT**

# *Importance of logistic regression in data mining*

- **Logit can be an excellent performer in classification**
- **In STATLOG project, variations of logistic discriminant analysis were most accurate in 5 out of 21 problems**
  - competition included neural nets, decision trees
  - among top 5 performers, 12 trees out of 21 problems
- **Provides a smooth, continuous predicted probability of class membership**
  - A small change in a predictor variable yields a small change in predicted probability
- **Effective capture of global features of data**
  - Main effects model reflects how probability responds to predictor  $x$  over entire range of  $x$
  - Some flexibility allowed with transformation, polynomials and interactions

# *CART and LOGIT excel at different tasks*

- **CART is notoriously weak at capturing strong linear structure**
- **CART recognizes the structure but cannot represent it effectively**
- **With many variables, several of which enter a model linearly, structure will not be obvious from CART output**
- **CART can produce a very large tree in an attempt to represent very simple relationships**
- **LOGIT easily captures and represents linear structure**
- **Many non-linear structures can still be reasonably approximated with a linear structure, hence even incorrectly specified LOGIT can perform well**

# *Natural question-can CART and Logit be combined?*

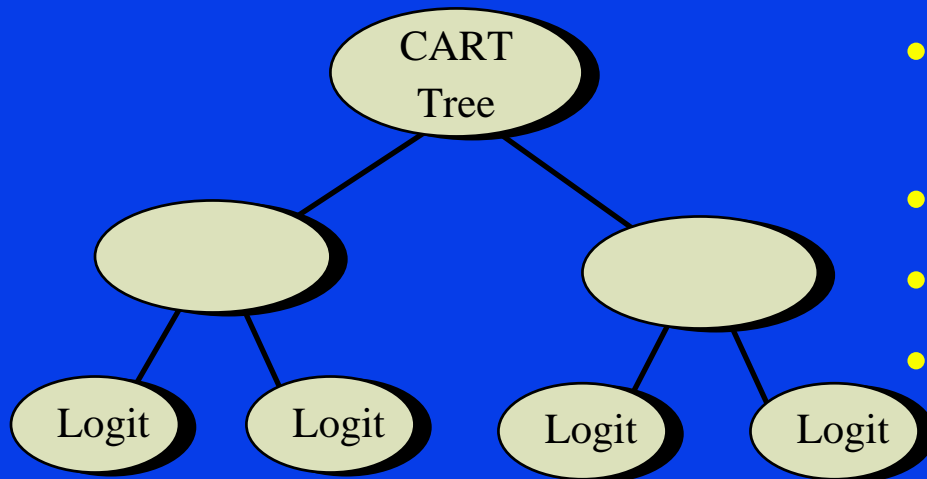
## CART

- Automatic analysis
- Surrogates for missing values
- Unaffected by outliers
- Discontinuous response
  - Small change in  $x$  could lead to a large change in  $y$
- Coarse-grained
  - A 17-node tree can only predict 17 different probabilities

## LOGIT

- Requires hand-built models
- Deletes records or imputes missing values
- Sensitive to outliers
- Continuous smooth response
  - Small change in  $x$  leads to a small change in  $y$
- Can have unique predicted probability for every record

# *Comparison of Hybrid models - Earlier methods*



- **Logit run in deliberately shallow tree**
- **Run in terminal nodes**
- **Results not successful**
- **Want to understand why**

# *How CART works*

- **CART excels in the detection of local data structure**
- **Once a database is partitioned into two sub-samples at the root node, each half of the tree is analyzed entirely separately**
- **As the partitioning continues, the analysis is always restricted to the node in focus**
  - **the discovery of patterns becomes progressively more local**
  - **information from different nodes is not pooled or combined**
  - **the “fit” at one node is never adjusted to take into account the fit at another node**

# *CART trees progressively truncate the variance of $x$ and $y$*

- **Goal of CART is to split data into homogeneous subsets**
  - the farther down the tree we go the less variability in the dependent variable
- **CART splits send cases with  $x \leq c$  to left and  $x > c$  to right**
  - Thus, the variance in predictor variables is also drastically reduced
  - if  $X$  is normally distributed and the cut point is at mean, variance in child nodes is reduced by about 64%; for subsequent mean splits the variance reduction will always be greater than 50%
  - thus, two splits on a variable could reduce variance to less than 25% of full sample variance
    - ♦ reduction will also apply to other correlated predictors

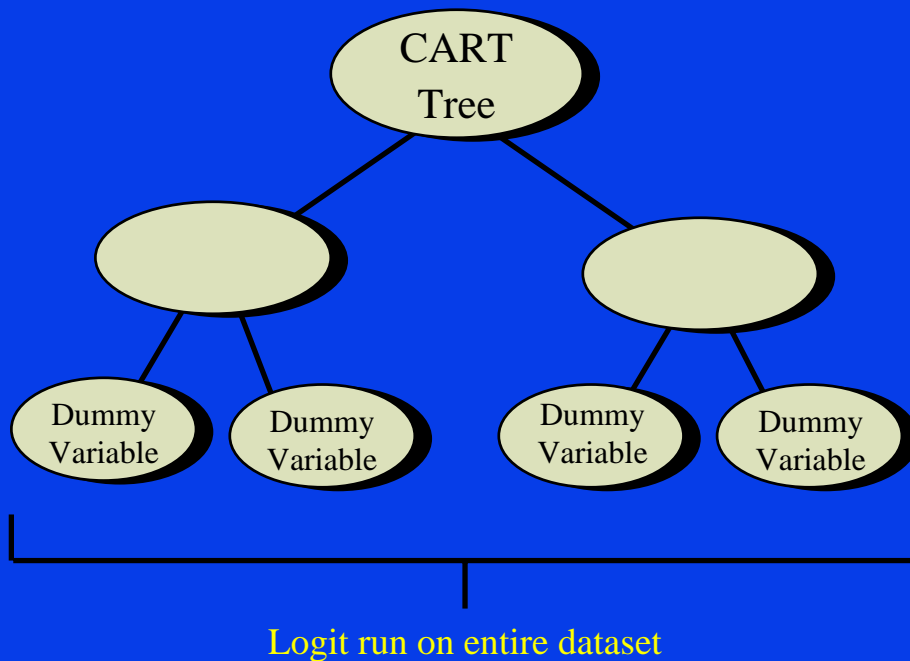
# *Running Logit models in CART terminal nodes is hopeless*

- **By the time CART has declared a node terminal the information remaining in the node is insufficient to support further statistical analysis**
- **The sample size is drastically reduced**
  - **The x and y variables have substantially reduced variance**
- **In a well-developed CART tree no parametric model should be supportable within terminal nodes**
- **Some analysts have suggested estimating logits or other parametric models earlier in the tree, after just a few splits**
  - **Same drawbacks as terminal node models but less extreme**
  - **At best provides a mechanism for finding switching regressions-- not terribly successful in practice**

# *Key to successful hybrid: run Logit in the root node*

- **Need to run the logit on all the data, thereby capitalizing on logit's strength in detecting global structure**
- **Implemented as follows**
  - **Run CART, assign every case to a terminal node**
    - Assignment possible even for cases with many missing values
    - Even a case with all missing data can be assigned a terminal node
  - **Terminal node assignment reported by categorical variable with as many levels as terminal nodes**
  - **Feed this categorical variable in the form of terminal node dummies to a logit model**

# Hybrid CART-Logit



- **Single Logit run**
  - 1) Uses all data available
  - 2) Has variation in predictor variables
  - 3) Dummy variables for terminal nodes represent CART tree
  - 4) added variables constitute the hybrid model

# Logit formulas for CART and Hybrid models

- CART only**

$$y = \beta_0 + \beta_1 \text{NODE}_1 + \beta_2 \text{NODE}_2 + \dots + \beta_K \text{NODE}_K$$

Where  $\text{NODE}_i$  is a dummy variable for the  $i^{\text{th}}$  CART node.

**NOTE:** Every observation is assigned to a CART terminal node regardless of whether any predictor variables are missing.

- CART-Logit Hybrid**

$$y = \beta_0 + \beta_1 \text{NODE}_1 + \beta_2 \text{NODE}_2 + \dots + \beta_R \text{NODE}_R + \beta_{R+1} X_1 + \beta_{R+2} X_2 + \beta_{R+3} X_3 + \dots + \beta_R X_R$$

$$= \beta_0 + \sum_{i=1}^R \beta_i Q_i + \sum_{R+1}^R \beta_i Z_j$$

CART node dummies    Hybrid covariates

# *Logit on terminal node dummies alone reproduces CART exactly*

- **The logit model fit to CART terminal node dummies converts the dummies into estimated probabilities**
- **Otherwise, it is an exact representation of the CART model**
- **Each dummy represents the rules and interaction structure discovered by CART, albeit buried in a black box**
- **Likelihood score for this model forms a baseline for further testing and model assessment**
- **Excellent way to incorporate sampling weights and recalibrate a CART tree**

*Hybrid allows baseline Logit to expand by adding other variables to model*

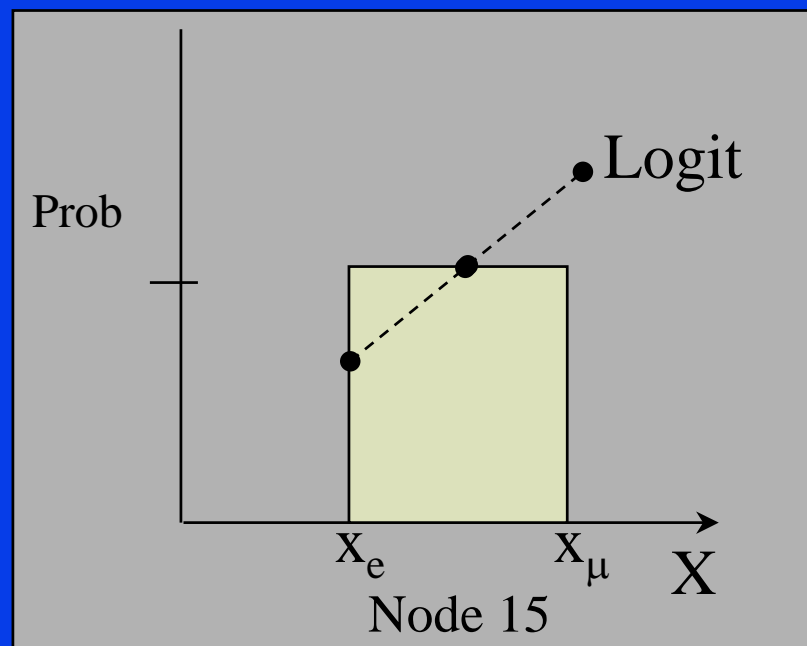
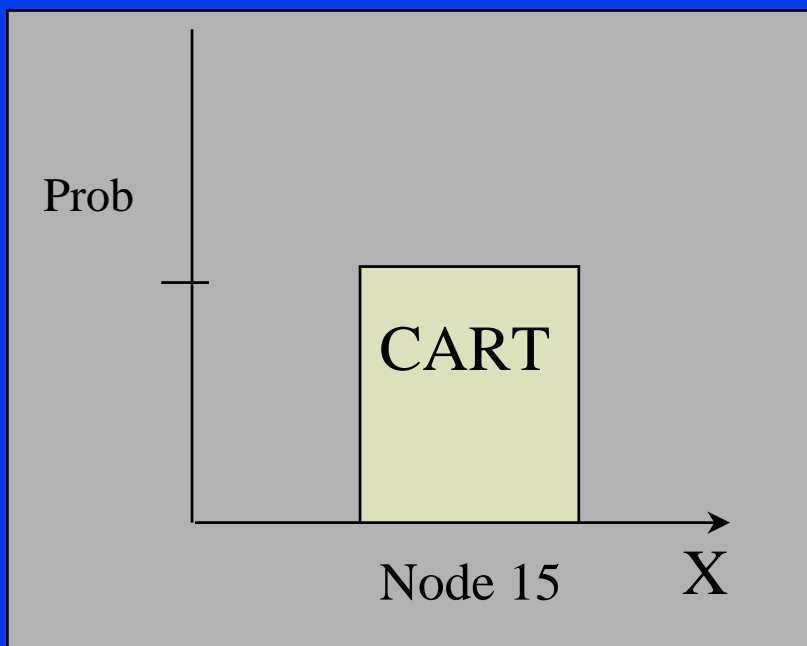
- **Variables added as main effects will capture effects common across all nodes**
- **Common effects are undetectable within terminal nodes because the signal to noise ratio within terminal nodes is too low**
- **Effects detected across terminal nodes are likely to be weak -- all strong effects already detected by CART**
- **Nevertheless, a collection of weak effects can be very significant**
- **Added variables can be tested as a group via likelihood ratio test**

# *Why Does the Logit Augment CART?*

- **By looking across nodes, Logit finds effects that CART cannot detect**
- **Because these effects are not terribly strong they are not picked up by CART as primary node splitters**
- **Once the sample is split by CART, these effects become progressively more difficult to detect as the subsamples become increasingly homogenous in the CART child nodes**
- **While these effects may not be the strongest individually, collectively they can add enormous predictive power to the model**
- **Results in a major enhancement in CART methodology**

# *How the Logit improves CART*

- CART assigns a single score (probability) to all cases arriving at a terminal node
- Logit imposes a slope on the cases in the node, allowing continuous differentiation of within-node probabilities based on variables
- Note: Logit is common to all nodes - so slope is common across nodes



# *LOGIT can also compensate for CART weaknesses*

- **CART sometimes produces a very coarse grained response image**
  - Might produce only a small number of terminal nodes
  - “Score” is shared by all cases in a node
  - A 12 terminal node tree produces only 12 distinct scores
- **Once CART finds a relatively rich group of responders it might stop splitting that node**
  - Identifying a large block of cases as responders could be a classification tree success
  - For targeting, we may want to distinguish the strongest responders from the merely strong responders (e.g., for targeting with a restricted budget, we might only mail to the strongest group)
  - For classification purposes, such a distinction is irrelevant
    - ♦ Both groups are responders so CART does not try to distinguish further

# *Building the Logit component of the hybrid*

- **Add variables already selected as important by CART**
- **Add competitor variables in the root node that never appeared as splitters or surrogates in the tree**
- **Possibly use a stepwise selection procedure**
- **Add variables known to be important from other studies**
- **Will need to consider variable transforms--log, square root**
- **Will need to deal with missing values**
  - **impute a value**
  - **zero and add missing value dummy indicator to model**
- **Can ignore interactions: already captured in CART terminal nodes**
  - **goal is to search for weak main effects**

# *Missing value handling in hybrid CART-Logit*

- **Simplest approach--ignore the problem!**
  - Drop all records with missing values on model variables
- **Assign CART-predicted probabilities to those cases**
- **Assign hybrid-predicted probabilities to all other cases**
- **More complicated approaches**
  - Missing value imputation
  - Dummy for missing value indicator plus nesting for non-missing
- **Complicated procedures not needed since CART tree will give good results anyway**

## *Beyond linear main effects: node-specific effects*

- **The only interactions worth considering in hybrid model are:**
  - terminal node dummy interactions with selected variables
  - interactions with missing value indicators
- **Has the effect of allowing a separate model for a group of terminal nodes, most likely restricted to just a few variables**
- **Some difficulty in locating such interactions**
  - With  $T$  terminal nodes there typically will be  $2^{T-1} - 1$  partitions of nodes into subsets
- **If these are pursued an automated search is needed**

# *Does a main effects Logit augmented CART impose too much structure on the sample?*

- **The CART methodology segments the data in very different sub-samples**
- ***Q:* Why should we expect that a single common Logit would be valid?**
- ***A:* First, the terminal node dummies capture all the complex interactions and non-commonality of the hyper-segments**
- **Second, we test the Logit developed in each node to see if it makes an improvement over the CART score**
  - **If the Logit does not improve the node likelihood (a rare event), we do not apply the Logit to that node.**
  - **In our AMEX models, the nodes have often been improved dramatically by the Logit model**

# Assessment of node-specific logit fit

Node ID	_TYPE_	_FREQ_	CART log-likelihood	Logit log-likelihood	Expected log-likelihood	Logit LL minus CART LL	Expected LL minus CART LL
.	0	107315	-7533.91	-7490.01	-7490.01	43.9041	43.9038
1	1	55727	-3105.76	-3080.93	-3086.75	24.8256	19.0118
2	1	7477	-438.04	-436.13	-437.11	1.9098	0.9347
3	1	24946	-2066.42	-2057.35	-2058.26	9.0711	8.1642
4	1	2221	-94.87	-94.68	-94.25	0.1950	0.6259
5	1	820	-40.31	-41.48	-40.07	-1.1644	0.2428
6	1	16124	-1788.50	-1779.43	-1773.58	9.0670	14.9244

- **Important to test whether the hybrid model shows lack of fit in any node or subset of nodes**
- **Simple likelihood test node by node**
  - **CART model assigns the mean probability to all cases in a node**
  - **Hybrid assigns record-specific probability**
  - **If CART likelihood is greater than hybrid main effects model, do not apply hybrid model to this node**

# *Monte Carlo model assessment*

- **Assume known true probabilistic data generation processes**
- **Draw samples of various sizes (N=2,000 and 20,000)**
- **Repeat experiment many times (we use R=100)**
- **Examine typical outcomes (mean) rather than single example**
- **Assess models on basis of**
  - **Fit (log-likelihood)**
  - **Lift (gains chart)**
- **Look at both training data and holdout samples**

# *Specific Monte Carlo experiments*

- **True DGP is**
  - **Simple Logit- one variable**
  - **CART process- one variable, highly non-linear**
  - **Hybrid process**
  - **Logit- several variables (Possibly missing)**
  - **Hybrid- several variables (Possibly missing)**
  - **Highly non-linear smooth function (not Logit)**
  - **Complex Logit with information missingness**

# *Model Development Sequence*

- **Generate TEST and LEARN samples of equal size.**
  - **N=1,000 for LEARN and TEST small experiments**
  - **N=10,000 for LEARN and TEST for large experiments**
- **Generate HOLDOUT sample (N=2,000 or N=20,000)**
- **Run CART on LEARN data; select optimal tree using TEST data.**
- **Run Logit on all available data (pooled TEST and LEARN samples).**
  - **Include dummies for CART nodes.**
  - **Include selected variables and transforms**
  - **Choose missing value handling procedure.**

# *Summary of results*

- **In smaller samples, (N=2,000) Logit performs very well even when Logit is not the true model**
  - **Simpler model reduces risk of over-fitting**
- **In larger samples, (N=20,000) Hybrid model dominates out-of-sample performance**
  - **Even when Logit is true model, Hybrid is almost as good**
- **In larger samples, CART and Hybrid manage problems with missing values quite well whereas logit performance collapses**
- **In large datasets with high frequencies of missings, Hybrid outperforms other models regardless of which model is true**

# *Log-likelihood results N=1,000*

Learn/Test Sample				
Experiment	CART	Logit	Hybrid	Truth
1	-1231.97	-1200.34	-1192.39	-1201.38
2	-1260	-1385.33	-1259.37	-1268.39
3	-1188.5	-1163.94	-1144.98	-1160.99
4	-1198.47	-1158.21	-1125.64	-1080.05
4 w/dummies	-1191.24	-1158.21	-1120.5	-1080.05
5	-1179.37	-1153.71	-1111.39	-1032.93
5 w/dummies	-1175.81	-1153.71	-1108.3	-1032.93
6	-1265.4	-1374.76	-1259.12	-1227.14
6 w/dummies	-1266.13	-1374.76	-1260.27	-1227.14
7	-1192.58	-1142.43	-1108.89	-1036.11
7 w/dummies	-1175.85	-1142.43	-1101.3	-1036.11

Holdout Sample				
Experiment	CART	Logit	Hybrid	Truth
1	-1263.49	-1197.91	-1214.28	-1196.95
2	-1316.8	-1387.31	-1316.72	-1266.95
3	-1261.25	-1164.93	-1204.03	-1158.63
4	-1311.01	-1170.84	-1224.17	-1079.54
4 w/dummies	-1321.32	-1170.84	-1234.33	-1079.54
5	-1271.29	-1161.18	-1192.12	-1030.16
5 w/dummies	-1276.11	-1161.18	-1197.16	-1030.16
6	-1576.29	-1384.05	-1578.55	-1227.67
6 w/dummies	-1546.62	-1384.05	-1549.09	-1227.67
7	-1283.42	-1157.93	-1191.85	-1037.49
7 w/dummies	-1228.37	-1157.93	-1198.81	-1037.49

# *Log-likelihood results N=10,000*

Learn/Test Sample				
Experiment	CART	Logit	Hybrid	Truth
1	-12384.8	-11980	-11972.2	-11981
2	-12671.4	-13861.8	-12670.9	-12658.7
3	-11931.5	-11589.3	-11563.5	-11543.4
4	-12025.6	-11633.4	-11497.4	-10789.1
4 w/dummies	-11979.6	-11633.4	-11484.5	-10789.1
5	-11822.1	-11548.4	-11274.8	-10277.8
5 w/dummies	-11728	-11548.4	-11230.8	-10277.8
6	-13127.9	-13788.5	-13085.9	-12267.8
6 w/dummies	-13069.2	-13788.7	-13046.3	-12268
7	-11868.7	-11507.5	-11283.5	-10374.7
7 w/dummies	-11625.4	-11507.5	-11217	-10374.7

Holdout Sample				
Experiment	CART	Logit	Hybrid	Truth
1	-12431.4	-11999.2	-12010.6	-11997.9
2	-12710.1	-13864.2	-12710.5	-12659.9
3	-12003.1	-11585.8	-11594.8	-11537.6
4	-12259.2	-11645.1	-11648.5	-10794.8
4 w/dummies	-12238.8	-11645.1	-11654.1	-10794.8
5	-12028.4	-11547	-11417.5	-10280.6
5 w/dummies	-11970.3	-11547	-11401.1	-10280.6
6	-13388.9	-13796.2	-13357.4	-12251.5
6 w/dummies	-13368.6	-13796.5	-13350.3	-12251.8
7	-12114.7	-11505.6	-11461.6	-10370.6
7 w/dummies	-11792.5	-11505.6	-11400	-10370.6

# *Performance on the holdout sample 2000 observations, 100 replications*

Experiment	CART	Logit	Hybrid	Truth
1	0.1857	0.2406	0.237	0.2406
2	0.1606	0	0.1615	0.1809
3	0.2091	0.2584	0.251	0.2597
4	0.1879	0.256	0.2474	0.3013
4 w/dummies	0.1867	same	0.2465	same
5	0.1995	0.2498	0.2497	0.3092
5 w/dummies	0.1989	same	0.2496	same
6	0.0665	0.0198	0.0707	0.2061
6 w/dummies	0.0656	same	0.0696	same
7	0.1799	0.2441	0.2394	0.2966
7 w/dummies	0.1977	same	0.2399	same

**Table 5**

- **Actuals gains**

Experiment	CART	Logit	Hybrid	Truth
1	0.1853	0.2393	0.2358	0.2393
2	0.1609	0.0003	0.1621	0.1815
3	0.2116	0.2613	0.254	0.2627
4	0.189	0.2566	0.2483	0.3017
4 w/dummies	0.1873	same	0.247	same
5	0.1993	0.251	0.2505	0.3112
5 w/dummies	0.1989	same	0.2504	same
6	0.0664	0.0205	0.0706	0.2068
6 w/dummies	0.0658	same	0.0693	same
7	0.1814	0.2452	0.2408	0.2968
7 w/dummies	0.1957	same	0.2414	same

**Table 6**

- **Expected gains**

# Performance on the holdout sample 20,000 observations, 100 replications

Experiment	CART	Logit	Hybrid	Truth
1	0.198	0.239	0.2386	0.239
2	0.1791	0.0003	0.1791	0.1819
3	0.2357	0.2619	0.2612	0.263
4	0.2241	0.2592	0.2601	0.302
4 w/dummies	0.226	same	0.26	same
5	0.2282	0.2536	0.2621	0.3114
5 w/dummies	0.232	same	0.2635	same
6	0.1114	0.0285	0.1209	0.2069
6 w/dummies	0.1167	same	0.1217	same
7	0.2152	0.247	0.251	0.2966
7 w/dummies	0.2323	same	0.2541	same

- **Table 12**
  - **Actual gains**

Experiment	CART	Logit	Hybrid	Truth
1	0.1987	0.2395	0.2391	0.2395
2	0.1793	0.0004	0.1794	0.1818
3	0.2354	0.2614	0.2608	0.2626
4	0.2243	0.2589	0.2599	0.3019
4 w/dummies	0.2265	same	0.2601	same
5	0.2282	0.2533	0.2622	0.3113
5 w/dummies	0.2321	same	0.2635	same
6	0.1118	0.0296	0.1213	0.2068
6 w/dummies	0.1171	same	0.1223	same
7	0.2159	0.2471	0.2514	0.2968
7 w/dummies	0.2316	same	0.2544	same

- **Table 13**
  - **Expected gains**

# *Performance as measured by gains of CART, Logit, and Hybrid models*

- **The relative gains pattern roughly matches the LL pattern (see tables 5 and 12)**
  - **The Hybrid seems to do slightly better in terms of gains than it does in terms of LL**
  - **Even when the Hybrid is not the best model, the Hybrid is never much worse than the best model**
  - **In some cases the Hybrid is much better than the other models**
  - **The larger the sample the better the hybrid performs relative to the other alternatives. (Compare table 5 with 2000 observations to table 12 with 20000 observations)**

# *Measurement of performance without a holdout sample*

- **In the Monte Carlo situation the true probabilities are known; thus, the expected performance can be computed**
  - **The expected performance on the LEARN/TEST sample or HOLDOUT sample and the actual performance on the HOLDOUT sample are all identical to within the random sampling error**
  - **Using the estimated probabilities to estimate the expected performance overestimates the true performance substantially for in-sample data and in some cases for out-of-sample data as well**

# *Performance measurement*

- **The actual integral gains computed on the TEST sample alone is a very good estimate of the out-of-sample performance**
- **The column marked "Truth" gives the best possible average out-of-sample performance. "Truth" gives the gains integral based on true probabilities. In some cases the models achieve almost the maximum possible gains. In others they achieve much less**
- **In-sample estimated models may do better than the "Truth" due to over fitting**
- **Overfitting means the performance will be worse out-of-sample**
  - ◆ **A very overfit model will perform much worse out-of-sample**
  - ◆ **In larger samples overfitting is reduced (compare Tables 1 and 8)**
- **NOTE: "Truth" has not been adjusted for missingness. No model could ever achieve the gains listed under the truth column for the cases with missing data (Experiments 4 through 7)**

# *Performance in the presence of Missing Values*

- **Including dummies for missing values in the CART improves CART's performance in large samples (see Table 12) and whenever the missingness is informative (see Experiment 7, Table 5)**
  - **In most real world situations individuals with missing data do differ from other individuals. Missingness is informative. This should not be surprising, considering the usual reasons for missing data**
    - ◆ **Refusal to answer questions**
    - ◆ **Failure to match data from other sources**
    - ◆ **Nonexistent data (credit bureau, driver's license records, census, etc.)**
    - ◆ **Inconsistent or impossible combinations**

# *Performance in the presence of Missing Values*

- **Including dummies for missing values in CART improves the Hybrid's performance in large samples (see Table 12) and whenever the missingness is informative (see Experiment 7, Table 5)**
  - **The improvement is much smaller. The Hybrid works quite well even if based on a mediocre CART model. (Examine the results for Experiment 7, as reported in Tables 5, 6, 12 and 13)**
  - **In large samples with missingness the Hybrid always outperforms Logit and CART. Logit, however, does surprisingly well in situations with missingness if the linear logistic model is a reasonable approximation to the DGP**

# *References*

- **Breiman, L., J. Friedman, and R. Olshen, and C. Stone (1994), *Classification and Regression Trees*, Pacific Grove: Wadsworth.**
- **Friedman, J. H. (1991a), *Multivariate Adaptive Regression Splines (with discussion)*, *Annals of Statistics*, 19, 1-141 (March).**
- **Michie, D., D. J. Spiegelhalter, and C. C. Taylor, eds (1994), *Machine Learning, Neural and Statistical Classification*, London: Ellis Horwood Ltd.**
- **Steinberg, D. and Colla, P. L., (1995), *CART: Tree-Structured Nonparametric Data Analysis*, San Diego, CA: Salford Systems.**