iberi

Institute of
Business and
Economic Research

University of
California at
Berkeley

CENTER FOR REAL ESTATE AND URBAN ECONOMICS

WORKING PAPER SERIES

WORKING PAPER NO. 93-213

CLASSIFICATION AND REGRESSION TREE ANALYSIS IN MARKETING RESEARCH

By

JAMES M. CARMAN
RAVI DHAR

These papers are preliminary in nature: their purpose is to stimulate discussion and comment. Therefore, they are not to be cited or quoted in any publication without the express permission of the author.

WALTER A. HAAS SCHOOL OF BUSINESS

CENTER FOR REAL ESTATE AND URBAN ECONOMICS
UNIVERSITY OF CALIFORNIA AT BERKELEY
Kenneth Rosen, Chair

The Center was established in 1950 to examine in depth a series of major changes and issues involving urban land and real estate markets. The Center is supported by both private contributions from industry sources and by appropriations allocated from the Real Estate Education and Research Fund of the State of California.

INSTITUTE OF BUSINESS AND ECONOMIC RESEARCH
Richard Sutch, Director

The Institute of Business and Economic Research is an organized research unit of the University of California at Berkeley. It exists to promote research in business and economics by University faculty. These working papers are issued to disseminate research results to other scholars. The authors welcome comments; inquiries may be directed to the author in care of the Center.

CLASSIFICATION AND REGRESSION TREE ANALYSIS

IN MARKETING RESEARCH

James M. Carman

Walter A. Haas School of Business

University of California, Berkeley
94720
(510) 642-1502
and

Ravi Dhar

School of Organization and Management

Yale University

Working Paper No. 93-213

November 1992

This paper grew out of a project generously funded by the Center for Real Estate and Urban Economics at Berkeley and conducted by the authors along with Professor Frederick E. Balderston of the Haas School of Business.

CLASSIFICATION AND REGRESSION TREE ANALYSIS IN MARKETING RESEARCH

ABSTRACT

This paper provides an evaluation of the Classification and Regression Tree (CART) model for marketing research applications. The paper reviews tree type multivariate classification and analysis of variance models from nonparametric classification models to logistic regression. The CART model and its computer software version are described and found superior to previous tree analysis procedures. An application and test of the model are then presented using a segmentation application with a large sample of mixed continuous and categorical variables. The model's results, robustness, and ease of use are compared with multinomial logistic regression. The CART model is less restrictive than the latter and produced more useful results. It is recommended for both academic and commercial work.

Seeking understanding of the interrelationships among variables within data with mixed levels of measurement is a common problem in marketing research in general and segmentation research in particular. For example, segmentation studies usually involve economic, demographic, usage rate, and attitudinal variables. The measures of these variables are going to be a mixture of ratio scales, rating scales, ranking scales, and nominal classifications. While progress in the development of statistical procedures has moved somewhat slowly, the rapid increase in computing power has made it possible to develop search procedures that increase our understanding of the relationships among variable with mixed measurement levels. This genre of search procedures attempts to find relationships among variables with minimum assumptions about the level of measurement or the nature of distributions as are common in parametric statistical techniques. Tree models are a family of such algorithms that have proven useful to marketers and continue to be improved.

The purpose of this paper is to provide an evaluation of a new tree model procedure, Classification and Regression Trees, (CART) developed by Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone (1984). Perhaps the most unique aspect of this procedure is that it is both a classification and regression tool. That is, it will accommodate nominal or interval data in any variable, be it dependent or independent. It employs decision rules that are "less parametric" than some other techniques, such as Analysis of Variance, while at the same time utilizing least squares or least absolute deviation criteria in the construction of regression trees.

The paper begins with a brief review of tree models. The CART model and computer program are then described. An application of the model in a search for market segments is presented next and the CART procedure is compared to other multivariate approaches. The CART results are then validated by comparison with multinomial logistic regression and by repeating the procedure on another data set. The paper concludes with some summary comments concerning the usefulness of tree analysis in marketing classification problems.

Tree Analysis in Marketing Research

Marketing research in general and segmentation research in specific are characterized by data at varying level of measurement. Intervally scaled data are usually the exception rather than the rule. Marketing research is making increasing use of measures, such as rating scales, that are often treated as intervally scaled data, but are, in fact, ordinal or nominal.

The problems of analyzing data with such mixed measurement has long been a frustration to researchers. During the 1960s when it became obvious that the R^2s one usually found in segmentation studies using regression were in the range from .1 to .3, there was a search for analytical approaches that would find "what must be stronger relationships in the data." Probably today researchers are more willing to accept the fact that measurement and noise are great enough that not all the blame should be laid at the feet of the analytical model.

The parametric multivariate methods used for mixed interval and nominal data are well known. If the dependent variable is intervally scaled and normally distributed, multiple classification analysis (dummy regression) can be used. If the dependent variable is nominal and the predictors are all intervally scaled, multiple discriminant analysis may be robust. However, the multiple discriminant model assumes all predictors are multivariate normal. These conditions are seldom met with marketing research data. A logistic transformation of even a polytomous categorical dependent variable combined with modern maximum likelihood estimation procedures has permitted a significant improvement in the efficiency of parametric regression estimates for this problem. Nonetheless, traditional multivariate methods of classification place restrictions on the data types in the analysis and on the homogeneity of the structures of the data over the measurement space.

Newer, nonparametric methods that do not place restrictions on the covariance matrices, such as the "kth nearest neighbor rule," employ Bayes Rule to search for structure (Hand, 1981). Unfortunately, they: are very sensitive to the selection of an initial starting value of central tendency; do not handle categorical variables and missing data; do not provide much insight into the structure of the data, which is, after all, the purpose of most segmentation research.

Nonetheless, models with fewer restrictions and assumptions remain of great importance for marketing research. One would like to use the most powerful analysis methods available, but the data often do not fit the assumptions of the analysis model. Classification analysis seems to give away too much, while regression analysis is too restrictive. As a result, tree models with large samples, that offer some advantages of both, have been widely discussed in the marketing literature, and a complete review here is unnecessary (Carman, 1970; Assael, 1970; Armstrong and Andress, 1970; Kinnear and Taylor, 1971; Currim, Meyer and Le, 1988). Before describing the new CART system, to aid in comparison, four tree techniques used in marketing will be briefly reviewed: Automatic Interaction Detector, Classification Trees, Logistic Regression Trees, the Concept Learning System.

Automatic Interaction Detector (AID)

The oldest of the widely available tree algorithms, Automatic Interaction Detector (AID), is often ideally suited for use in segmentation research (Sonquist, 1970). Its name derives from the fact that it will find interactions among the predictors that in a conventional ANOVA model, multiple classification analysis, or dummy regression would have to be hypothesized and tested for in an ad hoc procedure.

AID is fundamentally a classical parametric multivariate procedure in that it builds trees by repeated application of one-way ANOVA to make every branch. Thus, an intervally scaled dependent variable and all nominally scaled predictors are assumed. Further, the splitting criterion is based on the ratio of variance between a possible split to the total variance of the dependent variable. The program, as do all tree programs, proceeds sequentially through the data seeking the binary split that explains the greatest proportion of variance in the dependent variable. While not providing a solution to the problem of missing data, it provides some methods for dealing with the problem as well as with multicolinearity problems.

All tree techniques "use up" the sample as it proceeds through the splitting process. In addition to splitting rules, tree procedures need rules for stopping the splitting process. Some procedures also have rules for pruning branches already split. AID does not prune and stops when a branch contains five observations. Because of the parametric assumptions and focus on variance of the dependent variable, AID is very much a large sample procedure.

Classification Trees

Often the dependent variable in segmentation research is not intervally scaled, like quantity of a brand purchased, but nominal, like a consumer's purchase incident. In such cases, if the restrictions of multiple discriminant analysis cannot be met, the problem is a classification problem of aggregate data for which higher-order contingency table analysis models have been developed (Dillon and Goldstein, 1977; Perreault and Gwin, 1978; Green, Carmone and Wachspress, 1979).

A general strategy for the construction of such algorithms has been to use a maximum likelihood or Bayes rule to grow an initial tree solution and then to run a subset validation procedure, such as the Jackknife technique (Jones, 1956; Mosteller and Tukey, 1967; Gray and Schucany, 1972; Fenwick, et al., 1983), to prune the tree back to one that can be validated through repeated subsample testing. The initial tree is grown as a sequence of binary splits. The branch construction is just another geometric representation of doing repeated cross tabulations of binary splits on the sample to find the most "important" split and then to look for binary splits on each of the resulting two subsample cells. The procedure then continues in a recursive manner until some stopping rule, such as minimum remaining sample size in a branch, is met. Such a procedure will discover any interactions among variables and should identify multicolinearity.

Clearly, modern computer technology is up to such repeated search procedures. Thus, the interesting questions concern: the criteria for growing the initial tree, the robustness of the pruning procedure, and the ease of using and interpreting the output from such computer programs.

Perreault and Barksdale (1980) developed a Chi-Square based automatic interaction detector model (CHAID) to grow classification trees. The dependent variable is polytomous (not necessarily ordered); the trees are constructed by calculating Chi-Square statistics for every possible split rather than calculating the ratio of between to total variance. The most interesting growth strategy aspect of this procedure is that it collapses categories before growing the initial tree. That is, the program reduces the number of classes for each predictor variable before considering which successive branch splits will best explain the differences in the dependent variable. As a result, it is not constrained to binary splits. After each split, the program goes back to see if it needs to split any categories of predictor variables previously combined. The strategy can be characterized as pruning before growing rather than growing an initial tree and then pruning. This topic is one to which we will return in the discussion of CART.

Logistic Regression Analysis

An important subset of segmentation research deals with consumer choice rather than market share or brand loyalty. Purchase incidence models represent the dependent variable as dichotomous indicating whether the brand has been purchased or not. The currently popular class of such models assumes the particular logistic functional form of relationship between the binary dependent variable and a linear combination of predictor variables that are either intervally scaled or dummy. Given the assumptions of this model, the parameters can be estimated using maximum likelihood search routines. The logit model becomes a tree in its nested form because it assumes buyers make choices through a series of ordered attribute criteria (Bucklin and Lattin, 1991).

It is important to realize that the logistic transformation of a polytomous dependent variable has powerful application in the analysis of aggregate data in contexts completely outside of choice modelling. In the 1960s it was common to use OLS to estimate regression coefficients with a dichotomous dependent variable. Such estimates are not efficient because the error term is either 0 or 1, quite heteroscedastic, and seriously violate the least squares assumptions. Such is not the case with the error term of the logistic transformed dependent variable (Maddala, 1983; Dhrymes, 1986). Modern logistic regression also gains the advantages of maximum likelihood estimation over OLS. Of course, the structure of the dependence must still be linear and specified by the user. Interactions must be hypothesized and modelled as in any regression model. It is also likely that the output may not be as insightful as with trees.

Concept Learning System (CLS)

Currim, Meyer and Le (1988) have suggested the CLS tree model as an alternative to the logit model for consumer choice modelling. While CLS has its roots in the artificial intelligence literature (Quinlan, 1983), in essence it is a nonparametric classification tree model that accommodates polytomous dependent variables. In most respects, the output from the CLS software is similar to that provided by the CART software.

CLS employs the entropy measure of impurity as the splitting criterion used in the construction of trees.

$$(1) \qquad i(t) = - \sum_{j} p(j|t) \ln [p(j|t)]$$

where: $i(t)$ is the measure of node impurity resulting from assigning a case in class j incorrectly to node t;

$p(j|t)$ is the proportion of cases belonging to class j that fall in decendant node t.

The entropy measure has a long history of use in this and related areas. Since CLS was viewed as an artificial intelligence technique, using the information theory measure of information at a source seemed appropriate. It has previously been used in statistics as a measure of discriminatory information and is a natural extension of a likelihood ratio test. Carman (1970) was the first to use this measure in tree analysis of brand choice. He converted nominal purchase pattern measures to a continuous entropy measure of brand loyalty and then used AID to search for the correlates of brand loyalty.

Like AID and CHAID, CLS does not prune the initial tree. CLS continues to make binary splits until some stopping criterion, such as no important reduction in $i(t)$, is satisfied. Thus, the user must employ an ad hoc procedure for pruning. CLS provides a Chi Square test for this purpose. Such a procedure may be quite appropriate in exploratory research. CLS does not provide any automatic validation procedure through some holdout or jackknife technique.

The Cart Model

The CART model has several advantages over AID, CHAID and CLS. First, it offers alternative criteria for splitting. Second, it facilitates the pruning process rather than simply growing an initial tree. Third, it provides a method for validation through holdout samples. The CART software accommodates either an intervally scaled dependent variable (regression tree) or nominally scaled dependent variable (classification tree) by so stating at the input stage. Here, regression trees are discussed first and compared to AID before moving to the classification problem. A summary of the key features of the five multivariate polytomous dependence analysis procedures is shown in table 1.

Splitting

CART modifies the traditional regression standard error formulae somewhat and sometimes uses heuristics in order to allow for distributions in the dependent variable (*y*) that are more general than those assumed in traditional least squares regression. In addition, trees are likely to be more informative than linear regression when the relationships are nonlinear. However after comparing several rules, the developers of CART returned to a rather standard criterion for splitting in the construction of an initial tree. It is the binary split that creates the greatest decrease in the average total within node sum of squares, R(T):

$$(2) \qquad R(T) = 1/N \sum_{\underline{t}}^{T} \sum_{\underline{n}}^{\underline{t}} (\underline{y}_n - \bar{\underline{y}}(\underline{t}))^2 \quad .$$

Stopping

The initial tree is grown by iteratively splitting nodes so as to maximize the decrease in R(T). The process continues in this way until all nodes have five or less cases in them or are completely pure, i.e., have within node sum of squares equal zero. Since this latter condition is rare in regression trees (not the case in classification trees), initial regression trees tend to be very large.

Pruning

Pruning is done in exactly the same way in classification and regression trees. The criterion is to find the number of terminal nodes that minimize the total error complexity measure, $R_\alpha(T)$. This measure is defined as,

$$(3) \qquad R_\alpha(T) = R(T) + \alpha|\underline{T}^{\sim}|$$

where: R(T) in regression is defined above in (2) and in classification analysis is based on the probability and cost that cases in a particular node are misclassified (defined below);

α is a complexity parameter set by the user;

$\underline{T}^{\sim}$ is the number of terminal nodes.

In other words, CART considers both the cost of misclassification and the cost of complexity. Both of these parameters can be set by the user. This is so for R(T) because the user can assign different costs of misclassification for any particular class. In our work, we regularly pruned trees by increasing the complexity parameter, α because the value of parsimony seemed to be easy to fix. In market segmentations it may be more common to set a misclassifcation cost based on the cost of including too many people in a segment vs. excluding too many.

Validation

Since R(T) needs to be estimated from the sample, standard errors estimates are required. For this purpose, validation procedures are available within the CART package. One option, for large samples, is to use a randomly selected test sample. The other option available in the software is a jackknife cross-validation routine. From either, the program calculates the standard error of R(T). The graph of R(T) follows a very standard form when plotted against the number of terminal nodes, T. As T increases, R(T) drops sharply at first and then is relatively flat for a long period before beginning to rise slowly. Desired is a tree near the beginning of the flat part of the curve so that $R_{\alpha}(T)$ is near its minimum. However, R(T) is an estimate subject to sampling error in the test or jackknife sample. The rule used by CART is to select the smallest tree, $T_{\underline{k}=1}$, such that $R(T_{\underline{k}=1})$ is less than one standard error different from $R(T_{\underline{k}=0})$.

Splitting Classification Trees

The rules for splitting must be different in the classification problem since there are no sums of squares. CART provides two options: Gini and twoing.

The Gini diversity index is defined as,

(4) $$\underline{i}(\underline{t}) = [\sum_j p(\underline{j}|\underline{t})]^2 - \sum_j p^2(\underline{j}|\underline{t}) = 1 - \sum_j p^2(\underline{j}|\underline{t})$$

where the notation is the same as in (1).
The Gini index has desirable mathematical and computational properties and appears to perform about as well as the entropy measure.

The twoing criterion takes a somewhat different approach to the problem of classification. That approach is to search at each node for a binary split that results in two classes that are most dissimilar. Specifically, the objective is to split so as to maximize the decrease in impurity in the data ($\underline{i}$):

(5) $$\max \underline{i} = p_L p_R/4 \ [\sum_j |p(\underline{j}|\underline{t}_L) - p(\underline{j}|\underline{t}_R)|]$$

where: p_L and p_R are the proportion of the node that go to the left and right branch respectively;

$p(\underline{j}|\underline{t}_L)$ is the proportion of cases belonging to class $\underline{j}$ that go to the left decendant branch.

Validating Classification Trees

Since pruning of the initial tree is done exactly as in regression, the discussion here will move directly to the subject of validation. The difference is that pruning requires finding an unbiased estimate of R(T) which in (2) was defined as the within node sum of squares. More generally, CART refers to R(T) as the "misclassification cost." For the classification problem, the expected value of this cost, $R^*(d)$, for a particular classification rule, d, is defined as,

$$R^*(d) = \sum_{j} \{\pi(j) \sum_{i} [c(i|j)\ Q^*(i|j)]\} \tag{6}$$

where: d is a classification decision;

$Q^*(i|j) = P[d(\mathbf{X}) = i|Y=j]$ is an estimate of the probability that a case in j is misclassified into i by d;

$c(i|j)$ is the cost of misclassifying a case in j into i;

$\pi(j)$ is the prior probability of being in class j.

Both the cost of misclassification and the prior probabilities can be set by the user. Only $Q^*(i|j)$ remains to be estimated by the program. In our analysis, equal (unit) misclassification costs and prior probabilities based on proportions in the total sample were used. These setting have the effect of producing better classifications of the largest classes. Note that the user has two degrees of freedom here that can be adjusted. If there is some clear managerial reason why misclassifications of certain kinds are to be avoided, then the cost of misclassification should be adjusted. Otherwise, it is probably better to experiment with priors that are closer to being equal rather than proportional.

CART produces unbiased estimates of $Q^*(\underline{i}|\underline{j})$ based on the test sample or cross-validation approaches described above. The standard errors can be computed by assuming a binomial model for the estimate of $Q^*(\underline{i}|\underline{j})$. The estimates are based on simple counts of cases misclassified by the validation samples. For a single test sample, this estimate is unbiased and straightforward to estimate. For the cross validation approach, there could be a trade-off between computation time and accuracy. Large jackknife subsamples produce better estimates but take more time to compute.

The program then searches for the pruning classifier, $\underline{d}$, that minimizes the total misclassification cost for the tree, R(T). As for regression trees, it searches in that neighborhood for the tree where $R_\alpha(T)$ is minimum using the "one standard error rule" defined above. Breiman et al. (1984) derive the statistical properties of all estimates or report simulation results where mathematical properties are complex.

Clearly, compared to CLS, CART is far more elegant and places more emphasis on pruning. Indeed, because it offers two rules for building the initial tree, it also provides more information to the user than does the single entropy rule used by CLS. However, one might have liked the entropy option added to the choices for initial tree building in CART because of its extensive past usage.

Options

CART also provides a number of other options that should be convenient in some applications. One permits the user to specify that splits be done on certain combinations of variables. An example for categorical variables would be a data set where age of youngest child at home was one variable and child in university was another variable. CART can be instructed to make a Boolean combination of these two variables and treat them as a single variable for splitting. CART will also estimate linear combinations of continuous variables and split on the sum of the combination rather than on the individual variables. Of course, such combination variables could be constructed before running CART if the data were in a convenient statistical package. For example, a life cycle variable could be constructed before analysis or by the CART program.

A second option is a missing data algorithm that replaces the missing value with an estimate and then reports its classification or misclassification just as it would for any other case. The estimate is constructed as a linear combination of the nonmissing variables most likely associated with it. Note that this is a richer procedure than simply assigning an average value to missing cases.

A third useful option allows the user to store a tree and then to test predictive efficiency by running a new set of data down this tree. This procedure is analogous to testing the relationship on a new set of data. With this option, case-by-case assignment information can be obtained, so the user knows case ID, actual category, and predicted category.

An option the authors found useful was the reporting of a measure of the relative importance of each predictor variable. Variables that were very close to being used for splitting but were seldom or never selected, perhaps because of colinearity, could be identified with this information. Such a variable could be forced into the tree by eliminating a colinear variable or by relaxing the complexity parameter.

Finally for regression trees, the program will build an initial tree based on least absolute deviations rather than least squares. This option is recommended for data where considerable nonlinearity is expected. However, tests show it to be inferior in most cases since the nature of tree analysis is designed to deal with discontinuity.

The problem setup is menu driven in interactive mode with easy access to help and on-line documentation at each step. The CART software is not quite as user friendly as it might be. The problems were not major in Version 1.1, and some improvements have been made in Version 1.3. The documentation available to us was only in draft form and requires refinement (California Statistical Software, Inc., 1985).

Data file restrictions and specification were not overly restrictive. Data must be in an ASCII file that the program reads through a FORTRAN format statement and a series of questions about each variable asked through the menu.

Memory management requires some learning. The program requires about 400 KB plus the workspace required for the problem at hand. The example above required 64 KB of workspace and could not handle the cross-validation algorithm. Obviously, CART must store a great deal of information, at least temporarily, in the process of searching over many possible binary splits in multiple subsamples. The program should be run on a computer with generous virtual memory, e.g., IBM3090.

An Application in Segmentation Research

In this section some example applications of CART for a large data set are presented. These data are from the 1983 Survey of Consumer Finances and comprise returns from a complex survey of 2822 households conducted by the University of Michigan Survey Research Center for the Federal Reserve Board of Governors. The study collected enough financial data to construct a balance sheet for each household. The structure of this financial portfolio was the dependent variable of interest. In addition, considerable data were available on the socioeconomic-demographic characteristics of the household as well as some information on attitudes concerning saving, investing, and borrowing.

The financial portfolios were used in a cluster analysis in order to segment the households into groups exhibiting similar patterns of usage of various financial instruments for asset and liability management. CART was utilized in two ways. CART was first used to validate the cluster analysis. That is, given the segment assignments based on the cluster analysis and the same set of financial variables, would CART classify the households to the same cluster?

Next, CART was used in the more traditional way to determine what socioeconomic-demographic and attitudinal variables would correlate with cluster membership. Note that cluster assignment is a nominal variable and the predictor variables were a mixture of categorical, ranked, and continuous variables. Thus, these data provide an ideal setting for tree analysis.

Use in Validating Clusters

The cluster analysis produced 16 segments based on the composition of the financial portfolios. Some of these were quite small, making up less than one percent of the sample. The CART tree was constructed using the Gini rule, unit misclassification costs, and proportional prior probabilities. With a 16 category nominal dependent variable, the tree constructed was very large. $R_\alpha(T)$ did not start to flatten out until the tree had 29 terminal nodes, and when the complexity parameter, α, was set very small, classification improved up to a tree size of over 80 terminal nodes. The correct classification rates were quite high for the large clusters but less good for the small ones. Experimentation with the prior probabilities, $\pi(j)$, showed that moving away from proportional priors did not improve the overall fit. Two of the smaller clusters were simply too obscure to be useful and were similar enough to a single other cluster to be combined. In the end then, a 14 cluster solution produced 79 terminal nodes, had an overall correct classification of 93 percent, and had only two clusters where the correct classification rate was below 75 percent.

Cluster analysis is a procedure that usually leaves the user rather uncertain about the validity of the solution. In this case, two cluster routines and a holdout sample had been used in order to try to validate the 16 cluster solution. CART clarified any remaining doubts quite nicely. Twelve of the clusters were clearly identified, two others were somewhat suspect, and two were probably illusory. In addition, more was learned about the relative importance of the balance sheet variables in forming the clusters than had been learned from the clustering procedure itself.

Application to Segmentation

The 14 clusters then became the dependent categorical variable in a multivariate analysis designed to determine if the socioeconomic-demographic-attitudinal variables would predict cluster membership. There were 12 predictor variables, 7 continuous and 5 nominal, each with either 7 or 8 categories. The initial tree was constructed using the twoing rule, unit misclassification costs, and proportional priors. After some exploratory runs, the complexity parameter was used to prune the tree to one with 31 terminal nodes. An abbreviated version of the final tree, showing 25 of the terminal nodes, is provided in figure 1. The actual output produces a graphic similar to figure 1.

As a study in correlates of segment membership, this analysis produced the usual, low correlations. CART was able to correctly classify only 50.9 percent of the cases correctly. Because proportional prior probabilities were used, it did best in predicting the two largest segments, hitting on 89 and 72 percent of the cases respectively. In ten segments, correct classifications were under seven percent. Figure 1 shows only those terminal nodes for the four clusters with strong predictive validity. But one would not expect that a tree analysis would solve the problem of low correlations between segment membership and socioeconomic-demographic variables. The interest here is on what was learned, how reliable the solution was, and how the CART solution compared to other approaches, namely logit regression.

Figure 1 shows most of the terminal nodes that are classified as segments 1, 12, 13, and 14. Note that income is the most important variable in explaining segment membership and that segment 1 is the lowest income segment and segment 14 is the highest income segment. Indeed, segment 14 was identified with only two predictors and was composed of households with a 1982 income in excess of $395,000. One would expect such a high-income group to be easily identified.

Segment 12 is more interesting. Note that it generally was composed of households with incomes between $56,500 and $395,000, but that lower income households are also in this cluster when other characteristics are present. These households tended to save for the future and were over 65 years of age. Households with persons under 65 and incomes from $26,600 to $56,500 were in segment 12 when they were saving for real property or the future. Attitudes toward borrowing also were significant for this group.

These interactions are probably far too complex to be uncovered by regression type models. The advantages of the latter is that they will report the importance of the various predictors, but will say less about how the variable work together to form the segments. CART does both. Table 2 shows another output of the CART program that provides a measure of the relative importance of each variable. A few features of this measure merit comment.

Note that variables 1634 and 3104 receive low relative importance scores and are never used in building the tree. However, note that variable 1730 receive rather large relative importance scores but also was not used in the construction of these first 25 nodes. Variables 4559 and 4560 show how the CART tree deals with interactions and multicolinearity that can be very meaningful to the user. Variable 4559 is the estimated slope of wage growth for the head of the household while under age 35. This measure is correlated with wage growth slope over age 35, variable 4560. The latter is used first in the tree and then the former splits on one of the 4560 branches.

Education of spouse, variable 1730, behaved somewhat differently. Since education of head, 1630, was used in the analysis, 1730 was not used but was relatively important.

While the results of the tree analysis shown in figure 1 appear interesting and useful, the tree still contained a complex of interactions in the lower branches, some of which were not shown here. Just how reliable was the pruning? CART provides a number of approaches for investigating this question. First, the user can increase the complexity parameter and alter the prior probabilities to see how pruning will be revised.

Second, the Gini rule rather than twoing could be used to construct the initial tree. The authors of CART do not voice an opinion on which is better. Gini was run on these data. The relative importance measure, table 2, was very similar for the two classification rules but there were some differences between the trees. Gini did a somewhat better job at classifying segment 13, but did less well with the larger segments. For this problem, twoing appears to have done a better job.

Finally, the most powerful reliability check is to split the sample and use one of the three validation options available in the computer package (learning-test samples, jackknife cross-validation, second sample run on first tree). With the large sample available, the learning-test sample approach was used in the present example. The results changed considerably. The tree was pruned in such a way that the only variables used in splitting were those with a relative importance over 40 (See table 2.). Note in figure 1, this results in all splits on reasons for saving, variables 5401 and 5402, being eliminated. While the hit rate was about the same in this simplified tree, the structure changed considerably.

Was the change in result caused by lack of real importance of these attitudes or by the reduction in sample size? Was the price of parsimony too great? An answer to these questions should have been available by running the jackknife cross-validation procedure. However in the version of the program available to us, the memory was not sufficient to permit running this procedure.

Another way to investigate this matter would be to construct a tree using half the sample and then run the other half of the sample down this tree. Put another way, instead of using the learning-test sample validation approach described above, construct an exploratory tree using half the sample and then test the tree with the second half-sample. It may seem that these two methods are identical. However, this is not quite the case because the standard error of R(T) is estimated differently in the two approaches.

Comparison With Multinomial Logistic Regression Results

The SAS Logistic Regression procedure was used for the purpose of comparing the results of CART with one of the alternative methods mentioned earlier. The SAS procedure, too, does not provide the most user friendly results for a situation like this one where there are 14 clusters for a dependent variable. The model assumes each predictor has the same effect on each cluster and that the differences among clusters are reflected only in the intercept term (SAS Institute Inc., 1989). Thus, the procedure produces 13 intercepts with the dependent variable being expressed as a cumulative probability of a case being in a cluster on a cluster greater in number. These probabilities are based on the inverse of the cumulative logistic distribution, so the signs on the regression coefficients all have the reverse sign to the one hypothesized. For example, education has a positive impact on sophistication of portfolio, but will be reported with a negative sign in the regression output. In order to determine the cluster to which the prediction equations assign each case, one needs to compute differences in the predicted probability values and assign a case to a cluster based on where the cumulative probability increases the most.

In terms of overall goodness of fit measures, the SAS routine reports a number of statistics. The Chi Square tests of overall model fit were significant with risk of .0001. An observation is said to be concordant when it has a lower predicted probability than an observation is a cluster with a smaller cluster number (because of the use of the inverse logistic function). Seventy three percent of the observations in this sample were concordant. However, concordance alone overstates the proportion of observations assigned to the correct cluster. Based on computing concordance for all possible combinations of pairs in the sample, it is possible to calculate a rank correlation coefficient to measure overall goodness of fit of the model. The most conservative of these nonparametric R^2 measures are Goodman-Kruskal Gamma and Kendall's Tau-a. These produced fits of .46 and .37 respectively.

The relative importance of the predictors, as measured by standardized regression coefficients, are shown in table 3. The results are very similar to those found by CART. Note that the two slopes of wage growth, 4559 and 4560, are less important in the regression results, where they are treated as continuous variables, than in CART where they can be split so as to have discontinuities and nonlinearity.

From the point of view of prediction of correct cluster classification, CART performed somewhat better than the logistic regressions, 51 and 43 percent respectively. These are shown in table 4. Note that the regressions assigned all cases to just three clusters. Clearly, the absence of assumptions regarding functional relationships in CART makes for better assignment. Note particularly that the regression does not classify anyone as being in Cluster 14, a very high wealth cluster of some interest. On the grounds of ease of interpretation and ease of use of the software, CART certainly does no worse than logistic regression.

A Validation Test

The 1983 Survey of Consumer Finances was repeated as a longitudinal study in 1986. Thus, it was possible to investigate the stability of the CART and logit results with an analysis of the 1986 data that was parallel to the 1983 analysis. The same 2822 households were involved and the same predictor variables were involved. Again, 14 clusters were used from the cluster analysis of financial portfolio types. Again, clusters 2 through 11 produced small groups that were hard to predict and shifts within these clusters between the two years were hard to understand. Figure 2 and tables 5 and 6 report the results in a form similar to those used for the 1983 findings.

The trees are similar in that income, education, and life cycle are important variable in explaining cluster membership and that the attitudinal variables dealing with borrowing and saving also are important. While far from being identical, these trees do suggest the basis for a targeted marketing program to four of the segment clusters.

With regard to variable importance and prediction, the logistic regressions again faired somewhat less well than CART. Again, the two wage slope variables, that are important in CART and statistically significant, do not do well in the logistic regressions where they are treated as continuous variables. They had signs that are contrary to theoretical expectation and were not statistically significant. The correctly assigned cases were 47 percent for CART and 43 percent for the regressions. Again, the regressions assigned all households to just three clusters.

Conclusions

The test of the CART tree analysis used here was market segmentation research, an area where socioeconomic and demographic variables have a long history of low explanatory power. Thus, the fact that overall model fit was just short of .50 is not surprising. The more appropriate criterion is whether the trees provided more insight and understanding than did the regression models.

While the logistic regressions did not do badly in these applications, the CART trees consistently did a somewhat better job of producing results that could be meaningfully interpreted for understanding and application. The trees are simply more meaningful than regression coefficients. In addition, the freedom from assumptions about functional form also facilitated understanding. From the standpoint of overall fit and prediction, again CART performed somewhat better than the regressions.

While one would hope that the software for the CART program would be made a bit more user friendly, the use of the SAS logistic regression package was also not very convenient. Indeed, the convenience of SAS data bases and the SAS language were the real sources of convenience of the SAS package -- not the logistic regression procedure.

This research suggests that CART trees do represent a significant improvement over previous tree analysis programs and over logistic regression. CART does have important application in marketing research. The procedure is recommended for both academic and commercial work where the variables are at different levels of measurement. In addition to segmentation with socioeconomic variables, the procedure should have usefulness in behavioral segmentation based on identifying innovators, early adopter, brand loyals, or micro-responders to direct marketing. This list constitutes the core activities of marketing managers today.

REFERENCES

Armstrong, J. Scott and James G. Andress (1970), "Exploratory Analysis of Marketing Data: Trees vs. Regression," Journal of Marketing Research, 7 (November), 487-92.

Assael, Henry (1970), "Segmenting Markets By Group Purchasing Behavior: An Application of the AID Technique," Journal of Marketing Research, 7 (May), 153-8.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone (1984), Classification and Regression Trees. Monterey CA: Wadsworth, Inc.

Bucklin, Randolph E. and James M. Lattin (1991), "A Two-State Model of Purchase Incidence and Brand Choice," Marketing Science, 10 (Winter), 24-39.

California Statistical Software, Inc. (1985), "An Introduction to CART Methodology," Lafayette, California.

Carman, James M. (1970), "Correlates of Brand Loyalty: Some Positive Results," Journal of Marketing Research, 7 (February), 67-76.

Currim, Imran S., Robert J. Meyer and Nhan T. Le (1988), "Disaggregate Tree-Structured Modeling of Consumr Choice Data," Journal of Marketing Research, 25 (August), 253-65.

Dhrymes, Phoebus J. (1986), "Limited Dependent Variables," in Zvi Griliches and Michael D. Intriligator (eds.), Handbook of Econometrics, Vol. 3, New York: Elsevier Science Publishers, 1568-1631.

Dillon, William R. and Matthew Goldstein (1977), "VARSEL: A Stepwise Discrete Variable Selection Program," Journal of Marketing Research, 14 (August), 419-20.

Fenwick, Ian, D. A. Schellinck and Kenneth W. Kendall (1983), "Assessing the Reliability of Psychographic Analyses," Marketing Science, 2 (Winter), 57-73.

Gray, Henry L. and W. R. Schucany (1972), The Generalized Jackknife Statistic, New York: Marcel Dekker.

Green, Paul E., Frank J. Carmone and David P. Wachspress (1979), "On the Analysis of Qualitative Data in Marketing Research," Journal of Marketing Research, 14 (February), 52-9.

Hand, David J. (1982), Discrimination and Classification, Chichester: Wiley.

Jones, Howard L. (1956), "Investigating the Properties of a Sample Mean by Employing Random Subsample Means," Journal of the American Statistical Association, 51 (March), 54-83.

Kinnear, Thomas C. and James R. Taylor (1971), "Multivariate Methods in Marketing Research: A Further Attempt at Classification," Journal of Marketing, 35 (October), 56-59.

Maddala, G S (1983), Limited Dependent and Qualitative Variables in Econometrics, Cambridge: Cambridge University Press, 16-41.

Mosteller, Frank and John W. Tukey (1968), "Data Analysis Including Statistics," in Gardner L. Lindzey and Elliott Aronson (eds.), Handbook of Social Psychology, Reading MA: Addison-Wesley.

Perreault, William D. Jr. and John M. Gwin (1978), "CATCLASS: A Model to Classify Consumers Based on Multivariate Categorical Data," Journal of Marketing Research, 15 (February), 113-15.

_____ and Hiram C. Barksdale Jr. (1980), "A Model-Free Approach for Analysis of Complex Contingency Data in Survey Research," Journal of Marketing Research, 17 (November), 503-15.

Quinlan, J. Ross (1983), "Learning Efficient Classification Procedures and Their Application to Chess End Games," in R. S. Michalski, J. G. Carbonell, and T. M. Mitchell (eds.), Machine Learning: An Artificial Intelligence Approach, Palo Alto: Tioga Publishing, 463-82.

SAS Institute Inc. (1989), "The Logistic Procedure," SAS/STAT Users' Guide, Vol 2. Cary NC: SAS Institute, 1071-1126.

Sonquist, John A. (1970), Multivariate Model Building: The Validation of Search Strategy, Ann Arbor MI: University of Michigan Institute For Social Research.

TABLE 1. SUMMARY COMPARISON OF MULTIVARIATE DEPENDENCE ANALYSIS WITH POLYTOMOUS DEPENDENT VARIABLE

	AID	CHAID	LOGIT	CLS	CART
Dependent Variable	Polytomous	Polytomous	Mode at zero	Polytomous	Interval or Polytomous
Splitting	Reduction in variance	Chi Square	Reduction in variance	Reduction in entropy	Choice. Reduction in: variance; Gini; twoing similarity; absolute dev.
Stopping	Small node or small reduction	Small node	User specified model	Small node or small reduction	Small node
Pruning	None	Collapse categories	Significance tests	None	Total misclassification cost
Validation	None	None	None	Split sample or jackknife	Split sample or jackknife

TABLE 2. VARIABLE IMPORTANCE IN EXAMPLE CART TREE

VARIABLE NUMBER	VARIABLE NAME	VARIABLE TYPE	RELATIVE IMPORTANCE
1305	Income	Interval	100
1630	Education of Head	Interval	59
4559	Wage Growth Slope, Under 35	Interval	52
5513	Most important attribute of a loan	8 Classes	51
4560	Wage Growth Slope, 36-55	Interval	48
3116	Stage of Life Cycle	7 Classes	43
1730	Education of Spouse	Interval	39
5502	Attitude Toward Borrowing	8 Classes	38
5402	Attitude Toward Saving, 2d Reason	7 Classes	36
5401	Attitude Toward Saving, 1st Reason	7 Classes	34
1634	Condition of Health, Self-Reported	Interval	27
3104	Number in Household Under 18	Interval	27

TABLE 3. VARIABLE IMPORTANCE IN LOGISTIC REGRESSION, 1983

VARIABLE NUMBER	CART REL. IMPORTANCE	LOGIT STD. REGRESSION COEFFICIENTS
1305	100	-.53
1630	59	-.30
5513	51	-.18,-.18,-.18,-.15,-.08,-.06,-.04*
3116	43	-.19,-.15,-.13,-.08
1730	39	-.12
5401	34	.11,.11,.10,.09,.06
4559	52	-.11
5402	36	-.09,.05,.04
4560	48	-.09
1634	27	.06
5502	38	.04,.03,-.04
3104	27	-.03

* With categorical variables, dummy variables were created.

TABLE 4. CART AND LOGIT ASSIGNMENT ACCURACY, 1983

PREDICTED	CART					LOGIT					
CLUSTER	1	2-11	12	13	14	1	2-11	12	13	14	TOTAL
ACTUAL CLUSTER											
1	769	8	68	16	0	713	0	148	0	0	861
2-11	492	41*	253	45	1	469	0	363	0	0	832
12	173	2	534	11	21	186	0	484	71	0	741
13	113	6	124	51	2	114	0	176	6	0	296
14	2	0	35	0	55	2	0	23	67	0	92
TOTAL	1549	57	1014	123	79	1484	0	1194	144	0	2822

Correct Assignment = 1436 / 2822 = 50.9%   = 1203 / 2822 = 42.6 %

* 27 assigned to correct cluster in this group.
In the abbreviated tree shown in figure 1,
the correct classification was 49.9 %.

TABLE 5. 1986 VARIABLE IMPORTANCE IN CART OF LOGIT

VARIABLE NUMBER	VARIABLE NAME	CART REL. IMPORTANCE	LOGIT STD. REGRESSION COEFFICIENTS
1301	Income	100	-.42
1630	Education of head	60	-.41
1131	Stage of life cycle	42	-.16,-.11
1730	Education of spouse	38	-.26
1822	Wage growth slope, <35	62	.06
1823	Wage growth slope, 36-55	53	.08
5513	Important loan attributes	46	-.03
5502	Attitude toward borrowing	45	-.03
1218	Attitude toward saving, 1st	34	-.16,-.12,-.06
1219	Attitude toward saving, 2nd	43	-.07,-.04
1104	No. in household < 18	30	-.16
1634	Condition of health	29	.10

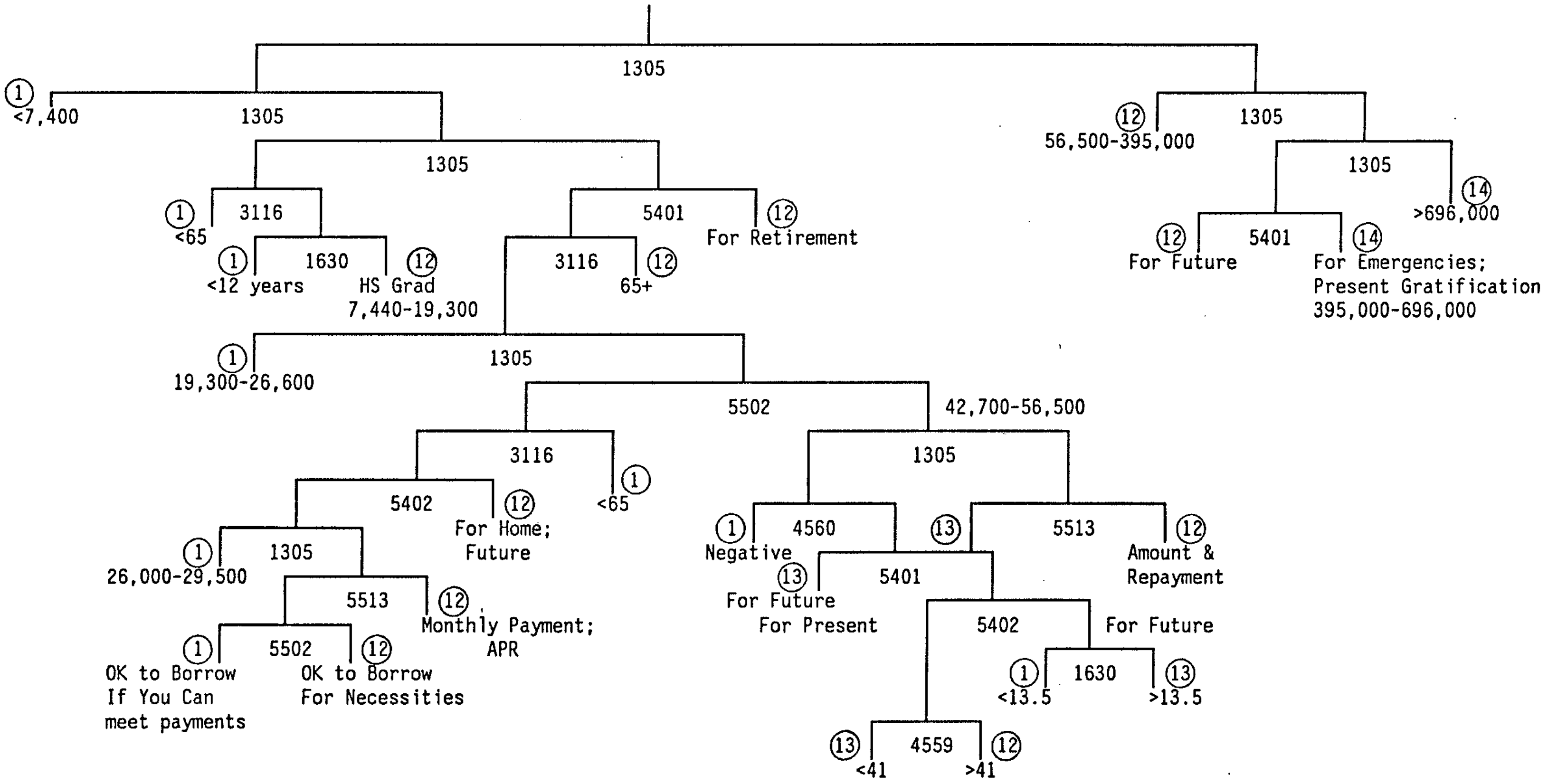
TABLE 6. CART AND LOGIT ASSIGNMENT ACCURACY, 1986

PREDICTED	CART					LOGIT					
CLUSTER	1	2-10	12,14	11	13	1	2-10	12,14	11	13	TOTAL
ACTUAL CLUSTER											
1	588	18	44	31	0	573	0	108	0	0	681
2-10	460	38*	285	197	4	495	0	489	0	0	984
12,14	129	18	559	41	26	154	0	612	0	7	773
11	91	9	91	112	0	112	0	191	0	0	303
13	0	0	37	0	44	0	0	64	0	17	81
TOTAL	1268	83	1,016	381	70	1334	0	1464	0	24	2822

Correct Assignment = 1338 / 2822 = 47.4 %   = 1202 / 2822 = 42.6 %

* 35 assigned to correct cluster in this group.
In the tree tree shown in figure 2,
the correct classification was 45.5 %.

FIGURE 1. 1983 CART TREE

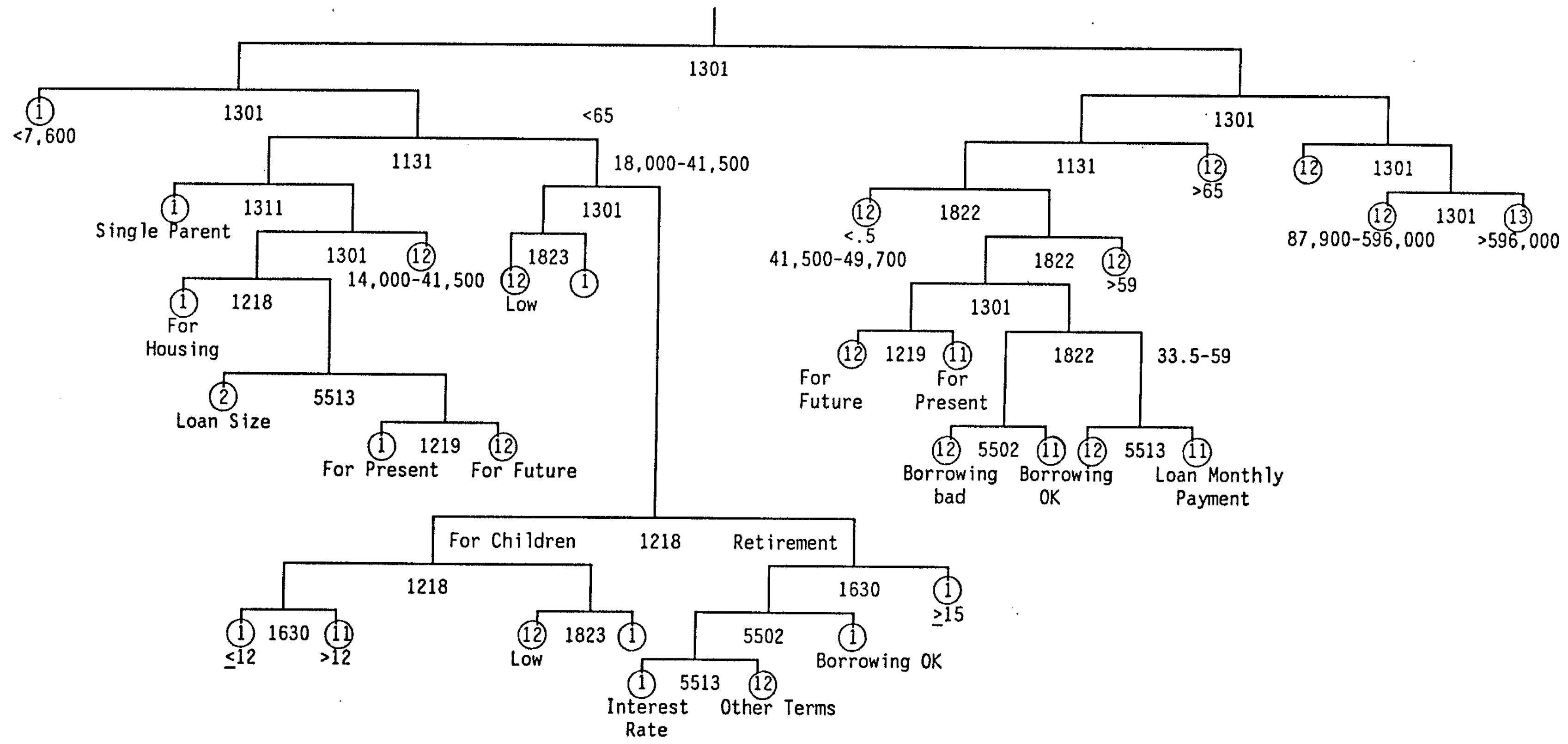


KEY:

Variable definitions shown in Table 2.

Numbers in circles indicate cluster number.

FIGURE 2. 1986 CART TREE



KEY:

Variable definitions shown in Table 5.

Numbers in circles indicate cluster numbers.