MORTGAGE DEFAULT:
CLASSIFICATION TREES ANALYSIS

David Feldman* and Shulamith Gross**#

March 24, 2003

Preliminary

Abstract. We introduce the powerful, flexible and efficient nonparametric Classification and Regression Trees (CART) algorithm to the analysis of mortgage default data while conducting the first academic study of mortgage default in Israel. 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. 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

*Department of Business Administration, School of Management, Ben-Gurion University of the Negev, P.O. Box 653, Beer-Sheva 84105, ISRAEL; telephone: +972-8-647-2106, fax +972-8-647-7691, email: feldmand@bgumail.bgu.ac.il.

**Department of Business Administration, School of Management, Ben-Gurion University of the Negev, P.O. Box 653, Beer-Sheva 84105, ISRAEL, and Department of Statistics and Computer Information Systems, Bernard M. Baruch College, The City University of New York; telephone: +972-8-647-7538, fax +972-8-647-7691, email: sgross@bgumail.bgu.ac.il.

#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, Ayala Cohen, Robert Edelstein, and Tatiana Umansky for helpful discussions, and Sivan Weiss for research assistance. We thank the Sapir Center for Development for financial support.

1 Introduction

We apply the powerful, flexible and efficient nonparametric Classification and Regression Trees (CART) algorithm[1] [Breiman, Friedman, Olshen, and Stone (1998)] to the analysis of mortgage default data while conducting the first academic study of mortgage default in Israel. 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 with outliers, is particularly appropriate for our data. Moreover, CART is intuitive and easy to interpret. 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.

Regression in general, and logistic regression in particular, require the elimination of a whole observation vector if one of its elements is missing, do not handle well cases where the number of explanatory variables is large relative to the number of cases, and, requires homogeneity, the same relations among the data all over the measurement space.

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

[1] In this paper we focus on the classification aspect of CART only. Regression analysis has been employed extensively in the literature.

securities, is in the trillions.[2] No wonder that the various aspects of mortgage contracting has 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 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 and Terrence (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).

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:

[2] 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.

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.

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. Both procedures, although very effective for small to medium data sets with a limited number of features, become unwieldy for data sets with a large number of features relative to the number of cases in the data, and for any large data set as might be found in risk assessment studies. Furthermore, under logistic regression, the classification process generally remains completely opaque, even when it provides as accurate a classification as the data warrant. Thus, for most users the estimated model is a black box. This limits the model effectiveness, the use of the model, and the appreciation for either its strength or its limitations. It also limits the development of intuition for the classification rules implied by the method.

[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))].

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] as well as less available procedures that mimic the parametric discriminant algorithms. All these methods yield classification rules that are unintelligible to the general user who is familiar with the subject matter, but is not an expert statistician.

The deficiencies of the commonly used traditional methods that we reviewed above, led us to choose a new method called CART: Classification and Regression Trees. CART is a powerful, flexible nonparametric data analysis tool. It uses binary trees, a method that Sonquist introduced in the sixties at the University of Michigan and Morgan and Messenger developed there in the seventies into an ancestor 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 in by Breiman, Friedman, Olshen, and Stone (1998).[6]

As far as we know this method has never been used in real estate data analysis. Unlike most of the above traditional methods, CART is a nonparametric method that does not require any model or functional form. It is particularly well adapted to large data sets, data sets that include a large number of features (high dimensionality), and

[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 Kth 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 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 first version of this book is from (1984).

nonstandard data structures. The latter include data sets that include a mixture of data types, data that has different relationships among variables in different parts of the measurement space (non-homogeneity), data with many outliers, and missing data. We explain below the way in which it handles very well missing data. These circumstances are precisely the circumstances that stomp CART's major traditional competitor, logistic regression. In addition, CART handles categorical variables as easily as continuous ones and is very efficient in terms of computer execution time and memory required.

CART handles missing data in a novel way, not used to our knowledge by any other classification method. First, the 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 the binary classification tree that CART produces, couching the description in our example of mortgage applicants' risk assessment when necessary.

In Sections 2 and 3 we elaborate on CART structure and methodology respectively. We do that for pedagogical reasons and to introduce a new methodology into the area of real estate data analysis. We believe that it can be used to great benefit. Section 4 describes the data, reports the results of the CART data analysis that we performed, and discusses the outcomes. Section 5 presents some general discussion and conclusion.

2 Classification Trees: Structure

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 in Section 3, CART offers several splitting methods.

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. It does not necessarily, however, minimizes 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.

For classifying data when some feature data is missing, CART provides at each node alternative splitting rules. For each missing feature CART offers an alternative splitting rule that is as close as possible to the main splitting rule in terms of minimizing the diversity of the resulting daughter nodes.

CART's efficiency and clever treatment of missing data, very prevalent in large data sets, are but two reasons for the enthusiasm with which it has been adopted in the fast developing field of Data Mining, and the field of Medical diagnosis. Another compelling reason for adopting CART over traditional model-based classifiers is its intuitive appeal. Most statistics consumers find nonlinear, generalized regression, such as logistic regression, far less intuitive, and far more indirectly related to their application, than the classification tree provided by CART. 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. No intensive model interpretation is required for understanding the output, as is the case in logistic regression for example.

Other, less obvious advantages of CART are its ability to use both continuous and categorical features, and its resistance 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 might 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.

Finally, 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.

In our risk assessment classification problem the choice of CART was indicated by two properties of our data: the large number of features (thirty three), and the large number and complexity of the missing feature data. In addition, the easy interpretability of the CART algorithms was very desirable.

Naturally, all these attractive features of CART do come at a certain cost. 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 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 Breiman, Friedman, Olshen, and Stone (1998), it was shown that in most simulated learning samples CART performed (in terms of true misclassification rate) as well or better that the 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.

Another limitation of CART is the instability of the tree structure that it provides. 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 problem exists in data sets with markedly correlated features. The problem 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 Breiman, Friedman, Olshen, and Stone (1998)].

3 Classification Trees: Method

In this section we first 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 Breiman, Friedman, Olshen, and Stone (1998) 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*. 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.

3.1 Splitting Rules

We first assign a prior probability, p_j, $0 \leq p_j \leq 1$, to every class *j* into which cases are classified, $j = 1,\ldots,J,$, 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*≤*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[7]. 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 Breiman, Friedman, Olshen, and Stone (1998)]. Let the cost, $C(i \mid j)$, of misclassifying a case that belongs to class *j* into class *i*, obey $C(i \mid j) \geq 0$ and $C(i \mid i) = 0$, and let $p(j \mid t),\ \ 0 \leq p(j \mid t) \leq 1$, $j = 1,...,J$, 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.

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

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 at going to nodes t_L and t_R respectively, is

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

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.

3.2 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

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

As was pointed out by Breiman, Friedman, Olshen, and Stone (1998), 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

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

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 $R^{TS}(T)$.

Breiman, Friedman, Olshen, and Stone (1998) 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 Breiman, Friedman, Olshen, and Stone (1998) 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

$$R_{\alpha}(\mathrm{T}) = R(\mathrm{T}) + \alpha\left|\tilde{\mathrm{T}}\right|, \qquad (10)$$

where α is a complexity coefficient, $0<\alpha$, and $|\tilde{T}|$ 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 is obtained using the concepts of sensitivity and specificity commonly used in 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 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 Breiman, Friedman, Olshen, and Stone (1998) 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.

4 Data Analysis with CART

Our data consists of end of the year 1998 information regarding 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 a subset of these data.

Our study seems to be the first Israeli academic open 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, 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. Despite this limitation, the data provided an excellent example of the use of the CART methodology, as well as a first, albeit limited, analysis of the Israeli mortgage market.

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:

Features related to the mortgage size and type

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. 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.

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.

First option combination

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.

- 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. 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 an engineer, an academic, 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 senior or regular 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 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. Prepayment terms are uniformly regulated in Israel to depend on market conditions such that, absent idiosyncratic reasons, the option to prepay 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.

5 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.

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

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.

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.

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.

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

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

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.

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.

Karolyi, A. and Sanders, A. B., 1998, "The Variation of Economic Risk Premiums in Real Estate Returns," with Andrew Karolyi, *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.

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.

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

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.

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

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

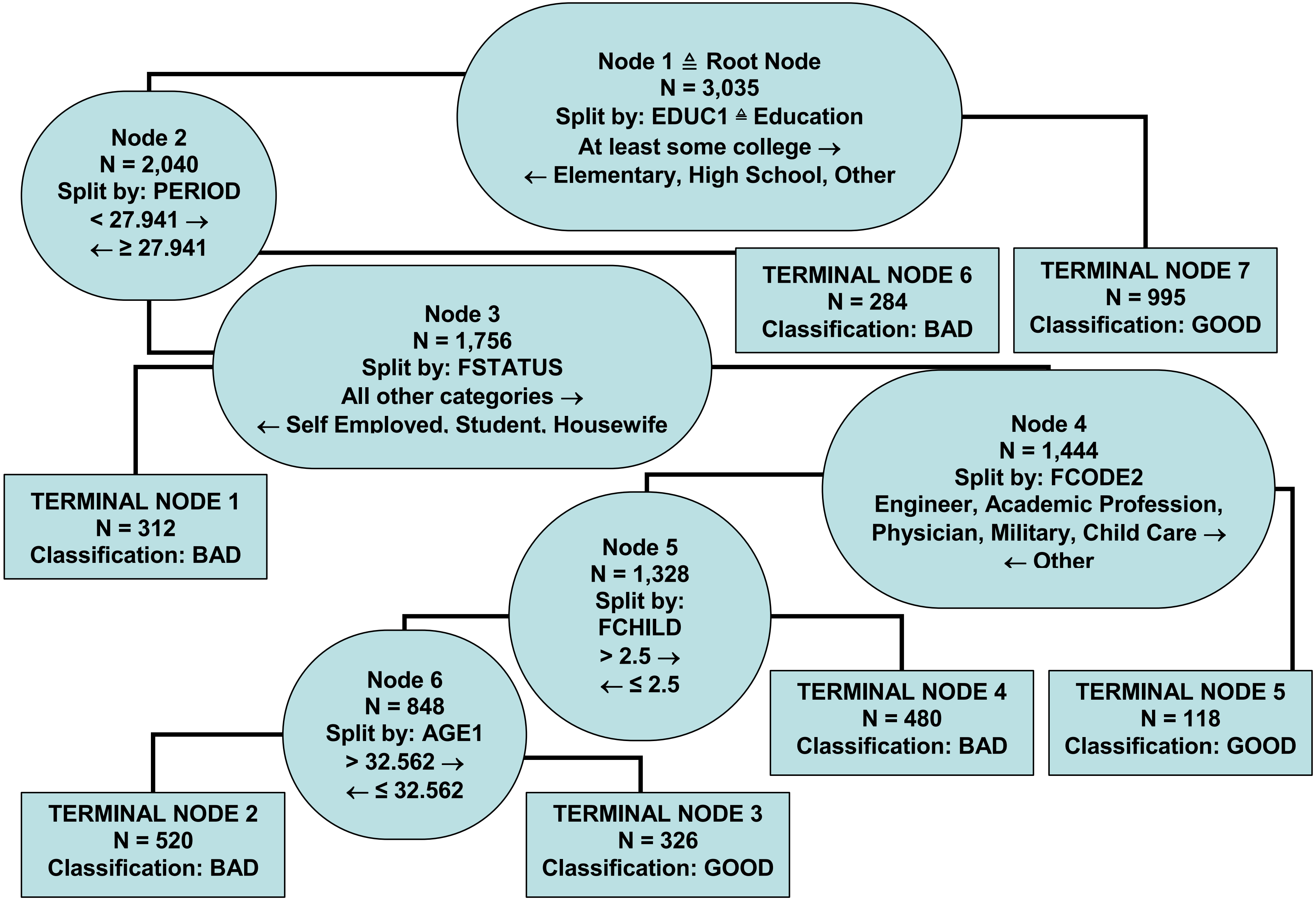
Vandell, K. D., 1993, "Handing Over the Keys: A Perspective on Mortgage Default Research," *Journal of the American Real Estate and Urban Economics Association*, 21, 211-246.

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

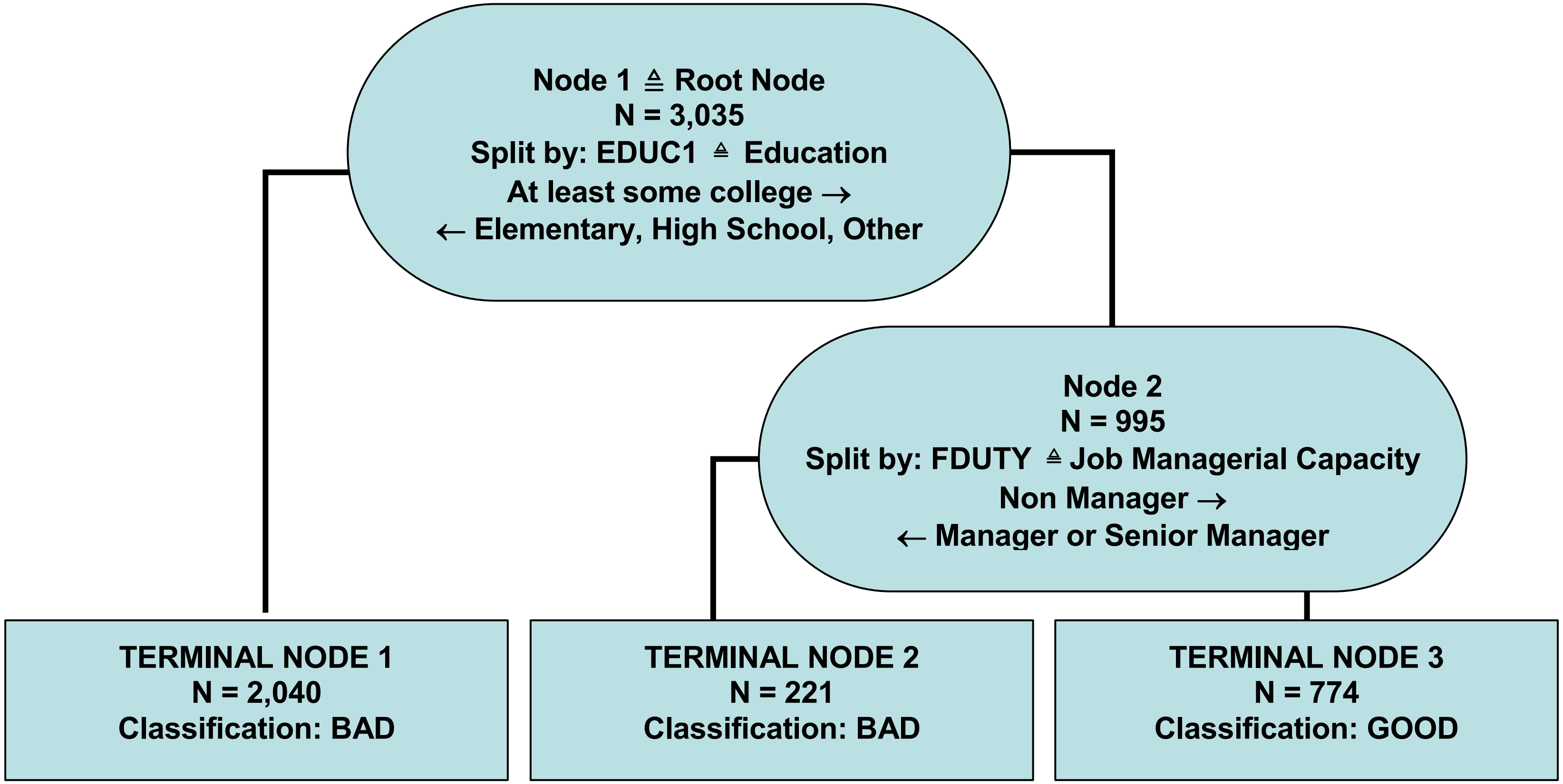
Table 1
Summary of the main characteristics of the five trees we selected for consideration

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
Node 3
N = 1,756
Split by: FSTATUS
All other categories →
← Self Employed, Student, Housewife
TERMINAL NODE 6
N = 284
Classification: BAD
TERMINAL NODE 7
N = 995
Classification: GOOD
Node 4
N = 1,444
Split by: FCODE2
Engineer, Academic Profession,
Physician, Military, Child Care →
← Other
TERMINAL NODE 1
N = 312
Classification: BAD
Node 5
N = 1,328
Split by:
FCHILD
> 2.5 →
← ≤ 2.5
Node 6
N = 848
Split by: AGE1
> 32.562 →
← ≤ 32.562
TERMINAL NODE 4
N = 480
Classification: BAD
TERMINAL NODE 5
N = 118
Classification: GOOD
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