CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Clute Institute: Journals

Rule Induction Methods For Credit Scoring

Jozef Zurada, (E-mail: jmzura01@louisville.edu), University of Louisville

ABSTRACT

Credit scoring is the term used by the credit industry to describe methods used for classifying applicants for credit into risk classes according to their likely repayment behavior (e.g. "default" and "non-default"). The credit industry has been using such methods as logistic regression, discriminant analysis, and various machine learning techniques to more precisely identify creditworthy applicants who are granted credit, and non-creditworthy applicants who are denied credit. Accurate classification is of benefit both to the creditor (in terms of increased profit or reduced loss) and to the loan applicant (avoiding overcommitment). This paper examines historical data from consumer loans issued by a financial institution to individuals that the financial institution deemed to be qualified customers. The data set consists of the financial attributes of each customer and includes a mixture of loans that the customers paid off or defaulted upon. The paper uses rule induction methods (decision trees) to predict whether a particular applicant paid off or defaulted upon his/her loan. The main advantage of decision trees is their ability to generate if-then classification rules which are intuitive and easy to understand. Rules could be explained to business managers who would need to approve their implementation as well as loan applicants as the reason for denying a loan. The paper compares the correct classification accuracy rates of several decision tree algorithms with other data mining methods proposed in earlier works.

INTRODUCTION

In the last several years, the financial services industry has experienced a rapid growth with significant increases in mortgages, auto-financing, debts to retailers, credit card debts, and home equity loans to name a few. With this growth, however, there have been mounting losses for delinquent loans. For example, in 1991, $1 billion of Chemical Bank's $6.7 billion in real estate loans were delinquent and the bank held $544 million in foreclosed property. Manufacturers Hanover's $3.5 billion commercial property portfolio was burdened with $385 million in non-performing loans (Rosenberg and Gleit, 1994). In 1994 in the UK about 12% of retail expenditure was made using credit cards, amounting to a total of about 36 billion British pounds (Hand and Henley, 1997). This dynamic is also highly relevant for the former East-block Central and East European countries, now members of the European Union. For example, in 2004 alone, Czek and Slovak banks recorded 33.8% and 36.7% increases in their retail loans, respectively (Vojtek and Kocenda, 2006). In response, many financial services institutions are developing new credit scoring models in addition to traditional statistical techniques to support their credit decisions. The ultimate objective of these models is to increase accuracy in loan-granting decisions, so that more creditworthy applicants are granted credit, thereby increasing profits, and non-creditworthy applicants are denied credit, thus decreasing losses. A slight improvement in accuracy translates into significant future savings.

Credit-risk evaluation decisions are inherently complex and unstructured due to the nonlinear relationships between independent variables that interact with each other and various forms of risks involved. The most harmful risk to the party approving credit is the nonpayment of obligations when they come due. Simultaneously, the payoff associated with a correct credit-risk decision is high. Due to the difficulty of credit risk assessment, the financial institution that provided data for this paper experienced a default rate of about twenty percent (20%), even though the financial institution must have used some credit scoring model to eliminate bad loans. Some of the defaults are governed by unforeseen factors (i.e. stability of marriage, health status and/or job stability) that may be difficult to reflect in the financial attributes of the consumer. However, some of the bad loans could be avoided by using more discriminating credit risk assessment techniques. As a result, any improvement in making a reliable discrimination, between those who are likely to repay the loan and those who are not, would be highly desired.

This paper examines and compares the effectiveness of three decision tree algorithms (chi squared, entropy reduction, and Gini reduction) to predict whether a consumer defaulted upon or paid off a loan. An original data set contains 5960 loan applicants. The data set is highly unbalanced and dominated by good loans. The proportion of good loans to bad loans is 4771 cases (80%) and 1189 cases (20%), respectively. We performed stratified sampling and divided the data set into the training set and the test set. The training set contained 4170 cases (70%) represented by 3339 good loans and 831 bad loans. The test set comprised 1790 cases (30%) divided into 1432 good loans and 358 bad loans. We recorded the results for 3 probability cutoffs: 0.3, 0.5, and 0.7. The decision trees used turned out robust and efficient at identifying both good loans and bad loans in the test set given that the financial institution that provided the data considered all of the loans contained in the data set to be good loans warranting an extension of credit. The paper assesses and analyzes the probability of default on a single loan and a group of loans, and compares the obtained results to the results published in the earlier studies (J. Zurada and M. Zurada, 2002).

The paper is organized as follows. Section 2 describes the rule induction methods fundamentals. Section 3 reviews the prior literature. Section 4 describes the data sample used in this study, whereas section 5 presents the experiments and simulation results. Finally, section 6 concludes the paper and gives some recommendations for future work.

RULE INDUCTION METHODS

A decision tree classifier is a relatively simple and widely used classification technique. A tree has three types of nodes: a root node, internal nodes, and terminal (leaf) nodes. In a binary tree, a top node is a root node that has no incoming edges and two outgoing edges (branches). Each internal node has exactly one incoming edge and two outgoing edges. Finally, each leaf or terminal node has exactly one incoming edge and no outgoing edges. Each leaf node is assigned a class label. Edges coming of the root and other internal nodes contain attribute test conditions to separate cases that have different characteristics. Classifying a test record is straightforward once a decision tree has been built. Starting from the root node, one applies the test condition to the case and follows the appropriate branch based on the outcome of the test. This will lead one either to another internal node, for which a new test condition is applied, or to a leaf node. The class label associated with the leaf node is then assigned to the record.

Efficient algorithms employing greedy search strategy exist to induce a reasonably accurate, but suboptimal, decision tree in a small amount of time. These algorithms recursively grow a decision tree by making a series of locally optimum decisions about which attribute to use for partitioning the data. The discussion below illustrates the Hunt's algorithm; on which other algorithms such as ID3, C4.5, and CART; are based (Tan *et al*., 2006).

Let D_t be the set of training patterns (examples) that are associated with node t and $y=\{y_1,y_2,\ldots,y_c\}$ be the class labels.

Step 1: If all the patterns in D_t belong to the same class y_t, then t is a leaf node labeled as y_t.

Step 2: If D_t contains patterns that belong to more than one class, an attribute test condition is selected to partition the records into smaller subsets. A child node is created for each outcome of the test condition and the patterns in D_t are distributed to the children based on the outcomes. The algorithm is then recursively applied to each child node.

Algorithms that build a decision tree should address two issues: (1) how should the training patterns be split? and (2) how should the splitting procedure stop? Selecting the best split is based on the degree of disorder/impurity of the child nodes. For example, a node which contains only cases of class 1 or class 0 (for binary classification) has the smallest disorder = 0. Similarly, a node that contains an equal number of cases of class 1 and class 0 has the highest disorder = 1. Examples of impurity measures include

$$Entropy(t) = -\sum_{i=0}^{c-1} p(i \mid t) \log_2 \, p(i \mid t)$$

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i \mid t)]^2,$$

$$Classification\ error\ (t) = 1 - \max_i [p(i \mid t)],$$

where c is the number of classes, $0\log_2 0=0$ in entropy calculations, and $p(i|t)$ is the fraction of cases belonging to class i at a given node t. The reference to node t can be omitted and the fraction can be expressed as p_i. In a 2-class problem, the class distribution at any node can be expressed as (p_0,p_1), where

$p_1=1-p_0$.

To find out how good is a test condition, the degree of disorder of the parent node is compared to the degree of disorder in the child nodes. The higher the gain Δ, the better the split.

$$\Delta = I(parent) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j),$$

where $I(\cdot)$ is the disorder measure of a given node, N is the total number of cases at the parent node, k is the number of attribute values, and $N(v_j)$ is the number of cases associated with the child node, v_j.

Chi-squared splitting criteria measure the reduction in variability of the target distribution in the branch (child) nodes. Specifically, the likelihood ratio Pearson chi-squared test statistic is a measure of association between the categories of the dependent variable and the branch nodes. This test statistic can be used to judge the worth of the split; it measures the difference between the observed cell counts and what would be expected if the branches and target classes were independent. We used a default significance level of 0.20.

For more details on decision trees, refer to (Mitchell, 1997; Han and Kamber, 2001; Giudici, 2003; SAS Enterprise Miner: http://www.sas.com; and Tan et al., 2006).

PRIOR RESEARCH

A variety of machine learning techniques have been applied to credit scoring and credit risk assessment in the recent years. For example, papers by Rosenberg and Gleidt (1994) and Hand and Henley (1997) were concerned with detecting and reducing loan defaults and serious delinquencies. Other credit risk management concerns discussed in the two papers were the maintenance of existing credit lines and determining the best action to be taken on delinquent accounts.

Barney *et al*. (1999) compared the performance of neural networks and regression analyses in identifying the farmers who had defaulted on their Home Administration Loans and those farmers who paid off the loans as scheduled. Using an unbalanced data, Barney found that neural networks outperform logistic regression in correctly classifying farmers into those who made timely payments and those who did not. Jagielska *et al*. (1999) investigated credit risk classification abilities of neural networks, fuzzy logic, genetic algorithms, decision trees, and rough sets and concluded that the genetic/fuzzy approach compared more favorably with the neuro/fuzzy and rough set approaches. In two of his papers, Piramuthu (1999) analyzed the beneficial aspects of using both neural networks and nuero-fuzzy systems for credit-risk evaluation decisions. Piramuthu used three real-world applications data that involved credit-risk evaluation in various forms: credit approval, loan default, and bank failure prediction. Neural networks performed significantly better than neuro-fuzzy systems in terms of classification accuracy, on both training as well as testing data. However, the neural network cannot explain the rationale behind its credit granting/denial decision, unlike the neuro-fuzzy systems that explain decisions using simple if-then rules. West (2000) investigated

the credit scoring accuracy of five neural network architectures and compared them to traditional statistical methods. The neural architectures and traditional models included multilayer perceptron, mixture-of-experts, radial basis function, learning vector quantization, and fuzzy adaptive resonance; and discriminant analysis, logistic regression, *k* nearest neighbor, kernel density estimation, and decision trees, respectively. Using two real world data sets and testing the models using 10-fold crossvalidation, the author found that among neural architectures the mixture-of-experts and radial basis function did best, whereas among the traditional methods regression analysis was the most accurate. Thomas (2000) surveyed the techniques for forecasting financial risk of lending to consumers, Yang *et al*. (2001) examined the application of neural networks to an early warning system for loan risk assessment, and J. Zurada and M. Zurada (2002) reported some preliminary results comparing the performance of data mining techniques in predicting the credit worthiness of customers. Also, Feldman and Gross (2005) applied decision trees for detecting mortgage default rates.

DATA SET USED IN THE STUDY

In our paper we used variables describing loan and consumer characteristics that were similar to the ones used in the previous studies (Quinlan, 1987; Fahrmeir and Hamerle, 1994; Desai *et al*., 1996; West, 2000; and Giudici, 2003).

We used a sample data provided by a money lending institution. The data set contains financial information about 5960 consumers allocated among 13 variables. The financial institution extended loans to all of the applicants in the data set since all of them seemed to be qualified customers. Those applicants who were denied a loan during the application process were not included in the data sample which we investigated. Out of the 5960 applicants, 4771 had paid off their loans and 1189 defaulted on the loans, resulting in a default rate of approximately 20%. Because decision trees tolerate missing values quite well, we have not used any imputation methods to replace missing values in the data set nor we have discarded cases with missing values. Out of these 13 variables, there were 12 independent variables, and one binary dependent variable whose two states we are trying to predict. Using this data set, we built three decision tree models to predict whether a future applicant will default upon a loan or pay it off. The 12 independent variables are: (1) Amount of the current loan request, (2) Amount due on existing mortgage, (3) Value of current property, (4) Debt-to-income ratio, (5) Years on current job, (6) Number of major derogatory credit reports, (7) Number of credit lines, (8) Number of delinquent credit lines, (9) Number of recent credit inquiries, (10) Age (in months) of oldest trade line, (11) Reason for loan (debt consolidation or home improvement), and (12) Applicants' job category. The binary dependent variable Loan_Status takes values of 1 (client defaulted on loan or seriously delinquent) or 0 (loan paid off).

It is worth noting that almost all data sets used in modeling credit scoring problems have an inherent bias as they contain only those creditworthy customers actually given a loan. In other words, such data sets very often do not include customers who did not get a loan and we do not know whether or not they would have been at risk. Although these remarks do not affect the validity of the analysis, we should keep them in mind.

EXPERIMENTS AND RESULTS

We performed computer simulation for the three different decision tree methods (chi squared, entropy reduction, and Gini reduction) and tested their classification accuracy in terms of identifying good loans and bad loans.

The data set contained 5960 loan applicants. It is highly unbalanced and dominated by good loans. The proportion of good loans to bad loans is 4771 cases (80%) and 1189 cases (20%). We performed stratified sampling and allocated 70% of cases to the training set and 30% of cases to the test set. As a result, the training set contained 4170 cases (70%) divided into 3339 good loans and 831 bad loans. The test set comprised 1790 cases (30%) divided into 1432 good loans and 358 bad loans. We built the models on the training set and tested the performance of the models on the test set.

The results from computer simulation are shown Tables 1 and 2 and Figures 1 through 5. We recorded the "overall", "repaid", and "defaulted" correct classification accuracy rates for the three cutoff probabilities, i.e., .3, .5,

and .7. The choice of the best model may depend on a cutoff probability threshold that a financial institution chooses to use. Because the target event was detecting loan defaults, the .3 cutoff implies that the cost of making an error of granting a loan, when it should not be granted, is 3.3 times higher than the cost of denying a loan, when it should be granted. This cutoff may be applicable to situations in which banks do not secure smaller loans, i.e., do not hold any collateral. In other words, a .3 cutoff would allow a financial institution to eliminate the most likely loan defaulters and extend a credit to the most creditworthy customers. Consequently, the .5 cutoff means that the cost of making an error of granting a loan, when it should be denied, is equal to the cost of denying a loan, when it should be granted. Finally, the .7 cutoff implies that the cost of making an error of granting a loan, when it should be denied, is 3.3 times smaller than the cost of denying a loan, when it should be granted. The latter two cutoffs may typically be used when a financial institution secures larger loans by holding collateral such as title on a car or house purchased by the customer.

It appears that for the .3 cutoff, detecting bad loans is paramount. The overall, good (loan paid off), and bad (loan defaulted upon) correct classification accuracy rates of the three algorithms appear to be statistically insignificant (Table 1). The results show that the chi squared method correctly classifies 77.1% of bad loans.

For the .5 cutoff, the correct classification accuracy rates (overall, bad, and good) are not statistically different between the three different methods. These results are not significantly different either from the results reported for the same cutoff probability in the study by J. Zurada and M. Zurada (2002).

The 0.7 cutoff, applied to the decision tree produced by the chi squared method, yields 88.1%, 96.4%, and 54.8% for the overall, repaid, and defaulted correct classification accuracy rates, respectively. The Gini reduction method generates very similar results. The small differences in classification accuracy rates between the two mentioned methods are statistically insignificant. The entropy reduction method is the only method that produces significantly inferior classification accuracy rates at the .01 significance level for loan defaults (38%) than the two remaining methods.

Table 2 shows that all three decision tree models identify a Debt-to-income variable as the most important in predicting the outcome of the target variable. The variable's relative importance is 1. Also, the three models are in solid agreement as to which next three variables have the most predictive power. These variables are Number of delinquent credit (trade) lines, Age (in months) of the oldest trade line, and Value of current property. Their averaged relative importance across the three methods is 0.35, 0.27, and 0.24, respectively.

Figures 1 through 3 show the actual decision trees generated by the three models. The depth of the trees is limited to 3 levels. Although the classification accuracy rates for the 3 decision tree models are statistically insignificant for the 3 probability cutoffs, it appears that the chi squared and entropy reduction methods would be preferable because they generated the simpler trees in terms of the number of rules (splits) used. We emphasize that decision rules are important because they enable compact explanation of data as well as explain the relationships between the independent variables and the dependent variable.

The properly interpret the cumulative percent response chart (Figure 4) one needs to understand how three curves on this chart are constructed. Our target event is detecting loan defaults; thus, a target event is defined as an individual who defaults on a loan (Loan_Status=1). For each individual the 3 decision tree models predict the likelihood that the individual will default. For each model, the test cases are first sorted by the predicted probability of response in the descending order, from the highest likelihood or response to the lowest likelihood of response. Next the test cases are grouped into ordered deciles, each containing about 10% of the data, and finally percentage of actual respondents in each decile is counted using the target variable Loan_Status. The horizontal line represents the baseline rate (20%) for comparison purposes. This line indicates that the probability of selecting a loan defaulter at random from the population (the original data set) is 20%. One can see that the performance of the 3 models is very similar and the three curves almost overlap each other. For example, the chi squared model detects 82.4% of defaulters in the 1st decile (10th percentile); more than 4 times more than the response rate in the population (20%). In the first two deciles (20th percentile), the Gini reduction model correctly identifies 73.4% of defaulters and it slightly

outperforms the other two models. The cumulative charts are very useful if a loan granting institution wanted to target only 10%, 20%, 30% or more of the most likely loan defaulters.

The above findings about the strengths of the three models are confirmed by the ROC chart. Without going into details, one can say that the performance of the three models is very similar. The more the curves push up and to the left, the better the models. The left and right parts of the chart represent higher and lower probability cutoffs, respectively.

Based on the overall analysis of the obtained classification rates, we recommend implementing the chi squared method because it produced simpler and fewer rules, which are easy to interpret, than the two other methods (Table 2 and Figure 1).

CONCLUSION AND SUGGESTIONS FOR FUTURE RESEARCH

The paper examined historical data from consumer loans issued by a financial institution to individuals whom the financial institution deemed to be qualified customers. Accurate classification is of benefit both to the creditor (in terms of increased profit or reduced loss) and to the loan applicant (avoiding overcommitment). The paper uses rule induction methods (decision trees) to predict whether a particular applicant paid off or defaulted upon his/her loan. The main advantage of decision trees is their ability to generate if-then classification rules which are intuitive and easy to understand. Such rules could be explained to business managers who would need to approve their implementation as well as loan applicants as the reason for denying a loan.

Further research should focus on refining the training and testing of the decision tree models, fine tuning the models as well as using various balanced and unbalanced data sets to improve the classification performance. It would allow one to establish the most profitable lending policy from a credit risk perspective based on the projected profits on good loans, average losses on bad loans, and the fixed and variable costs of lending operations.

REFERENCES

1. Barney, D.K., Graves, O.F., and Johnson, J.D., The Farmers Home Administration and Farm Debt Failure Prediction, *Journal of Accounting and Public Policy*, Vol. 18, pp. 99-139, 1999.
2. Desai, V.S., Crook, J.N., and Overstreet, G.A.Jr, A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment, *European Journal of Operation Research*, Vol. 95, pp. 24-37, 1996.
3. Fahrmeir, L., and Hamerle, A., *Multivariate Statistical Modeling Based on Generalized Linear Programs*, Springer-Verlag, Berlin, 1994.
4. Feldman, D., and Gross, S., Mortgage Default: Classification Tree Analysis, *Journal of Real Estate Finance and Economics*, Vol. 30, 369-396, 2005.
5. Giudici, P., *Applied Data Mining: Statistical Methods for Business and Industry*, John Wiley & Sons Ltd., Chichester, West Sussex, England, 2003.
6. Han, J., and Kamber, M., *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, CA, 2001.
7. Hand, D., and Henley, W., Statistical Classification Method in Consumer Scoring: A Review, *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, Vol. 160, 523-541, 1997.
8. Hand, W., and Vinciotti, V., Choosing *k* for Two-Class Nearest Neighbor Classifiers with Unbalanced Classes, *Pattern Recognition Letters*, Vol. 24, 1555-1562, 2003.
9. Henley, W., and Hand, D., A *k*-nearest Neighbor Classifier for Assessing Consumer Credit Risk, *Statistician*, Vol. 45, 77-95, 1996.
10. Jagielska, I., Matthews, C., and Whitfort, T., An Investigation into the Application of Neural Networks, Fuzzy Logic, Genetic Algorithms, and Rough Sets to Automated Knowledge Acquisition for Classification Problems, *Neurocomputing*, Vol. 24, pp. 37-54, 1999.
11. Mitchell, T.M., *Machine Learning*, WCB/McGraw-Hill, Boston, Massachusetts, 1997.

12. Piramuthu, S., Financial Credit-Risk Evaluation with Neural and Neurofuzzy Systems, *European Journal of Operational Research*, Vol. 112, 310-321, 1999.
13. Piramuthu, S., Feature Selection for Financial Credit-Risk Evaluation Decisions, *INFORMS Journal on Computing*, Vol. 11, No. 3, pp. 258-266, 1999.
14. Quinlan, J.R., Simplifying Decision Trees, *International Journal of Man-Machine Studies*, Vol. 27, pp. 221-234, 1987.
15. Rosenberg, E., and Gleidt, A., Quantitative Methods in Credit Management: A Survey, *Operations Research*, Vol. 42, 589-613, 1994.
16. Tan, P-N., Steinbach, M., and Kumar, V., *Introduction to Data Mining*, Addison-Wesley, 2006.
17. Thomas, L.C., A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Consumers, *International Journal of Forecasting*, Vol. 16, pp. 149-172, 2000.
18. Vojtek, M, and Kocenda, E, Credit Scoring Methods, *Czech Journal of Economics and Finance*, Vol. 56, Issue 3-4, 152-167, 2006.
19. West, D., Neural Network Credit Scoring Models, *Computers & Operations Research*, Vol. 27, pp. 1131-1152, 2000.
20. Yang, B., Li, L.X., Ji, H., and Xu, J., An Early Warning System for Loan Risk Assessment Using Artificial Neural Networks, *Knowledge-Based Systems*, Vol. 14, pp. 303-306, 2001.
21. Zurada, J., and Zurada, M., How Secure are Good Loans: Validating Loan-Granting Decisions and Predicting Default Rates on Consumer Loans, *The Review of Business Information Systems*, 6(3), pp. 65-83, 2002.

Table 1. Correct Classification Accuracy Rates For The Three Decision Tree Models.

	DT – Chi Squared		**DT – Entropy Reduction**		**DT – Gini Reduction**	
	Count	%	Count	%	Count	%
Cutoff=0.3						
Overall	1560	87.2	1595	89.1	1598	89.3
Paid off	1284	89.7	1333	93.1	1332	93.0
Defaulted	276	77.1	262	73.2	266	74.3
Cutoff=0.5						
Overall	1596	89.2	1595	89.1	1600	89.4
Paid off	1347	94.1	1333	93.1	1336	93.3
Defaulted	249	69.6	262	73.2	264	73.7
Cutoff=0.7						
Overall	1577	88.1	1536	85.8	1576	88.0
Paid off	1381	96.4	1400	97.8	1380	96.4
Defaulted	196	54.8	136	38.0	196	54.8

Table 2. Relative Importance Of The Variables Used In The Three Decision Tree Models.

Variable Name	DT – Chi Squared		DT – Entropy Reduction		DT – Gini Reduction	
	Importance	Used in x Rules	Importance	Used in x Rules	Importance	Used in x Rules
Debt-to-income ratio	1.0	2	1.0	2	1.0	2
Number of delinquent credit (trade) lines	0.33	2	0.37	3	0.33	2
Age (in months) of oldest trade line	0.29	2	0.26	1	0.29	3
Value of current property	0.24	1	0.23	1	0.24	1
Number of recent credit inquiries	0.09	1	0.0	0	0.13	0
Reason for a loan (home improvement or debt consolidation)	0.09	1	0.0	0	0.0	0
Amount due on existing mortgage	0.0	0	0.0	0	0.09	1
Years on current job	0.0	0	0.12	1	0.12	1
Applicant's job category	0.0	0	0.08	1	0.15	3
Number of trade (credit) lines	0.0	0	0.0	0	0.08	1

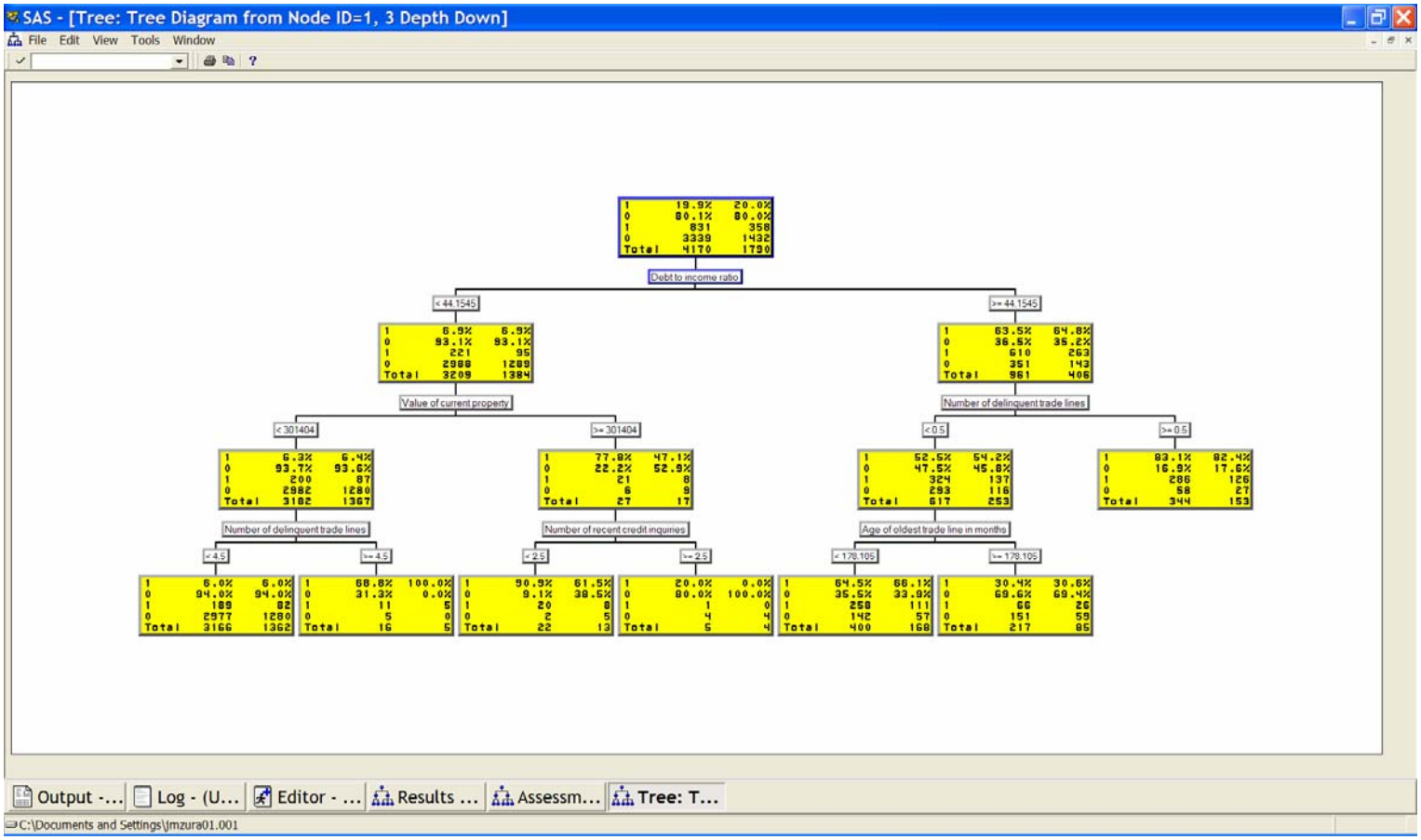

Figure 1. The Decision Tree Diagram For The Chi Squared Method.

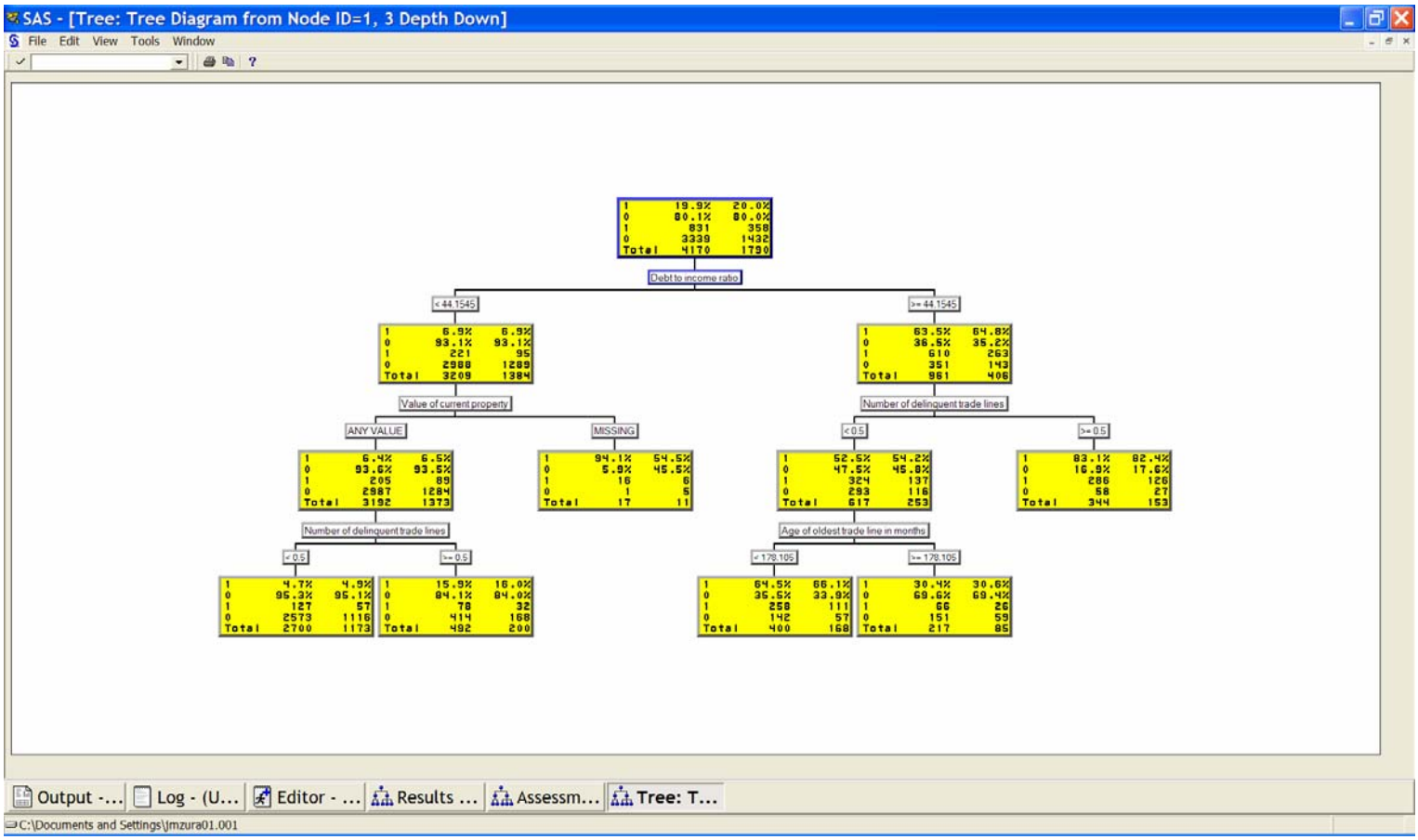

Figure 2. The Decision Tree Diagram For The Entropy Reduction Method.

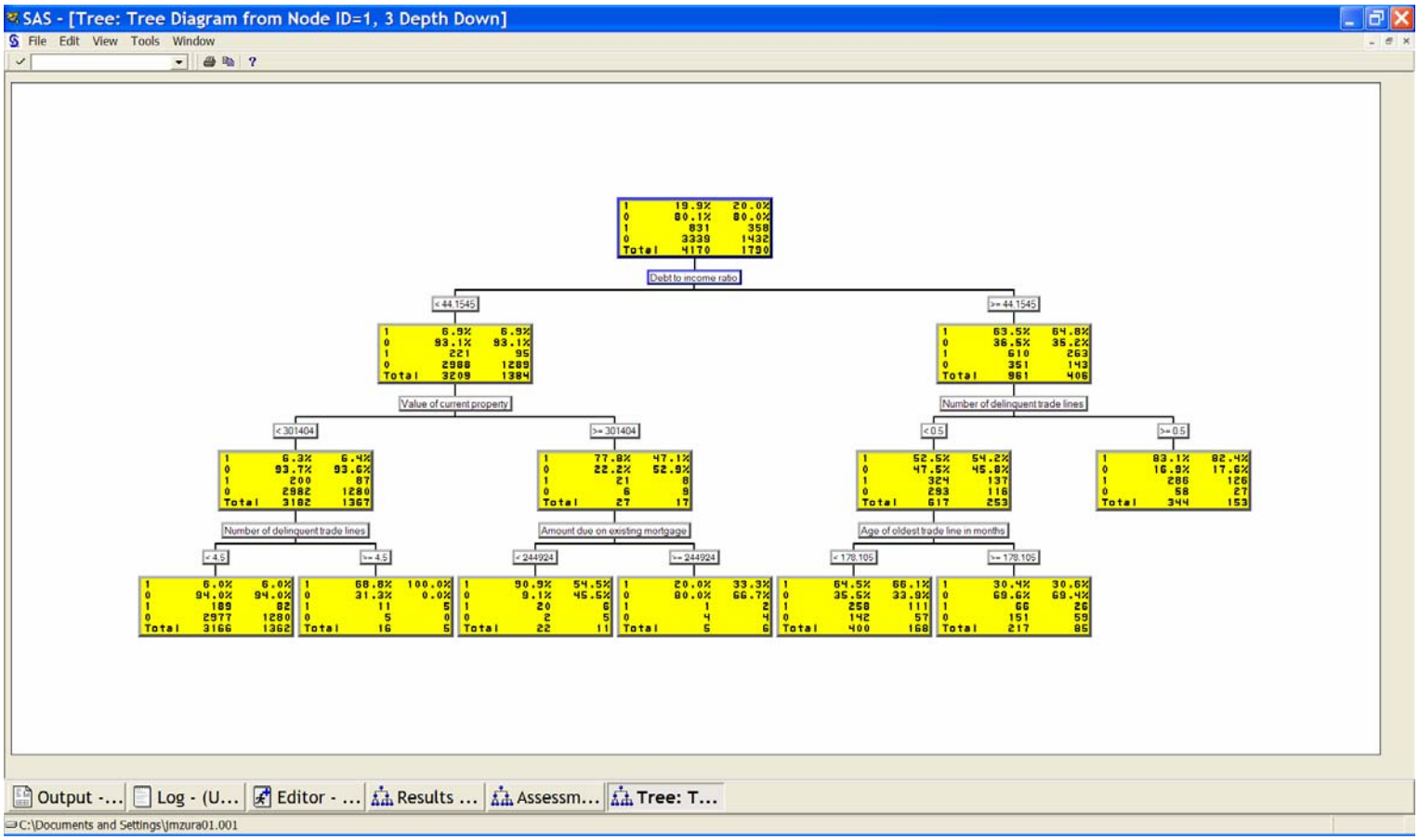

Figure 3. The Decision Tree Diagram For The Gini Reduction Method.

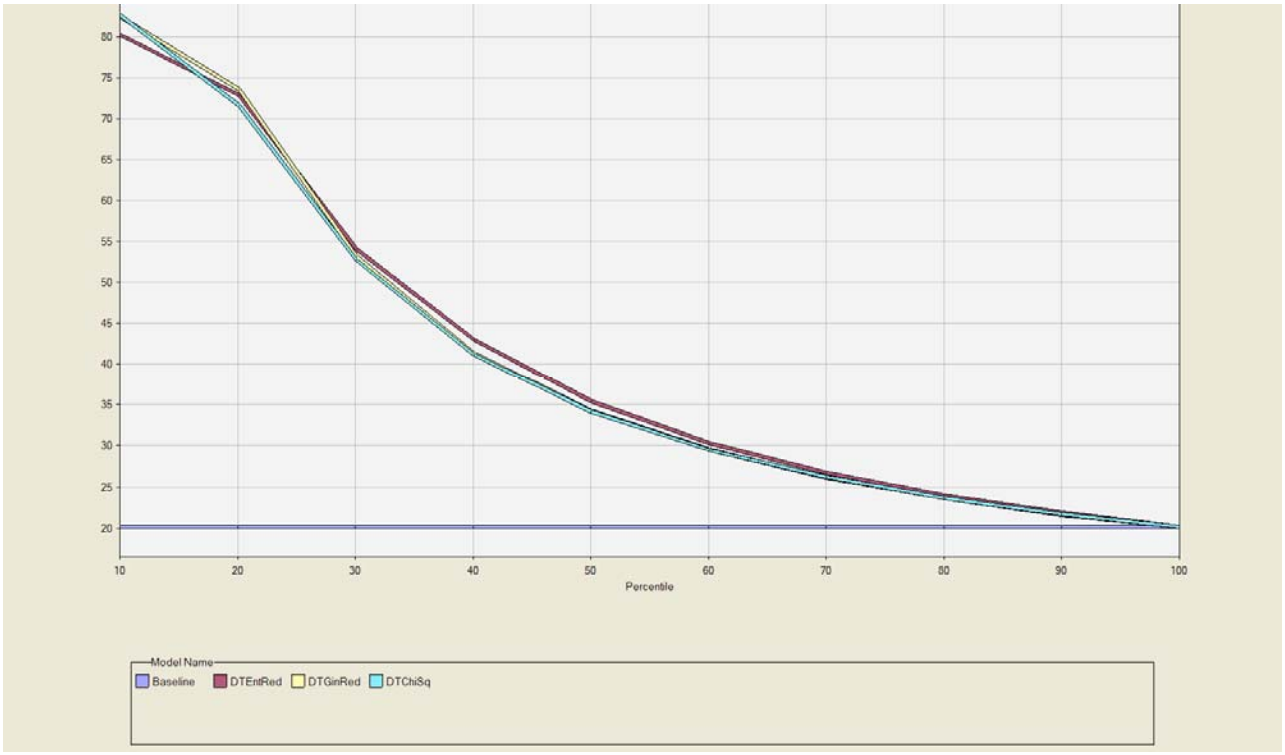

Figure 4. Cumulative Percent Characteristics Chart For The Three Decision Tree Models

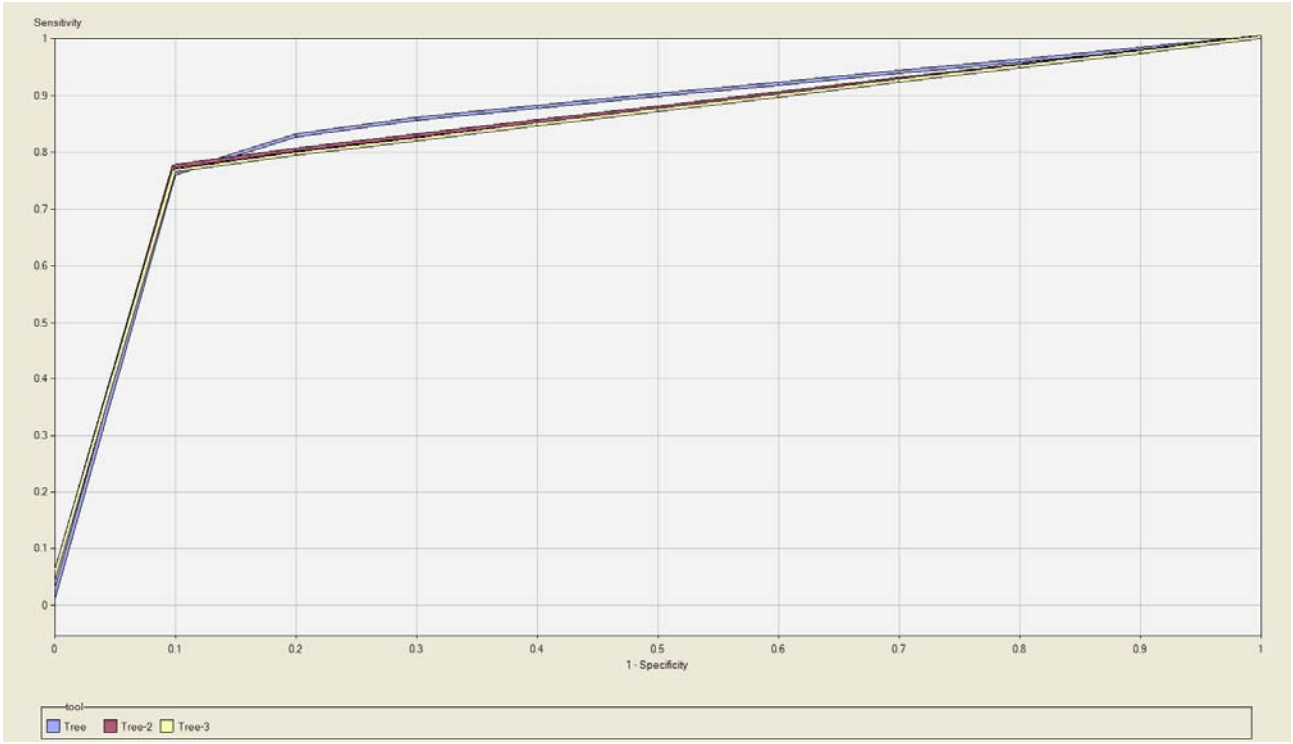

Figure 5. ROC Chart for the Three Decision Tree Models.

NOTES