SOFTWARE REVIEW

CART 4.0

CART 4.0 is a decision tree software distributed by Salford Systems, Inc., a data mining software and consultancy firm (Salford Systems, 2000). Figure 1 shows a small decision tree created by CART from some simulated data. This decision tree seeks to predict job status 6 months after hire based on variables known at the time of hire. The three categories of job status are fired, quit, and still on the job. The variables used to predict job status are conscientiousness test scores, job classification (sales or customer service), and age.

CART has a strong history of use in financial and marketing research. Most of the applications in finance, such as the selection of stocks, result in proprietary reports that are not publicly available. Likewise, the CART applications in marketing are also usually proprietary. However, there are many published applications of CART, ranging from language development (Delaney-Black et al., 2000), memory recovery from traumatic brain injury (Stuss et al., 2000), to relapse in schizophrenia (Doering et al., 1998). CART has also been used to conduct research in psychology on topics ranging from identifying risk for functional impairment (Lemsky, Smith, Malec, & Ivnik, 1996) to examining the relationship between job stress and mental health (Brook & Brook, 1995). In addition, CART has been used in management research to examine major league baseball salaries (Hoaglin & Velleman, 1995) and the prediction of business failures (Dimitras, Zanakis, & Zopounidis, 1996).

In the realm of decision tree science, CART is known as a binary recursive partitioning algorithm. It is a partitioning algorithm because it seeks to partition a set of observations into subgroups. In the tree in Figure 1, CART has partitioned the data into 4 groups called terminal nodes. The terminal nodes are the nodes that are not further subdivided. CART is a binary algorithm because it splits each parent node into exactly two child nodes. Thus, at the top of the tree, there is a parent node containing all 384 observations in the data set. If CART splits a node, it always splits the node into two child nodes. In Figure 1, CART has split the parent node containing 384 observations using the conscientiousness test score. Those who have a conscientiousness score of 4 or lower are sent to the left node, and the remaining observations are sent to the right node. CART is a recursive algorithm because once it creates a child node, it then treats the child node as a parent node and attempts to split it. Thus, after the initial split created a child node containing all cases with a conscientiousness score of 4 or lower, CART treated that child node as a parent node and sought (successfully) to split the node again (this time on whether or not the job classification was a call center).

One of the biggest problems in building decision trees is capitalization on chance. It is possible and likely that one can build a tree that will perfectly (or almost perfectly) describe a set of data. Such trees are very likely to have very poor predictive value because they find patterns that only fit the current data but do not generalize to new

Organizational Research Methods, Vol. 5 No. 1, January 2002 129-132

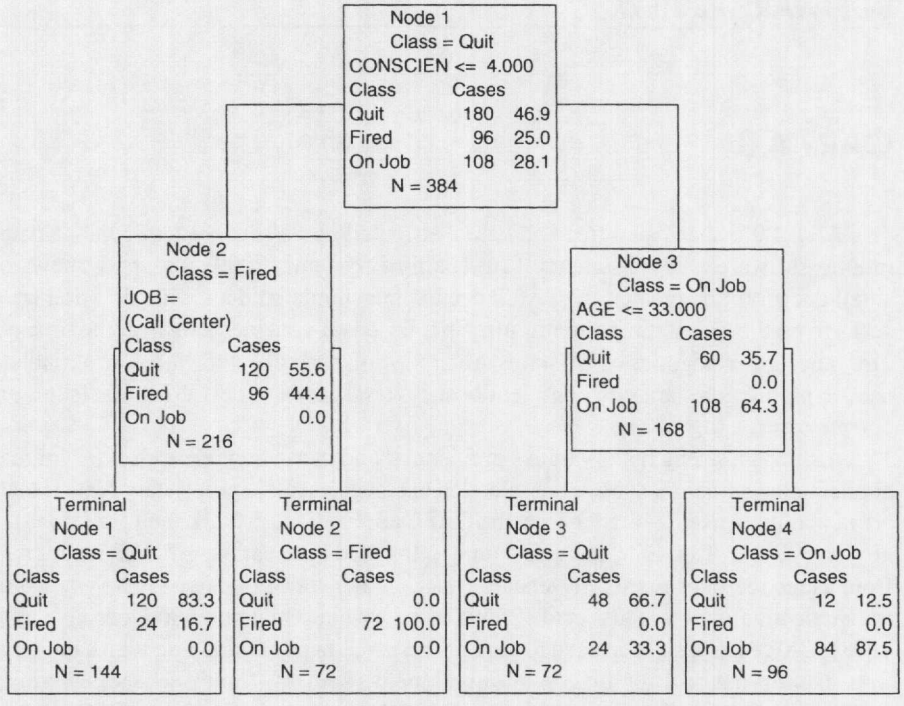

Figure 1: Decision Tree to Predict Job Status 6 Months After Hire as a Function of Job Classification, Conscientiousness, and Age

data. By way of metaphor, a multiple regression equation may show good predictive ability on the sample from which it was created but far worse predictive validity in a new sample. In the decision tree sciences, such trees are referred to as *overfitted* trees. To guard against overfit models, CART incorporates cross-validation into its tree derivation algorithm. If one has large amounts of data, CART builds a tree on part of the data and uses the other part of the data for cross-validation. If one has smaller data sets, CART engages in what it calls tenfold cross validation. In this process, available data are randomly assigned to 1 of 10 groups. In each cross-validation replication, 9 of the groups data are used to build the tree, and one subset is used to test the tree. This process is repeated 10 times. CART's approach to cross-validation is one of its major strengths.

Most decision tree algorithms create very large trees and then prune them back in search of the optimal tree. Oates and Jensen (1998) have shown that for most recursive partitioning tree algorithms, the size of the final tree is linearly related to the size of the data set. Specifically, trees continue to grow in complexity even when smaller tress are equally or more accurate. Oates and Jensen have shown that this problem is due to characteristics of many of the pruning algorithms. They show that the pruning algorithm used in CART 4.0 is superior to most in use.

In Figure 1, the variables conscientiousness, age, and job classification are known as splitters because they are used by the tree in splitting the data into partitions. Problems can arise when the splitter variable is missing for an observation. Many decision

tree programs require that the user either interpolate missing data before submitting the data to the program or drop cases that include missing data. When CART encounters missing data on a variable, it uses a surrogate splitter to assign the observation to one child node or the other. A surrogate splitter is one that closely mimics the behavior of the splitter. For example, in Figure 1, when an observation is missing age, the algorithm splits the data based on years of experience. In the data set used for Figure 1, years of experience is a reasonable surrogate for age, and splitting on years of experience less than or equal to 13 provides about the same split as age 33 or younger.

Some misclassification errors are more serious than others. For example, consider the prediction of police officer corruption. From an organization's perspective, it is very bad to misclassify a job applicant who will be corrupt as a police officer as one who will be noncorrupt. It is less bad to classify a person who is noncorrupt as a person who is corrupt because there are many more applicants for police jobs than people who are actually hired. (We recognize that the perspective of the falsely rejected job applicant is different from the perspective of the organization). CART permits one to identify some misclassification errors as more costly than other errors. Many other decision tree programs cannot differentially weight classification errors.

Our issues with using CART are few. Our first issue concerns data complexity. The first author of this review primarily conducts research in personnel selection where most relationships are linear. In such applications, CART produces poor decision trees. Linear regression produces much more reasonable models of the data. CART excels in data sets of high complexity. For organizational behavior/human resources researchers, high complexity refers to data sets with interaction effects and nonlinear relationships; the more obscure the better. This is not so much a criticism as a boundary condition on when to and when not to use CART.

A second issue concerns sample size. Is size important? Many applications of CART are in disciplines such as finance or marketing, where there are hundreds of thousands (or millions) of cases of data. For example, a common application is to determine which million or so people should receive preapproved credit card applications. The credit card company can make money from several distinct market segments. For example, those who pay off the balance monthly but who use the card extensively generate profits through the fees charged to merchants. Those with more modest purchases who do not pay off their balance in full monthly but who pay the minimum payment every month generate profits through paid interest. Many potential users of CART are falsely discouraged from using the software because their idea of a large sample size is 500 not 500,000. We are aware of applications with a small sample size by CART standards ($n = 500$) where the resulting decision tree was very poor. We are also aware of applications with sample sizes of a few hundred where very useful trees have been developed. Thus, although we would believe, in general, that CART is not very useful for small data sets, analyses involving a few hundred cases may often yield very informative results.

In general, we are pleased with CART. It is a market leader in the decision tree business. Its tree development algorithm is well known and well studied. CART 4.0 has overcome most previous problems associated with the software. The manual is very useful, and the program defaults get novice users up and running quickly. The software is quick, and the output is easy to understand. If one has a categorical dependent variable, several hundred observations, and a topic area with few linear relationships, we recommend exploring CART 4.0.

References

Brook, R. J., & Brook, J. A. (1995). Sequential tree method of examining the relationship between job stress and mental health. *Perceptual and Motor Skills, 80*, 287-290.

Delaney-Black, V., Covington, C., Templin, T., Kershaw, T., Nordstrom-Klee, B., Ager, J., et al. (2000). Expressive language development of children exposed to cocaine prenatally: Literature review and report of a prospective cohort study. *Journal of Communication Disorders, 33*, 463-481.

Dimitras, A. I., Zanakis, S. H., & Zopounidis, C. (1996). A survey of business failures with an emphasis on prediction methods and industrial applications. *European Journal of Operational Research, 90*, 487-513.

Doering, S., Muller, E., Kopcke, W., Pietzcker, A., Gaebel, W., Linden, M., et al. (1998). Predictors of relapse and rehospitalization in schizophrenia and schizoaffective disorder. *Schizophrenia Bulletin, 24*, 87-98.

Hoaglin, D. C., & Velleman, P. F. (1995). A critical look at some analyses of major league baseball salaries. *The American Statistician, 49*, 277-285.

Lemsky, C. M., Smith, G., Malec, J. F., & Ivnik, R. J. (1996). Identifying risk for functional impairment using cognitive measures: An application of cart modeling. *Neuropsychology, 10*, 368-375.

Oates, T., & Jensen, D. (1998). Large datasets lead to overly complex models: An explanation and a solution. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (pp. 294-298).

Salford Systems. (2000). *Cart for windows user's guide*. San Diego, CA: Author.

Stuss, D. T., Binns, M. A., Carruth, F. G., Levine, B., Brandys, C. F., Moulton, R. J., et al. (2000). Prediction of recovery of continuous memory after traumatic brain injury. *Neurology, 54*, 1337-1344.

Michael A. McDaniel and W. Lee Grubb, III
Virginia Commonwealth University