THE PINHAS SAPIR CENTER FOR DEVELOPMENT
TEL-AVIV UNIVERSITY

Mortgage Default:
Classification Trees Analysis

David Feldman[1] and Shulamith Gross[2]

Discussion Paper No. 3-2003

October 2003

[1] Department of Business Administration, School of Management, Ben-Gurion University, E-mail: feldmand@bgumail.bgu.ac.il.
[2] Statistics Program Director, The National Science Foundation, email: sgross@nsf.gov.

The paper can be downloaded from http://econ.tau.ac.il/sapir

Abstract

We introduce the powerful, flexible and computationally efficient nonparametric Classification and Regression Trees (CART) algorithm to the real estate analysis of mortgage data. CART.s strengths in dealing with large data sets, high dimensionality, mixed data types, missing data, different relationships between variables in different parts of the measurement space, and outliers, is particularly appropriate for our data set. Moreover, CART is intuitive and easy to interpret and implement. We discuss the pros and cons of CART vis-à-vis traditional methods such as linear logistic regression, nonparametric additive logistic regression, discriminant analysis, partial least squares classification, and neural networks, with particular emphasis on real estate. We apply CART to produce the first academic mortgage default study of Israeli data. We find that borrowers. features, rather than mortgage contracts features, are the strongest predictors of default if accepting "bad" borrowers is more costly than rejecting .good. ones. If these costs are equal, mortgage features are used as well. The higher (lower) the ratio of misclassification costs of bad risks versus good ones, the lower (higher) are the resulting misclassification rates of bad risks and the higher (lower) are the misclassification rates of good ones. This is consistent with real world stylized facts of rejection of good risks in attempt to avoid bad ones.

JEL Codes: C12, D12, G21, R29
Key Words: mortgage default, Classification and Regression Trees, misclassification error

#This paper uses the data that was collected for the proposal, .Mortgage Default in Israel,. by D. Ben-Shahar and D. Feldman, submitted to the Sapir Center for Development, Tel-Aviv University. We thank a major Israeli mortgage bank and Ephraim Goldin from GStat Ltd. for the data and cooperation. We thank D. Ben-Shahar for essential help in getting the data. We thank Brent Ambrose, D. Ben-Shahar, Leo Breiman, Ayala Cohen, Yongheng Deng, Robert Edelstein, David Nickerson, Richard Olshen, Boaz Rottenberg, Mordechai Rottenberg, and Tatiana Umansky for helpful discussions, Bank of Israel for providing documentation and information, and Sivan Weiss for research assistance. We also thank workshops participants at Ben-Gurion University of the Negev, University of Haifa, The Cambridge-Maastricht Real Estate Finance and Investment Symposium, Cambridge, and The French Finance Association Annual International Conference, Lyon. We thank the Sapir Center for Development for financial support.

1 Introduction

We introduce the powerful, flexible and computationally efficient nonparametric Classification and Regression Trees (CART) [Breiman, Friedman, Olshen, and Stone (1998)[1] (BFOS)] algorithm to the real estate analysis of mortgage data. CART's strengths in dealing with large data sets, high dimensionality, mixed data types, missing data, different relationships between variables in different parts of the measurement space, and outliers, is particularly appropriate for our data set. Moreover, CART is intuitive and easy to interpret and implement. We discuss the pros and cons of CART vis-à-vis traditional methods such as linear logistic regression, nonparametric additive logistic regression, discriminant analysis, partial least squares classification, and neural networks, with particular emphasis on real estate.

As far as we know this is the first application of CART in an academic study of real estate data and the first academic mortgage default study of Israeli data. We find that borrowers' features, rather than mortgage contracts features, are the strongest predictors of default if accepting "bad" borrowers is more costly than rejecting "good" ones. If these costs are equal, mortgage features are used as well. The higher (lower) the ratio of misclassification costs of bad risks versus good ones, the lower (higher) are the resulting misclassification rates of bad risks and the higher (lower) are the misclassification rates of good ones. This is consistent with real world stylized facts of rejection of good risks in attempt to avoid bad ones.

CART classifies individuals or objects into a finite number of classes on the basis of a collection of features, or independent variables. CART uses binary trees, a method that Morgan and Sonquist introduced in the sixties at the University of Michigan and Morgan and Messenger developed there in the seventies into an ancestor

[1] The first version of this book is from (1984).

classification method. CART strengthens and extends these original methods. It was first introduced independently by Breiman and Friedman in 1973, who later joined forces with Stone and then with Olshen. CART was first introduced to the general reader and is fully described by BFOS.[2] Although we use CART as a classification tool, it is also a regression tool. In fact, any guided classification, including the CART algorithm, may be regarded as a regression method where the response variable is categorical. Presented this way, it becomes evident that CART chief competitors are discrimination methods in general, and polytomous logistic regression in particular.

Our main purpose in analyzing the mortgage data is the binary classification of borrowers into two risk classes: potential defaulters and those unlikely to default. We use a data base, which we refer to as a *learning sample*, to develop the decision rule for the classification. Our learning sample consists of data both on the predictors, which we also call independent variables or *features*, and on the binary *outcome* variable: defaulted, or did not default. Our learning sample consists of data on 3,035 mortgage borrowers. The features include asset value, asset age, mortgage size, number of applicants, the main applicant's occupation, income, and family information, and other characteristics of the asset and the applicant: thirty three features in all.

1.1 Why CART?

A particularly important CART feature that deserves special mention is its *treatment of missing data*. Regression, including logistic regression, and other classification methods that use feature data to associate individual cases with one of two or more classes, require the elimination of whole observation vectors when even one of their elements is missing. CART seems to have introduced a novel way to deal with missing data efficiently, particularly for classification and prediction. First, the

[2] The first version of this book is from (1984).

classification algorithm creates a simple binary tree structure. Then, it uses this tree structure to classify new cases. In the likely event that a case with missing features is presented to be classified, CART offers alternative trees for each combination of missing features. To describe this important feature of CART, we will schematically describe (in Section 2) the binary classification tree that CART produces, couching the description in our example of mortgage applicants' risk assessment when necessary.

Among the important facilities that CART offers is a *weighting facility*. This facility is particularly relevant when the learning sample does not represent a simple random sample from the population, e.g., when the sample is stratified. For example, when the tree is intended to discriminate between members of a very rare class in the population, and the remainder of the population, it is often advantageous to "over-sample" the rare subset of the population. Weighting the different classes in a way that compensates for their proportion in the population allows CART to produce a consistent classification procedure. We have used this facility in analyzing the Mortgage data, because although the proportion of defaulters in the data is under 10%, in fact approximately equal numbers of defaulters and non-defaulters were selected from the bank database of mortgage customers.

A Bayesian decision maker will also find a *Bayesian classification feature* in CART, where the user provides subjective class probabilities that the algorithm uses to evaluate error rates of candidate trees using its *Cross Validation facility* (see below), before making its final tree choice. These prior probabilities serve in effect as user-selected class weights, and are therefore useful for analyzing data from complex samples, even when the researcher is not an avowed Bayesian.

For selecting the best classification tree for a particular set of requirements, and to evaluate the classification performance of a selected tree, CART uses robust methods

such as Cross-Validation. As is well known, a naive classification error-rate that is computed directly on the entire data set tends to be over-optimistic. It is usually recommended that a certain portion of the data be kept out of the classification tool selection process, and then be used for testing the selected classification tool. When CART constructs a classification tree, it performs this procedure, usually called Cross-Validation, automatically. CART divides the data into K (usually 10) equal parts, using K-1 parts to construct the tree, and testing it on the remaining data, repeating this procedure K times. Section 2 explains this procedure in more detail.

CART handles independent categorical variables as easily as continuous ones, and is resistant to outlying values present in one or more continuous features. CART's resistance to outliers is due to its use of splits of the form $X \leq s$ or $X > s$. Such splits hardly depend on outlying values. Furthermore, the splits considered by CART are invariant under monotone transformations. That is, any monotone transformation such as log or square root, of one or more of the features, does not alter the final tree. Therefore, CART does not require any pre-transformation of the data.

Because the selection of candidate variables for splitting may be too limiting, CART permits the expansion of the set of candidate variables to include linear combinations of variables in the feature set. Naturally, any user who wishes to use a different function of existing features may define it and add it to the feature set. Moreover, the choice of features to be included in the feature space will depend on the subject matter, and is left to the user to select.

The process of selecting the features to be included in the tree, and the structure of the binary tree itself is completely automatic. No expert statistician is required to reduce the number of features to a manageable number, and no transformations are required.

Another advantage of CART stems from the tree structure of the decision algorithm: decision processes of subgroups in the population may reveal themselves. In addition, CART is computationally efficient, and has unusual ability to find quasi-efficient combinations of features for classification.

1.2 CART and Its Competitors

The task of predicting a binary outcome from a collection of relevant features is traditionally carried out using well known tools such as logistic regression. There are two main types of logistic regressions: the completely parametric linear one, and the nonparametric additive one, see Hastie, Tibshirani, and Friedman (2001). In the latter, functions of the features are inserted into the logit function[3] additively, and the form of each function is left open and is estimated by the data. In our case, the logit would have been the log of the odds of being classified a likely defaulter. These two logistic procedures may be considered complementary. When the dependence of the logit on the collection of features is patently nonlinear, the additive logistic procedure is usually adopted. Another class of classifiers are the linear, quadratic, or nonparametric discriminant analyzers[4] [see Hastie, Tibshirani, and Friedman (2001)]. The first two procedures divide the feature space into two complementary subspaces assuming normality of the features. This assumption is unlikely to hold in most cases, particularly when many of the features are ordinal or nominal categorical variables, as is common in business data. The nonparametric procedures include K-nearest neighbor rules,[5] partial least squares classifiers, or neural networks.

---

[3] The logit function is the log of odds function. Thus if the odds are n:k (p/1-p), the logit function is log(n/k) [log(p/(1-p))]. The logit function is also the inverse function of the logistic cumulative distribution function, $f(x) = \left(1 + e^{-\frac{x-\mu}{\sigma}}\right)^{-1}$.

[4] Roughly speaking, linear, non-linear, and non-parametric analyzers divide the space of features, linearly, non-linearly, and by ordinal ranking, respectively.

[5] The K-nearest neighbor rule due to Fix and Hodges (1958) may be succinctly defined as follows: Let $d(X,Y)$ be a distance function, say Euclidian distance, between two points, X,Y in the feature space. Fix

When we compare CART to traditional methods, we note that, as is the case of CART (see below), traditional methods do not truly search for an optimum model in an organized fashion. Consider logistic regression, or discriminant analysis (of any type). Given a group of features, these procedures will find the optimal coefficients for the linear or quadratic function that will split the feature space into subsets that are predicted to belong to different classes. But 'optimality' here is definitely model-dependent. Model parameters that are optimal under the assumption of a logistic model, are not, strictly speaking, optimal under a probit[6] model. Thus optimality is contingent on the model assumed. In order to find the optimal model, logistic regression and discriminant analysis may, depending on the software used, search for the optimal subset of independent variables that minimizes the Akaike information criterion (AIC),[7] or similar criteria, among all models built on the given features. In the case of logistic regression for example, that choice, optimal for estimating the probability of belonging to a given class (e.g., being a potential defaulter in our example) provided the logistic model is correct, may not be optimal for predicting class identity (e.g., potential defaulter). As is well known, the use of logistic regression for classification usually involves the application of ROCs (Receiver Operating Curves),[8] and the use of the latter is not fully understood in terms of optimal classification. The curve helps determine the cutoff probability p* that separates class predictions (in the binary classification case). If the estimated conditional probability of being a 'case' exceeds p*, the individual is classified as a 'case', and a 'non-case' otherwise. However, the rules governing the choice of p* are not clearly associated with any single optimality criterion. It is also unclear that the

---

an integer $K>0$. Classify a new point X into class j if the largest number of points among the K points nearest to X that belong to one class, belong to class j.

[6] The probit function is the inverse normal cumulative distribution function.

[7] AIC is a likelihood related criterion used to compare parametric statistical models (particularly non nested ones).

[8] A ROC curve is a plot of the sensitivity versus one minus the specificity as a function of the splitting value, for a binary classifier. See next paragraph for the definitions of sensitivity and specificity.

optimal estimated logits, and the subset of features selected, lead directly to 'optimal' classification.

The various discriminant procedures lead directly to classification, without the estimation procedure required by logistic regression. Nonetheless, the latter is usually found to be more efficient when the specificity (the probability of classifying non-cases as such) and sensitivity (the probability of classifying cases as such) achieved by the two procedures are considered. The fact that linear and quadratic discriminant analysis are based on the assumption of normal data may explain their lack of efficiency in real data.

Several authors have addressed the question of the relative efficiency of tree-based methods such as CART, neural networks classifiers, and logistic regression, including spline-based logistic regression.[9] For comparative studies of the various methods, see for example Rousu, Flander, Suutarinen, Autio, Kontkanen, and Rantanen (2003), and Moisen and Frescino (2002). Of the many remaining traditional classification methods, we mention in particular those that are reported in the literature as being particularly effective (See Breault, Goodall, and Fos (2002) for a study that considers probably most methods of classification in use, but uses a questionable method of comparison on real data). Two that we find particularly interesting are the Partial Least Squares (PLS) discrimination procedure, and neural networks for Discrimination. Both methods start out with the complete set of features to predict a response variable with a finite number of classes, but create a smaller set of "factors" on which they define a classification rule. PLS sequentially selects "factors" that maximize the correlation between the response (corrected for previously extracted factors)

[9] In the spline based logistic regression, spline functions (piecewise polynomial functions) are fitted to each independent variable before it is entered into the linear form in the logit function. This may increase the efficiency of the method as a classifier, although this has not been definitely shown, but it certainly renders the method even more remote from practical experience, and renders interpretation far harder than traditional linear logistic regression.

variable and the features (also corrected for previously extracted factors). The number of factors thus defined is usually left to the user. Neural networks algorithms for discriminations usually build a simple feed-forward network, in which variables are divided into layers. The input layer contains all the features, or independent variables. The output layer contains all the response variables, and the sandwitched layer contains the unobservable, or latent, variable layer. Arcs connecting variables in different layers describe the general functional structure of the neural networks that optimizes the prediction of the output layer from the input layer by a nonlinear function of weighted linear combinations of input variables. The structure is reminiscent of factor analysis, with the important difference that the latter does not allow non-linear functions. See Goel, Prasher, Patel, Landry, Bonnell and Viau (2003) for a detailed comparison of CART with neural networks in the field of agricultural economics. Markham, Mathieu, and Wray (2000), analyzed a just-in-time kanban production system using CART and neural networks. They found the two methods "comparable in terms of accuracy and response speed, but that CARTs have advantages in terms of explainability and development speed." (There, abstract.)

De'ath and Fabricius (2000) analyzed ecological data of soft coral taxa from the Australian central Great Barrier Reef. They found that for their data, CART dominated its competitors, primarily linear models in their case, because (see, there, page 3178),

> "1) the flexibility to handle a broad range of response type, including numeric, categorical, ratings, and survival data; invariance to monotonic transformations of the explanatory variables; 2) ease and robustness of construction; and 5) ability to handle missing values in both response and explanatory variables. Thus trees complement or represent an alternative to many traditional statistical techniques, including multiple regression, analysis of variance, logistic regression, log-linear models, linear discriminant analysis and survival models."

The circumstances under which CART is particularly recommended are precisely the circumstances that stomp CART's major traditional competitor, logistic regression. The traditional competitors to CART do not in general handle well data sets that include a large number of explanatory variables relative to the number of cases; they also require data homogeneity, i.e., the same relations among the features all over the measurement space. Another compelling reason for adopting CART over traditional model-based classifiers is its intuitive appeal. Most statistics consumers would find nonlinear, generalized regression, such as logistic regression, far less intuitive, and far more indirectly related to their application than CART's classification tree. The latter represents in a simple and accessible tree structure the decision process associated with the classification. Generally the tree involves only a small fraction of the features available in the data, and gives a clear indication of the importance of the various features in predicting the outcome. CART requires no intensive interpretation for understanding the output, as is the case, for example, in logistic regression.

We do not argue, however, that under any circumstances, using CART dominates using one of CART competitors, or a combination of CART and alternative methods. For many data sets CART produces trees that are not stable. A slight change in the learning sample data may alter the structure of the tree substantially, although it will not alter its discrimination ability very much. This property exists in data sets with markedly correlated features. This property is of course shared by other methods, and is well recognized by users of linear or logistic regression. In CART, the problem translates into the existence of several splits at a single node that are almost equivalent in reducing the total diversity of the daughter nodes. The selection of a particular split is then rather arbitrary, but may lead to widely different trees. This instability implies that

users must beware of over-interpreting the location of certain features in the tree produced by CART, despite the temptation to do so (see BFOS). On the other hand, this property implies the availability of different trees of similar discrimination capacity which allows flexibility in the choice of the features used by the tree, an advantage under many circumstances.

CART is not a fully efficient (in the statistical decision sense) alternative to traditional classification methods. CART's occasional reduced relative efficiency stems primarily from its recursive nature, which is also the secret to its transparency and simplicity, and the fact that it does local optimization on single variables at a time. At each node, CART considers all available features, and all possible splits on those features, to choose the best feature and the best split that will create the least internally diverse pair of daughter nodes.[10] This is done with complete disregard for the history of splits carried out in previous tree nodes, leading to the present node. The recursive nature of the CART algorithm then, and its consideration of one feature at a time, instead of working on multiple features at a time, as most other parametric and nonparametric methods do, suggests that CART cannot be as efficient in predicting class affiliation as truly multivariate methods. However, the truly multivariate methods will also tend to be more opaque than the recursive, single-variable at a time CART. It is important to note here, however, that CART does allow the user to select linear combinations of features, precisely to overcome the locally single-variable feature of the method.

When should CART be preferred to traditional methods then? For small data sets CART tends to provide somewhat less accurate classifications, when compared to logistic regression for instance. For most users, however, and certainly in applications

[10] See Section 2, first paragraph.

such as default risk classification, where transparency and ease of use are of paramount importance, a small loss in accuracy is not decisive. In simulation experiments carried out by BFOS, it was shown that in most simulated learning samples CART performed (in terms of true misclassification rate) as well or better that the K-nearest neighbor rule, except for one data set. They also compared CART to a stepwise (in deciding which features to retain in the discriminant function) linear discriminant rule. The latter was found slightly more accurate than CART, but of course its form is less appealing than CART's decision tree rule.

1.3 CART and Traditional Classification Methods in Management Applications

Classification has found various applications in business areas both as a sole tool of analysis and in combination with other analysis tools. Frydman, Altman, and Kao (1985) report on the use of decision trees for financial analysis of firms in distress, and compare it to discriminant analysis. Trostad and Gum (1994) describe the use of CART following a dynamic programming solution to a range cows culling decisions. Finally, CART is used as a data pre-processor, before the data is submitted to systems such as neural networks. Kennedy (1992) discusses the importance of classification in accounting, and examines the performance of seven methods of multiple classification, including classification trees. He stresses that the comparison of classification trees with logistic regression have yielded mixed results. That situation remains true to this day. Simulation results seem to prefer logistic regression, but in real data the differences are minimal, and not all research appears to use robust methods, such as cross validation, to carry the comparisons in real data. In the field of Health Care management, Fu (2003) reports on combining CART with log-linear analysis of birth data, where CART was used to select variables to do the log-linear analysis on. Abu-Hanna and de Keizer (2003) have used CART and compared it to

logistic regression classification in evaluating the efficacy of intensive care models for predicting patients' survival from important indicators assessed at admission to the intensive care unit. Here the authors suggest using CART to split the patient population into subpopulations where a local logistic regression may be used to do better prediction. Faraggi, LeBlanc, and Crowley (2001) report on an interesting use of CART following a neural networks analysis of censored regression data. The output (predictions) from the neural networks was fed into CART, and a classification procedure resulted, despite the incompleteness of the data. For more on the topic of hybrid methods, see Michie, Spiegelhater and Taylor (1994), Kuhnert, Do, and McClure (2000), and Averbook, Fu, Rao, and Mansour (2002).

In Marketing, CART could be useful in analyzing data consisting of price, product information, and consumer information together with brand choice. O'Brien and Durfee (1994) use and compare classification tree software for market segmentation. Haughton and Oulabi (1997), compare CART and CHAID (Chi-Square Automatic Interaction Detector) in analyzing direct marketing data and find them comparable. CART has been extensively used in the fast developing field of Data Mining, and the field of Medical diagnosis. Pomykalski, Truszkowski, and Brown (1999), suggest an approach to developing an expert classification system.

In the finance literature, Hoffman (1990) reports (in German) on the use of tree methodology for credit scoring. Chandy and Duett (1990) use CART, multiple discriminant analysis, and logistic regression to rate commercial paper and report 85% success. Mezrich (1994) uses CART to develop a decision rules for the attractiveness of buy-writes.[11] DeVaney (1994) used CART and logistic regression to examine the usefulness of financial ratios as predictors of household insolvency and Sorensen,

[11] The simultaneous writing of a stock call option and purchase of the underlying stock.

Miller, and Ooi (2000) use CART to select outperforming stocks. In addition, The Salford Systems web site reports on the use of CART software in the financial services industry to retain customers by making preemptive offers to mortgage holders identified as most likely to refinance their homes. Additional practitioners applications are in Gerritsen (1999) and Thearling (2002), and additional references in Komorad (2002).

Kolyshkina and Brookes (2002), use CART to evaluate insurance risks in workers compensation and hospital costs. In the first case they find that CART performs better than Logistic regression and in the second case they use MARS (Multivariate Adaptive Regression Splines) a modification of the CART methodology designed to improve performance where the response is continuous rather than binary or categorical. For more information on MARS, see Friedman (1991).

1.4 Our Application: Mortgage Default in Israel

Our analysis of the mortgage data is of some interest in its own right. Mortgage financing is an essential decision for both borrowers and lenders. Not only is this decision qualitatively important, it is quantitatively significant: aggregate outstanding mortgage balances, and thus the capitalization of various mortgage related securities, is in the trillions.[12] No wonder that the various aspects of mortgage contracting have been one of the most extensively researched topics in real estate finance and economics, both theoretically and empirically. Amongst these aspects, mortgage default has been one of the leading topics. Understanding mortgage default is necessary for appropriately valuing mortgages and for borrowers' and lenders' optimization. Indeed, there is a steady flow of theoretical and empirical studies including new approaches, methodologies, and perspectives in mortgage default

[12] Rough extrapolation of Miles's (1990) several estimates of U.S. real estate value puts today's value at the order of magnitude of 7 trillion dollars.

research and there seems to be a general consensus that more research is needed beyond accounting for the dynamic changes in markets. In this paper, we attempt to contribute to this effort by suggesting a new approach: the use of the CART methodology in analyzing mortgage default.

For related results and references, please see the following very partial sample of recent related works: Foster and Van Order (1984), Clauretie (1990), Kau, Keenan, Muller, and Epperson (1992), Kau and Keenan (1993), Lekkas, Quigley, and Van Order (1993), Vandell (1993), Kau, Keenan, and Kim (1994), Quigley and Van Order (1995), Vandell (1995), Ambrose, Buttimer, and Capone (1997), Deng (1997), Capozza, Kazarian, and Thomson (1997), Capozza, Kazarian, and Thomson (1998), Karolyi and Sanders (1998), Stanton and Wallace (1998), Ambrose and Buttimer (2000), Deng, Quigley, and Van-Order (2000), Ambrose, Capone, and Deng (2001), Sanders (2002), and Ambrose and Sanders (2003).

For reasons that we discuss below, there does not seem to be a previous academic mortgage default study that uses Israeli data. As we also discuss below the data that we received is comprehensive on one hand but suffers from some limitations on the other, and the Israeli market has particular characteristics and nuances. Our choice of CART, by and large, neutralizes the limitations of the data and fits some of the particular characteristics of the Israeli market, see Section 3.

The rest of the paper is organized as follows. In Sections 2 we elaborate on CART structure and methodology. In Section 3 we describe and analyze the Israeli mortgage data as an illustration of the use of CART in a real estate setting, report the results and discuss its conclusions. In Section 4 we present some general discussion and conclusions.

2. Classification Trees: Structure and Method

The CART binary tree consists of a root node, internal nodes and leaf (terminal) nodes. Each root and internal node is a parent node with two daughter nodes. Each node, say t, is described by the subset of the original learning sample that it contains. For all but the leaf nodes, this subset is divided into two groups, going to daughter nodes t_l and t_r. The split at each node is described by a rule that depends on one selected feature. Let this feature be X, and assume, first, that the X is continuous. Then, the split is of the form $X \leq s$ or $X > s$, for some constant s. If X is categorical, then the split is of the form $X \in S$ or $X \notin S$, where S is some nonempty subset of X's possible categories. The feature X is selected among all possible ones, and *s (or S)* is selected among all possible splits, with a view towards minimizing the *diversity* of the resulting subsamples forwarded to the two daughter nodes. Diversity of a subsample, roughly speaking, is a measure of its heterogeneity. We define specific measures of diversity below. As we will see Section 3.1 below, CART offers several splitting methods. We point out at the outset that none of these splitting methods corresponds to an optimal test that controls/optimizes error probabilities in any known way.

Initially, CART produces a large maximal tree and then prunes it into a simpler final tree. Although node splits are selected by maximizing the local reduction in diversity, this procedure also minimizes the overall tree diversity, please see Section 3.2. It does not necessarily, however, minimize the risk or cost of misclassification. CART offers several pruning procedures that we will discuss in Section 4. The choice of a splitting rule and the choice of a pruning procedure are both important for achieving a stable tree yielding as small a risk/cost of misclassification, as is possible for a given data. It turns out that the class assignment problem is relatively simple. The critical choices are those of selecting splits and in determining when to stop splitting.

We now provide a more detailed description of the classifier CART that we use on our mortgage data. We use general terms, and refer the reader to BFOS for more technical details. Our description aims to provide the reader with sufficient understanding of the method to make educated decisions in selecting the CART options that are appropriate for a certain data set. We will then specify the particular options in CART that we applied to our data. In the following section we describe the data and the results.

As we explained in the introduction, the CART algorithm is a recursive procedure; starting at the root node, and then at every internal node, it selects a single feature, and a threshold value *s* to split the group of individuals at the node into two groups to be placed at two new daughter nodes. CART grows the largest tree possible, called a maximal tree, that is the tree whose leaves (terminal nodes) cannot be split any further. A node may not be split any further either because it contains only cases that belong to a single class, or because no reduction in total diversity can be obtained by further splitting.

CART provides three possible splitting methods: *Entropy*, *Gini*, and *Twoing*. Each of these choices may be adopted along with a structure of classification error costs, $C(i \mid j)$, the cost of classifying a case into class *i*, when in fact it belongs to class *j*. CART's user chooses levels of misclassification costs, $C(i \mid j)$, with great flexibility to fit the particular application. Once the tree is complete, CART offers various options for pruning the large tree and reducing it to a tree with far fewer nodes but with a similar discrimination ability.

2.1 Splitting Rules

We first assign a prior probability, p_j, $0 \le p_j \le 1$, to every class j into which cases are classified, $j = 1,...,J,$, with $\sum_{j=1}^{J} p_j = 1$. In case the user does not provide prior probabilities, the relative frequencies of the classes in the learning sample are used as prior probabilities. In order to create a tree one needs to specify:

1. A criterion of diversity.
2. A goodness of split criterion function at node *t,* for feature *X*, and threshold split value *s*, $\Delta d(s,t)$, that determines how good the split is in reducing diversity of the two daughter nodes for feature *X*.
3. A splitting rule.
4. A "stop splitting" rule.
5. A rule for assigning a terminal node (a leaf) into one of the *J* classes.
6. A misclassification cost structure for evaluating the resulting tree performance.

The splitting rules are of the form $X \le s$ or $X > s$, for some constant s when the feature X is quantitative or at least ordinal. When X is qualitative with L categories, CART tries all possible distinct binary splits, $2^{L-1} - 1$ in number[13]. At each node of the tree the program searches through the features one by one, determines the best split for each X, and then the best X to split on at that node. Each split causes the resulting groups into which the data is split to be more homogeneous (less diverse) than the parent group.

A splitting rule is derived from a diversity function (called impurity function by BFOS). Let the cost, $C(i \mid j)$, of misclassifying a case that belongs to class j into class i,

[13] There are 2^L total combinations, when order does not matter and excluding the "all-nothing" split we have $2^{L-1} - 1$.

obey $C(i \mid j) \geq 0$ and $C(i \mid i) = 0$, and let $p(j \mid t), \quad 0 \leq p(j \mid t) \leq 1$, $j = 1,...,J$, be the proportion of class *j* cases present at node *t* of the tree. *J* denotes the number of classes. Thus, for each node *t*, $\sum_{j=1}^{J} p(j \mid t) = 1$.

We shall now present the three major diversity functions that CART uses at some node *t*. We shall distinguish between two different cases. In the first case, the cost of misclassification of any item, regardless of its actual class, and regardless into which class it was misclassified, is uniform. In the second case, the cost of misclassifying a case belonging to class *j* into class *i*, denoted by $C(i|j)$, may depend both on *i* and on *j*.

1. The *Entropy function* under uniform costs is

$$d_E(t) = -\sum_{j=1}^{J} p(j|t) \log[p(j|t)], \tag{1}$$

and is, under non-uniform costs

$$d_E(t) = -\sum_{j=1}^{J} \sum_{i=1, i \neq j}^{J} C(i|j) p(j|t) \log[p(j|t)], \tag{2}$$

where *i* stands for the class into which the case is classified and *j* stands for its true class.

2. The *Gini index of diversity* under uniform costs is

$$d_G(t) = \sum_{j=1}^{J} \sum_{i=1}^{J-1} p(i \mid t) p(j \mid t) = \frac{1}{2}\left(1 - \sum_{j=1}^{J} p^2(i \mid t)\right), \tag{3}$$

which, in the binary case, simplifies to

$$d_G(t) = p(1 \mid t) p(2 \mid t), \tag{4}$$

and is, under non-uniform costs

$$d_G(t) = \sum_{j=1}^{J} \sum_{i=1}^{J-1} p(j|t) p(i|t) [C(i|j) + C(j|i)]. \tag{5}$$

3. The *twoing function*, with daughter nodes t_L and t_R, and where the probabilities p_L and p_R are the proportions of cases going to nodes t_L and t_R respectively, is

$$d_T(t) = \frac{p_L p_R}{4} \sum_{j=1}^{J} \left| p(j|t_L) - p(j|t_R) \right|. \quad (6)$$

We remark that the Entropy and the Gini index diversity functions refer to the diversity of cases at a given node. Therefore as a tool for splitting cases at a node, a change in diversity from that of the parent node, to the sum of diversity at the daughter nodes is required. The twoing function, on the other hand, measures a class-prevalence distance between the daughter nodes, anticipating that the diversity within the daughter nodes will decline when the split achieves a higher degree of difference in the prevalence of the different classes in the two daughter nodes. Thus, to achieve the highest reduction in diversity, one chooses the split *s* that maximizes the towing function.

Note that both the Entropy function and the Gini index achieve their maximum value at node *t* when the distribution of cases to classes is uniform. Both achieve their minimum, zero, when all cases at the node fall into a single class. In contrast, the twoing function which measures the heterogeneity between the daughter nodes, achieves its minimum when the daughter nodes contain exactly the same distribution of classes, and its maximum when all cases belonging to a given class are found in one node. Thus if there are two nodes, all cases of class 1 belong to one node, and of class 2, to the other node.

Once the Gini or Entropy diversity functions is chosen, a splitting rule, that is a splitting value *s** is adopted at node *t* that maximizes the reduction in diversity obtained by the split. Using the notation just developed, we define the gain in (reducing)

diversity reduction obtained by splitting node t into two nodes, L and R using the threshold s, for some feature, as

$$\Delta d(s,t) = d(t) - p_L d(t_L) - p_R d(t_R), \tag{7}$$

where p_L and p_R are the proportions of cases going to nodes t_L and t_R respectively. This gain in diversity reduction is also referred to as the goodness of the split s for node t. Splitting is continued as long as the goodness of the best split at t is positive. We reemphasize that this procedure applies to the Gini index and Entropy functions only.

2.2 Selecting and Pruning a Tree

Suppose that a tree T has been generated with terminal nodes T^t, we then define the tree diversity as

$$D(T) = \sum_{t \in T^t} d(s,t). \tag{8}$$

As was pointed out by BFOS, although we select a tree by choosing the best splitting feature, and the best split for that feature at each node, the resulting tree is also the tree that minimizes the diversity $D(T)$. It is not necessarily the best tree from the point of view of misclassification.

The goodness of the tree as a classification instrument may be characterized in terms of its estimated misclassification rate. When misclassification costs are not uniform, a reasonable definition of the (generalized) expected misclassification cost is

$$R(T) = \sum_{j=1}^{J} \sum_{i=1, i \neq j}^{J} C(i|j) Q(i|j) \pi(j), \tag{9}$$

where $Q(i|j)$ denotes the proportion of class j cases misclassified into class i, and $\pi(j)$ is the prior probability of a case being in class j.

Of course these estimated misclassification rates are highly underestimated, because they depend on the data that produced the classification rules to begin with.

Two better methods of estimating misclassification costs are available in CART: The Cross-Validation method, and the Test-Sample method. In the former, the learning sample is randomly split into K equal size subsamples. K is usually set to be ten, but may be changed for very small or very large data sets. A CART tree is produced K times, each time from a different group of K-1 (usually 9) subsamples. The rule is used to classify the cases in the tenth subsample left out in the tree construction, and the resulting misclassification rates are noted. The K (usually 10) misclassification rates thus obtained are then averaged to obtain the Cross-Validation misclassification rates $Q^{CV}(i|j)$. These are then plugged into the $R(T)$ formula above to obtain the overall Cross-Validation misclassification rate $R^{CV}(T)$ that takes into account prior probabilities and non-uniform misclassification costs.

When the data set is sufficiently large we do not have to resort to Cross-Validation to produce a misclassification rate estimate that is not severely downward biased. In that case we simply take a single random test subsample from the learning sample and take the misclassification rates of the cases not included in the Test-Sample as our estimates of $Q(i|j)$. The resulting overall misclassification rate estimate is denoted by $R^{TS}(T)$.

BFOS proceed to estimate the standard errors (SE) of $R^{CV}(T)$ and of $R^{TS}(T)$. Here standard errors refer to the distribution of $R^{CV}(T)$ and of $R^{TS}(T)$ produced by the random selection of subsamples in both the Test-Sample case and in Cross-Validation. The purpose of these SE estimates is to be used in pruning the maximal trees. A maximal tree is initially produced by splitting nodes until they are pure in the sense that each terminal node contains only cases that belong to a single class, or nodes whose diversity cannot be reduced by further splitting.

It turns out that in trying to select a subtree of the maximal tree that minimizes the estimated misclassification cost, a large number of subtrees will yield approximately the same estimated misclassification cost. It is then reasonable to stop the search for the best pruned tree once a subtree is found that is within one SE of the minimum estimated misclassification cost subtree. This is called in CART the 1 SERULE. Once the subtree is selected, that is pruning is completed, CART uses another Cross-Validation to estimate the expected misclassification error of the pruned tree. In simulation experiments carried out by BFOS the final R^{TS} came within one SE of R^{CV}.

It is evident that using different diversity measures, different misclassification cost structures, Cross-Validation versus Test-Sample, and various levels for SERULE (0 or 1), generally, various classification trees are obtained. Criteria for selecting the 'best' tree are then required. One criterion is the cost-complexity of a tree.

The cost-complexity of a tree is defined by

$$R_{\alpha}(\mathrm{T}) = R(\mathrm{T}) + \alpha\left|\mathrm{T}^{\mathrm{t}}\right|, \tag{10}$$

where α is a complexity coefficient, $0<\alpha$, and $\left|\mathrm{T}^{\mathrm{t}}\right|$ is the number of terminal nodes of the tree. Because the estimated misclassification rate tends to decrease as the number of terminal nodes of a tree increases, the proposed cost-complexity measure penalizes a tree for the proliferation of its terminal nodes; the complexity parameter α may be thought of as complexity per node. This cost-complexity may then be used to compare the small number of trees obtained via the carefully selected methods described above.

Another useful comparison of classification trees in the binary case uses the concepts of sensitivity and specificity, commonly used in statistical test evaluation. In binary classification, we identify as "bad" the category that we most want to identify. In our example, that category would be the more likely-to-default category. The other category will be referred to as "good." Sensitivity and specificity now split the overall

correct classification rate into its essential components. Sensitivity of the tree is the (estimated) probability that a new "bad" case will be classified as "bad" when processed by the tree. Specificity (of the tree) is the (estimated) probability that a new "good" case will be identified as "good" by the tree.

This completes our concise description of the main components of CART. For a more accurate and detailed description of the method please see BFOS or Hastie, Tibshirani, and Friedman (2001). See also Bloch, Olshen, and Walker (2002) work on misclassification estimation, which contains some illuminating general comments on CART. We also recommend the latter for further references.

3 Data Analysis with CART

Our data consists of end of the year 1998 information regarding fixed rate residential mortgage contracts that were issued during the years 1993 through 1997 by a major Israeli mortgage bank. The bank contracted the consulting firm GStat Ltd. to analyze these data, providing them with some electronic but mainly paper files of several dozens of thousands mortgage contracts. About 1500 of these contracts were delinquent during the period. Out of the of non-delinquent mortgages, GStat Ltd. chose about 1500 mortgage contracts at random. This defined a set of 3,035 mortgage contracts. GStat Ltd., keyed in a subset of mortgage and borrowers and features from the bank's paper files, and merged it with electronic bank data and created the data base. Following a suggestion from the bank, GStat Ltd. gave us all these data records excluding in each record some identifying features such as names, addresses, etc. complying with banking privacy legal acts.

Our study seems to be the first Israeli academic mortgage default study. The surprising absence of previous studies stems probably from lack of mortgage default data, which, in turn, is probably a consequence of the non-competitive nature of the

Israeli banking industry in general, and mortgage banking in particular. The two largest Israeli banks control about 80% of the Israeli banking retail market. The data that we received suffers however from some important limitations. For example, although a single mortgage contract could have several delinquencies (being late in paying for at least ninety days), no information on the time, size, and number of these contract delinquencies was available in our data. For that reason, delinquency became a binary attribute, with no time dimension. In addition, because of the monopolistic nature of the Israeli banking market no credit histories are available. In fact, the major banks opposed the establishment of a national credit history data base.

There were no prepayments in our data. This is a consequence of the Bank of Israel regulation that allows the banks to charge borrowers a prepayment fee which is equal to the economic benefit of refinancing the unpaid principal under the prevailing rate of interest.[14] This fee sets the value of prepayment and refinancing to zero. Adding to this fee stamp duty on the new loan which is about half a percent of the principal, and fixed bank fees for "opening a new loan file," the value of prepayment and refinancing to the borrower becomes negative. Thus, in Israel, the option to prepay a fixed rate loan is usually worthless. The Bank of Israel regulation actually constitutes a ceiling on the fee but this ceiling is the realized market fee, reflecting, probably, the (low) level of competition. Unlike banks, insurance companies when engaged in mortgage loans are not bound by Bank of Israel regulations. In their case, however, these regulations would not have been binding. They many times offer loans with no prepayment fees whatsoever, but their market share is insignificant.

There were no foreclosures in our data. This probably is a consequence of three factors. First, the relatively low LTV ratio of Israeli mortgage loans. Second, the fact

[14] See the Appendix of the Bank of Israel Banks Supervisor Circular No. 1673-06-H pp. 87-92.

that borrowers are responsible for their loans, thus banks can enforce the use of borrowers' non mortgaged property and various sources of income for paying the mortgage loan. Third, the common requirement, particularly for the higher LTV loans, that guarantors, in additional to the borrowers, sign the mortgage loan.

Despite its limitations, the data provided a very good example of the use of the CART methodology, as well as a first, albeit limited, analysis of the Israeli mortgage market. We note that Israeli banks require mortgage borrowers to have property and life insurance to cover mortgage liability, and Israeli law now, calls a mortgage delinquent only if delinquency lasted at least ninety days.

We first ran a descriptive analysis of the features: means, univariate analyses, and frequencies. Then, we checked correlations to assess the pair wise associations among the features. We also examined the relationships between the dependent variable and each of the independent variables using t-tests, or nonparametric tests. These did not raise any particular issue with any of the features. We then ran the CART analysis using the CART program that Salford Systems (www.Salford.com) distributes.

The thirty three features were:

<u>Features related to the mortgage size and type</u>

CSUM – mortgage total size

CROOMS – number of rooms in the property

MONTHRET – monthly payment

GRANT_PR – % of the property value given to borrower as a grant

RETINC_P – % of monthly payment from monthly income

VALNECSN – present value of property

YTR_HA – balance of the mortgage

YTR_HA_O – balance of the government supplementary mortgage

YIT_SILK – balance of the mortgage including late fees and penalties

VAL_NECS – original value of the property

SHETACH – area of the property

SIL_MUKD – 1 if mortgage prepaid, 0 otherwise

NGUARANT – number of guarantors

PERIOD – term to maturity of mortgage

CDESIG – designation of property

1 – living quarters

2 – apartment to rent

3 – property for business use

CTARGET1 – purpose of mortgage

1 – buy an apartment

2 – buy an apartment second-hand

3 – build own apartment

4 – other real estate purpose

5 – renovation purpose

6 – refinancing mortgage

7 – not for living or remodeling

8 – other

Features describing the borrower(s)

CLOANERS – number of borrowers on the mortgage

FCHILD – number of children of first borrower

FINCOME – monthly income of first borrower

NETINCOM – monthly net income

AGE1 – age of first borrower

CSPOUSE - 1 if first borrower is married, 0 otherwise

EDUC1 – education of first borrower

1 – elementary

2 – high school

3 – some college

4 – college degree

5 – other

FCODE2 – first borrower's occupation

1 – teacher

2 – driver

3 – engineer

4 – academic: social sciences

5 - practical engineer

6 – professional (worker) laborer

7 – unprofessional laborer

8 – salesman

9 – clerical worker

10 – clerical/religious student

11 – agricultural worker

12 – pilot

13 – medical doctor

14 – paramedical worker

15 – sales worker

16 – policeman

17 – army personnel

18 – care giver

19 – businessman

FEXP – first borrower's work experience

FDUTY – first borrower job's managerial capacity

1 – top manager

2 – manager

3 – not a manager

FFAMCON – first borrower’s marital status

1 – married

2 – divorced

3 – widow/widower

FSTABLE – first borrower job permanence

1 – permanent worker

2 – not permanent

3 – other

FSTATUS – first borrower job status

1 – employee

2 – self-employed

3 – both 1 and 2

4 – student

5 – Yeshiva student

6 – house-person (housewife)

7 – retired

8 – on (public assistance) some assistance

9 – receives alimony

10 – unemployed

11 – not working

12 – other

RUSSIA – borrower from Russia?

1 – yes

2 – no

ETHIOPIA – borrower from Ethiopia?

1 – yes

2 – no

FINC_CHI – first borrower monthly income divided by number of children

FSUM_CHI – first borrower mortgage size divided by number of children

The original data included variables associated with the second borrower. Because these contained much missing data and because we could not tell whether there was a second borrower in these cases, we decided to eliminate them from the analysis. We believe that this elimination has no systematic implications. Also, the last two variables were added on the suspicion that they may turn out to be more predictive of default than FINCOM and NETINCOM, respectively. Generally speaking, our data includes features related to property values, loan values, payments, income, and demographics that are commonly used in mortgage default studies.

We ran CART on the n=3,035 borrowers data using different options for creating and pruning the final trees. Our aim was to classify these borrowers into good: non-defaulters, and bad: defaulting borrowers.

We ran CART five times creating five trees, each under different option combinations, as follows.

<u>First option combination</u>

Misclassification costs: uniform

Splitting criterion: Gini index

Misclassification estimation: Cross-Validation

Pruning criterion: SERULE=0 (search for 'best' subtree with minimum estimated weighted misclassification rate)

Second option combination

All options remain as in 1, save for SERULE=1 (search for subtree that is within 1 SE of the 'best' subtree). This change was expected to lead to a tree that shares many of the qualities of the tree obtained under 1, but is less expensive to obtain and implement.

Third option combination

All options remain as in 1, except that the following non-uniform misclassification costs were used:

C(classify as bad | borrower is good) = 1

C(classify as good | borrower is bad) = 1.5

Here the cost of misclassifying a bad borrower as a good risk is considered 1.5 times more costly than the reverse. With this misclassification cost structure, the same tree was obtained with pruning using SERULE=1.

Fourth option combination

All options remain as in 1, save for Cross-Validation being replaced by Test-Sample. With a large sample, such as we have, it was deemed possible to replace the more costly Cross-Validation misclassification estimation by the Test-Sample method.

Fifth option combination

All options remain as in 4 (Test-Sample method), but with cost structure as in 3 (non-uniform cost structure) and pruning using SERULE=1. The tree obtained using these specifications with SERULE=0 was too unwieldy (36 terminal nodes, or leaves) and was dropped.

Several points are raised by the results displayed in Table 1.

- Trees possessing high sensitivity relative to specificity are obtained when the misclassification cost of a 'bad' borrower into a 'good' one is taken to be higher than the reverse misclassification. Trees 3 and 5 display this characteristic.
- The smallest tree, tree number 3, also possesses the smallest overall (penalized) cost complexity. It possesses remarkably high sensitivity, as measured by Cross-Validation, and relatively low specificity. In risk-control application, such as ours, this ratio of sensitivity to specificity may be desirable.
- If a more balanced treatment of the two possible misclassification: 'bad' to 'good' and 'good' to 'bad' is desired, then tree number 2, which has a slightly higher overall cost-complexity, may be the proper choice.
- The estimated cost-complexity, sensitivity and specificity of the fourth tree were obtained via a random sample of borrowers, rather than by the more robust Cross-Validation method. Since it does not have any particular feature to recommend it over trees 3 and 2, we did not attempt to estimate its cost-complexity, sensitivity and specificity using Cross-Validation.
- CART's analysis is, of course, blind to political concerns. Classification Trees, thus, might be "politically incorrect," and, therefore, hard to

implement. It is likely, however, that politically correct trees with similar properties exist.

- Regarding features that have surfaced as predictive in many of the trees:
    1. Most of the primary features are associated with the borrower and not with mortgage attributes.
    2. EDUC1 (some college versus no college) appears as the first splitting variable in all five trees.
    3. If we select the most parsimonious tree 3, only borrower characteristics really matter, and the second feature is FDUTY (manager or top manager, versus non-manager). Surprisingly, managers (with some college education) are classified as bad risk, as are borrowers with no college education. FDUTY appears as a significant splitting variable in tree 1. In trees 2, 4, and 5 it appears to be replaced by other work features associated with it: FSTATUS, borrower job status, and FCODE2, borrower's occupation.
    4. The period of the mortgage appears as the second splitting feature in all trees that use uniform costs. It seems that non-uniform costs, such as those used for trees 3 force borrower features in, and mortgage features out. Re: trees 1 and 2. In this risk identification application, this may be very desirable. This is not quite the case with tree 5, but the use of Test-Sample there, makes all cost evaluations and variable choices somewhat suspect.
    5. Important borrower features appear to be: education, status at work: FSTATUS, FCODE2 or FDUTY, # of children (FCHILD) or income per child: FINC_CHI. Finally, AGE1 appears in trees 1 and 2.

6. One has to be careful in interpreting our results because our paper does not allow for a changing environment. If the real-world equilibrium is dynamic, the sample will capture dynamic effects as well as endemic cross-sectional attributes during the sample period. Examining the sample period, we could not think of events that could be considered "regime switching" during the sample period. Neither could we think of events that would have changed the nature of the Israeli real estate market. In addition, the atemporal nature of the data makes it less than ideal to evaluate conditional dependency of default. However, judging our conclusions *ex post*, none of our findings seem especially sensitive to dynamic effects.
7. Our data looks at the status and history of many contracts at a certain date. Thus, one has to be concerned with truncation consequences. If the probability distributions are iid or even if the population is in steady state with respect to the measured attributes, then we should not have a truncation bias problem. Moreover, although a measure of contract age might have helped reduce (not eliminate) a possible truncation bias, this is not a relevant issue here because of special characteristics of the Israeli real estate market and of our data set. Israeli lenders tend to avoid foreclosures at all costs. Thus, guarantors are co-signed on the each mortgage contract. In case of delinquency, the bank captures the owed value from the guarantors. Consequently, none of the roughly 1500 delinquent properties in our sample were repossessed, and delinquency is therefore an ageless, binary attribute in our data.

8. Based on this study, we would recommend tree 2 or 3 for classification of future borrower into 'good' or 'bad' risks. Tree 3 is more conservative, but seems so parsimonious that its intended users of the procedure may shy away from it.
9. It is interesting to study the two candidate classification trees 2 and 3. Briefly, classification via tree 2 prescribes the following rule sequence:
    i. If applicant has at least some college education, stop and rate him a good risk.
    ii. Otherwise, if the period of the mortgage is over 27.5 years, stop and declare the applicant a bad risk.
    iii. If the applicant has at most high school education (or other for EDUC1), and the mortgage period is under 27.5 years, then if the applicant is either a student, or a housewife, or self employed (or other for variable FSTATUS) then stop and declare the applicant a bad risk.
    iv. Otherwise (to iiii) check applicant's job classification FCODE2. If applicant is employed by the army, Yeshiva student, care taker, or a para-medical worker, then stop and declare the applicant a good risk.
    v. Otherwise (to iv), if he has three or more children (FCHILD), stop and declare him a bad risk.
    vi. If he has 2 or fewer children, and is under 32.6 years of age (AGE1), then stop and declare him a bad risk. If he is over 32.6 years of age with 2 or fewer children, declare him a good risk.

    For tree 3 the decision process proceeds as follows:

i. If the applicant has at most a high school education, (or is other for EDUC1), stop and rate him a bad risk.

ii. Otherwise, check his employment type FDUTY. If he is a manager or a senior manager, stop and declare him a bad risk. Otherwise stop and rate him a good risk.

10. Tree 3 described above is rather surprising: a manager or senior manager with at least some academic education is considered a bad risk, but a non-manager with the same educational level is considered a good risk. However, as might be expected, an applicant with at most a high school education is considered a bad risk. An explanation of the classification of senior and regular managers as bad risks and of non-managers as good ones is consistent with higher rate of ruthless default of the former. This, in turn, is consistent with lower reputation default costs of the managers vis-à-vis non-managers. Non-managers might find ruthless default too costly in the long run.

11. Tree 2 seems to conform to expectations, except possibly for some results such as: A business person, a policeman, or a professional electrician, without college education, with a mortgage for under 27.5 years, and at least 3 children is considered a bad risk, but an academic with any number of children, and any length mortgage is considered a good risk.

12. The decision processes described in points 9 and 10 are clearly attractive for direct application in a bank or lending institution setting. It is also clear that decision tree 3, because of its limited use of both mortgage and applicant characteristics, may not find many takers. Tree 2 on the other

hand contains fewer surprising choices, and is far more likely to be chosen.

We remark that the CART analysis we have performed directly on the data without any pre-analysis that might narrow down the field of potential predictors of good risk customers, may now itself be used as input to other classifiers. For example, logistic regression that would be stomped by the number of features in the data, by the huge number of categories in some of the nominal categorical predictors, and by the large number of missing values, can now be attempted using predictors that have been identified as useful by CART. This post-processing by another classifier could potentially improve somewhat the accuracy of the CART classifier. Here we mention also the post-processing proposed by Freund and Schapire (1997) called boosting, and bagging proposed by Breiman (1996); both procedures enhance the accuracy of the CART classifier.

Finally, we would like to comment about prepayment in relation to default. As we explained above, absent idiosyncratic reasons, the option to prepay in our sample is worthless. Thus, we can safely say that in our data prepayment is not a substitute for default. We cannot say the opposite, however. Actually, the higher rate of default of managers versus non-managers of the same education level suggests that default may sometime substitute prepayment.

4 Conclusion

We have provided a concise introduction of CART, its main features, and guidelines for its implementation as a classification tool. We have also applied the method to mortgage default data from a major Israeli bank. Our data had special features, most of which are intimately connected to the nature of the rules governing the Israeli mortgage market. Valuable information was gleaned from the data using

CART with various option choices. We emphasized the process of selecting a final classification tree, which depends both on the CART method, and the particular subject matter at hand. We consider this work preliminary, and hope to receive more complete data in the future that will enable us to refine our findings and perform a comparative analysis of parametric and nonparametric methods.

If the cost of accepting bad risks exceeds that of rejecting good ones CART uses borrowers' features only. If the cost of accepting bad risks equals that of rejecting good ones, CART uses mortgage features such as term and property value as well. The higher (lower) the ratio of misclassification costs of bad risks versus good ones, the lower (higher) are the resulting misclassification rates of bad risks and the higher (lower) are the misclassification rates of good ones. This is consistent with real world stylized facts of rejection of good risks in attempt to avoid bad ones.

The classification process allows the examination of hypotheses. For example, Tree 3 is consistent with higher rate of ruthless default by senior and regular managers vis-à-vis non-managers. This is consistent, for example, with lower reputation penalties of default for managers. Moreover, as we elaborated earlier, CART generates many trees that are of similar quality, on the one hand, but that use different features and splits on the other. Thus, one could examine those trees and determine whether they negate various insights/hypotheses or are consistent with them.

References

Abu-Hanna, A. and de Keizer, N., 2003, "Integrating Classification Trees with Local Logistic Regression in Intensive care Prognosis," *Artificial Intelligence in Medicine*, forthcoming.

Ambrose, B. W. and Buttimer, R. J. Jr., 2000, "Embedded Options in the Mortgage Contract," *The Journal of Real Estate Finance and Economics,* 21, 95-111.

Ambrose, B. W., Buttimer, R. J., Jr. and Capone, C. A., Jr., 1997, "Pricing Mortgage Default and Foreclosure Delay," *Journal of Money, Credit, and Banking*, 29, 314-325.

Ambrose, B. W., Capone, C. A., Jr. and Deng, Y., 2001, "Optimal Put Exercise: An Empirical Examination of Conditions for Mortgage Foreclosure," *Journal of Real Estate Finance and Economics*, 23, 213-234.

Ambrose, B. W. and Sanders, A. B., 2003, "Commercial Mortgage Backed Securities: Prepayment and Default," *Journal of Real Estate Finance and Economics,*26, 175-192.

Averbook, B. J., Fu, P., Rao J. S. and Mansour, E. G., 2002, "A long-term analysis of 1018 patients with melanoma by classic Cox regression and tree-structured survival analysis at a major referral center: Implications on the future of cancer staging," *Surgery*, 132, 589-604.

Breault, J. L., Goodall, C. R. and Fos, P. J., 2002, "Data Mining a Diabetic Data Warehouse," *Artificial Intelligence in Medicine*, 26, 37-54.

Breiman, L., 1996, "Bagging Predictors," *Machine Learning*, 24, 123-140.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J., 1998, *Classification and Regression Trees*, Chapman and Hall / CRC, New York.

Capozza, D. R., Kazarian, D. and Thomson, T. A., 1997, "Mortgage Default in Local Markets," *Real Estate Economics*, 25, 631-655.

Capozza, D. R., Kazarian, D. and Thomson, T. A., 1998, "The Conditional Probability of Mortgage Default," *Real Estate Economics*, 26, 359-390.

Chandy, P.R. and Duett, E. H., 1990, "Commercial Paper Rating Models," *Quarterly Journal of Business and Economics,* 1990, 29, 79-101.

Clauretie, T., 1990, "A Note on Mortgage Risk: Default vs. Loss Rates," *AREUEA Journal*, 18, 202-206.

Daniel A. L., Olshen R. A. and Walker, M. G, 2002, "Risk Estimation for Classification Trees," *Journal of Computational and Graphical Statistics*, 11, 263-288.

De'ath, G. and Fabricius, K. E., 2000, "Classification and Regression Trees: A Powerful yet Simple Technique for Ecological Data Analysis," *Ecology*, 81, 3178-3192.

Deng, Y., 1997, "Mortgage Termination: An Empirical Hazard Model with a Stochastic Term Structure," *Journal of Real Estate Finance and Economics*, 14, 309-331.

Deng, Y., Quigley, J. M. and Van Order, R., 2000, "Mortgage Terminations, Heterogeneity and the Exercise of Mortgage Options," *Econometrica*, 68, 275-307.

DeVaney, S., 1994, "The usefulness of financial ratios as predictors of household insolvency: Two perspectives," *Financial Counseling and Planning,* 5, 15-24.

Fix, E. and Hodges, J., 1951, "Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties," Technical Report, Randolph Field Texas, USAF School of Aviation Medicine.

Faraggi, D., LeBlanc, M. and Crowly, J., 2001, "Understanding Neural Networks Using Regression Trees: an Application to Multiple Myeloma Survival Data," *Statistics in Medicine*, 20, 2965-2975.

Foster, C. and Van Order, R., 1984, "An Option-Based Model of Mortgage Default," *Housing Finance Review*, 3, 351-372.

Freund, Y. and Schapire, R. E., 1997, "A Decision-Theoretic Generalization of On-Line Learning and an Application to boosting", Journal of Computer and System Sciences, 55, 119-139.

Friedman, J. H., 1991, "Multivariate Adaptive Regression Splines," *Annals of Statistics*, 19, 1-141.

Frydman, H., Altman, E. I. and Kao, D. L., 1985, "Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress," *The Journal of Finance*, 40, 269-292.

Fu, C. Y., 2003, "Combining Loglinear Models with Regression Tree (CART): an Application to Birth Data," *Computational Statistics and Data Analysis*, forthcoming.

Gerritsen, R., 1999, "Assessing Loan Risks: A Data Mining Case Study," Exclusive Ore, Pennsylvania.

Goel, P.K., Prasher, S. O., Patel, R. M., Landry, J. M., Bonnell, R. B. and Viau, A. A., 2003, "Classification of Hyperspectral Data by Decision Trees and Artificial Neural Networks to Identify Weed Stress and Nitrogen Status of Corn," *Computers and Electronics in Agriculture*, 39, 67-93.

Hastie, T., Tibshirani, R. and Friedman, J. H., 2001, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer Verlag, New York.

Haughton, D. and Oulabi, S., 1997, "Direct Marketing Modeling with CART and CHAID," *Journal of Interactive Marketing*, 11, 42-52.

Hoffman, H. J., 1990, "Die Unwendung des CART-Verfahrens zur Statistichen Bonitatanalyse von Konsumentenkrediten" *ZeitSchrft-fur-Betriebswirtschaft*, 60, 941-62.

Karolyi, A. and Sanders, A. B., 1998, "The Variation of Economic Risk Premiums in Real Estate Returns," *Journal of Real Estate Finance and Economics,* 17, 245-262.

Kau, J. B. and Keenan, D. C., 1993, "Transaction Costs, Suboptimal Termination, and Default Probabilities for Mortgages," *AREUEA Journal*, 21, 247-63.

Kau, J. B., Keenan, D. C. and Kim, T., 1994, "Default Probabilities for Mortgages," *Journal of Urban Economics*, 35, 278-296.

Kau, J. B., Keenan, D. C., Muller, W. J., III and Epperson, J. F., 1992, "A Generalized Valuation Model for Fixed-Rate Residential Mortgages," *Journal of Money, Credit, and Banking*, 24, 279-99.

Kolyshkina, I. and Brookes, R., 2002, "Data Mining Approaches to Modeling Insurance Risk," Report, PriceWaterhouseCoopers.

Komor´ad, K., 2002, "On Credit Scoring Estimation," Master's Thesis, Institute for Statistics and Econometrics, Humboldt University, Berlin.

Kuhnert, P. M., Do, K. A. and McClure, R., 2000, "Combining Non-Parametric Models with Logistic Regression: an Application to Motor Vehicle Injury Data," *Computational Statistics and Data Analysis*, 34, 371-386.

Lekkas, V., Quigley, J. M. and Van Order, R., 1993, "Loan Loss Severity and Optimal Mortgage Default," *Journal of the American Real Estate and Urban Economics Association*, 21, 353-371.

Mezrich, J, J., 1994, "When is a Tree a Hedge?" *Financial Analysts Journal*, 50, 75-81.

Michie, D, Spieglehalter, D. J. and Taylor, C. C., Editors, 1994, *Machine learning, Neural and statistical Classification*, Ellis Horwood Ltd., London.

Miles, M., 1990, "What is The Value of U.S. Real Estate?" *Real Estate Review,* 20, 69-75.

Moisen, G. G. and Frescino, T. S., 2002, "Comparing Five Mmodelling Techniques for Predicting Forest Characteristics," *Ecological Modelling*, 30, 209-225.

O'Brien, T. V. and Durfee, P. E., 1994, "Classification Tree Software," *Marketing Research*, 6, 36-39.

Pomykalski, J. J., Truszkowski, W. F. and Brown, D. E., 1999, "Expert Systems," in *Wiley Encyclopedia for Electrical and Electronics Engineering*, Webster, J., Editor, John Wiley & Sons, Inc., New York.

Quigley, J. M. and Van Order, R., 1995, "Explicit Tests of Contingent Claims Models of Mortgage Default," *The Journal of Real Estate Finance and Economics*, 11, 99-117.

Rousu, J., Flander, L., Suutarinen, M., Autio, K., Kontkanen, P. and Rantanen, A., 2003, "Novel Computational Tools in Bakery Process Data Analysis: a Comparative Study," *Journal of Food Engineering*, 57, 45-56.

Sanders, A. B., 2002, "Government Sponsored Agencies: Do the Benefits Outweigh the Costs?" *Journal of Real Estate Finance and Economics,* 25, 121-127.

Sorensen, E. H., Miller, K. L. and Ooi, C. K., 2000, "The Decision Tree Approach to Stock Selection," *Journal of Portfolio Management*, 27, 42-52.

Stanton, R. and Wallace, N., 1998, "Mortgage Choice: What is the Point?" *Real Estate Economics,* 26, 173-205.

Thearling, K., 2002, "Scoring Your Customers," http://www.thearling.com.

Tronstad, R. and Gum, R., 1994, "Cow Culling Decisions Aadapted for Management with CART," *American Journal of Agricultural Economics*, 76, 237-249.

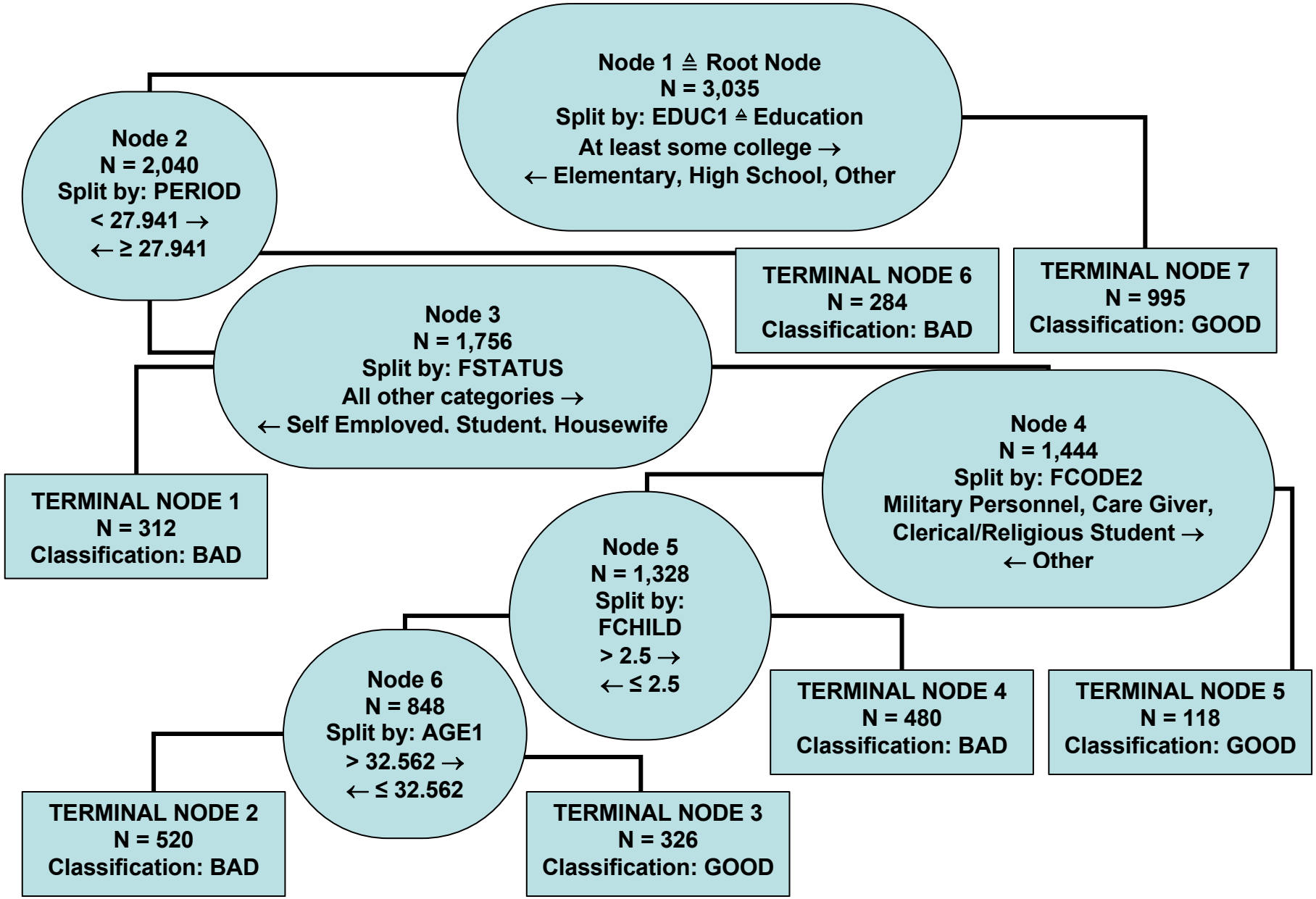
Vandell, K. D., 1993, "Handing Over the Keys: A Perspective on Mortgage Default Research," *Journal of the American Real Estate and Urban Economics Association*, 21, 211-246.

Vandell, K., 1995, "How Ruthless is Mortgage Default?" *Journal of Housing Research*, 6, 245-264.

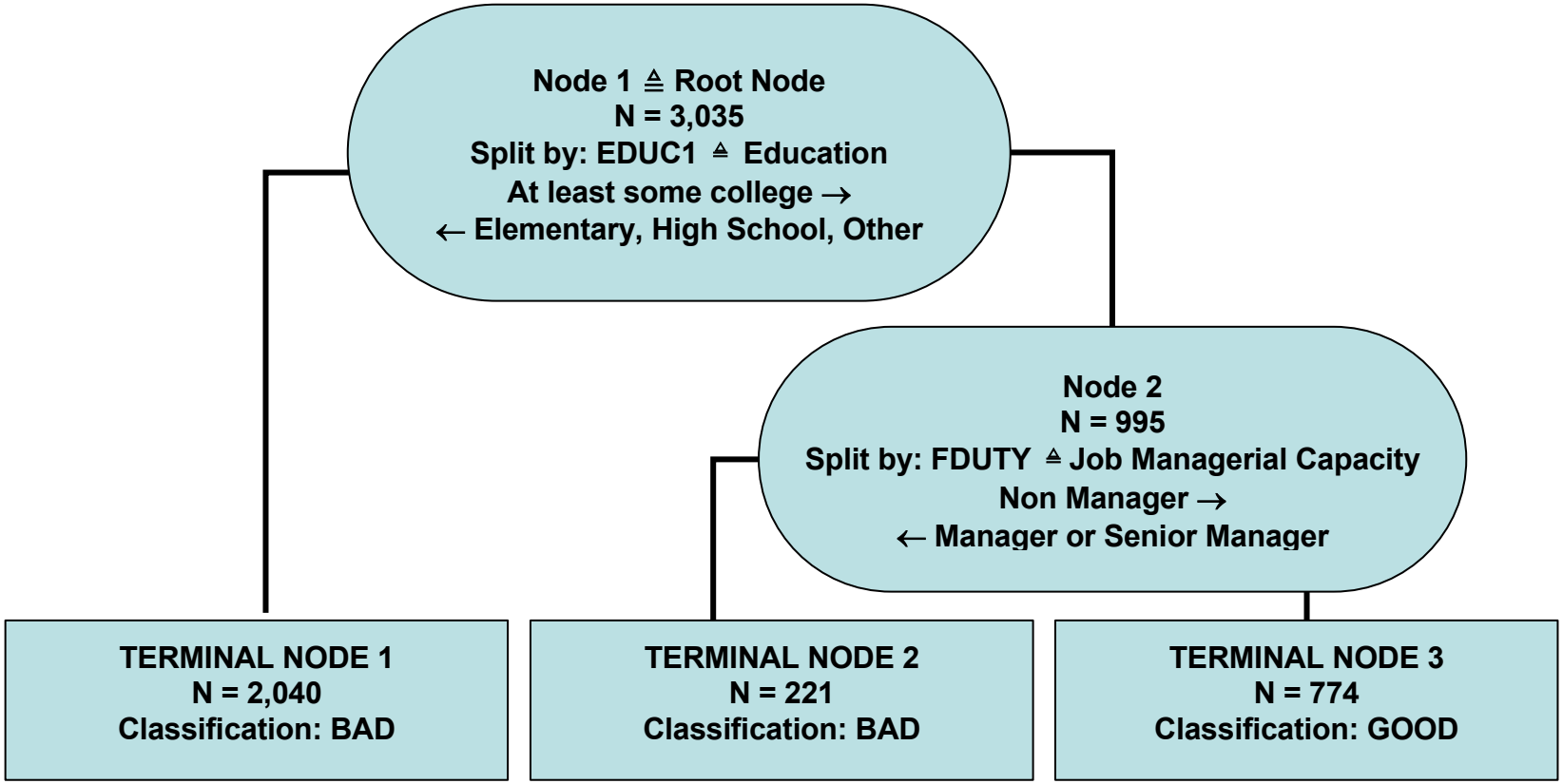
Table 1
Summary of the main characteristics of the five trees we selected for consideration

TREE	SPECIFICATIONS	# Internal Nodes: # Terminal Nodes	α=0.004, Cost Complexity	$\hat{p}(0\,|\,0)$, "0"="bad" Sensitivity	$\hat{p}(1\,|\,1)$, "1"="good" Specificity	Splits on Variables
1	C(1|0)=C(0|1)=1 GINI, CV, SERULE=0	12 : 13	.4475	.587	.662	EDUCI, PEIROD, FSTATUS, FCODE2, FCHILD, AGE1, VAL_NECS, FINC_CHI, YIT-SILK, FDUTY, ECODE2
2	C(1|0)=C(0|1)=1 GINI,CV, SERULE=1	6 : 7	.4300	.619	.577	EDUC1, PEIROD, FSTATUS, FCODE2, FCHILD, AGE1
3	C(1|0)=1.5 C(0|1)=1 GINI, CV, SERULE=0 or 1	2 : 3	.4250	.840	.334	EDUC1, FDUTY
4	C(1|0)=C(0|1)=1 GINI, TEST-SAMPLE, SERULE=0	4 : 5	.4385	.446	.717	EDUC1, PERIOD, FSTATUS, VALNECSN
5	C(1|0)=1.5, C(0|1)=1 GINI, TEST-SAMPLE, SERULE=1	5 : 6	.4620	.890	.234	EDUC1, RETINC_P, FINC_CHI, FCODE2, FSTATUS

Tree #2


Node 1 ≙ Root Node
N = 3,035
Split by: EDUC1 ≙ Education
At least some college →
← Elementary, High School, Other
Node 2
N = 2,040
Split by: PERIOD
< 27.941 →
← ≥ 27.941
TERMINAL NODE 6
N = 284
Classification: BAD
TERMINAL NODE 7
N = 995
Classification: GOOD
Node 3
N = 1,756
Split by: FSTATUS
All other categories →
← Self Employed, Student, Housewife
TERMINAL NODE 1
N = 312
Classification: BAD
Node 4
N = 1,444
Split by: FCODE2
Military Personnel, Care Giver,
Clerical/Religious Student →
← Other
Node 5
N = 1,328
Split by:
FCHILD
> 2.5 →
← ≤ 2.5
TERMINAL NODE 4
N = 480
Classification: BAD
TERMINAL NODE 5
N = 118
Classification: GOOD
Node 6
N = 848
Split by: AGE1
> 32.562 →
← ≤ 32.562
TERMINAL NODE 2
N = 520
Classification: BAD
TERMINAL NODE 3
N = 326
Classification: GOOD

Tree #3


Node 1 ≙ Root Node
N = 3,035
Split by: EDUC1 ≙ Education
At least some college →
← Elementary, High School, Other
Node 2
N = 995
Split by: FDUTY ≙ Job Managerial Capacity
Non Manager →
← Manager or Senior Manager
TERMINAL NODE 1
N = 2,040
Classification: BAD
TERMINAL NODE 2
N = 221
Classification: BAD
TERMINAL NODE 3
N = 774
Classification: GOOD