How Secure Are "Good Loans": Validating Loan-Granting Decisions And Predicting Default Rates On Consumer Loans

Jozef Zurada, (Email: jmzura01@gwise.louisville.edu), University of Louisville
Martin Zurada, (E-mail: mzurada@hotmail.com)

Abstract

The failure or success of the banking industry depends largely on the industry's ability to properly evaluate credit risk. In the consumer-lending context, the bank's goal is to maximize income by issuing as many good loans to consumers as possible while avoiding losses associated with bad loans. Mistakes could severely affect profits because the losses associated with one bad loan may undermine the income earned on many good loans. Therefore banks carefully evaluate the financial status of each customer as well as their credit worthiness and weigh them against the banks' internal loan-granting policies. Recognizing that even a small improvement in credit scoring accuracy translates into significant future savings, the banking industry and the scientific community have been employing various machine learning and traditional statistical techniques to improve credit risk prediction accuracy.

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 consists of the financial attributes of each customer and includes a mixture of loans that the customers paid off and defaulted upon. The paper uses three different data mining techniques (decision trees, neural networks, logit regression) and the ensemble model, which combines the three techniques, to predict whether a particular customer defaulted or paid off his/her loan. The paper then compares the effectiveness of each technique and analyzes the risk of default inherent in each loan and group of loans. The data mining classification techniques and analysis can enable banks to more precisely classify consumers into various credit risk groups. Knowing what risk group a consumer falls into would allow a bank to fine tune its lending policies by recognizing high risk groups of consumers to whom loans should not be issued, and identifying safer loans that should be issued, on terms commensurate with the risk of default.

1. Introduction

In the last several years, the financial services industry has experienced a rapid growth with significant increases in single-family mortgages, auto-financing, 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 response, many financial services institutions are developing new credit scoring models 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.

Determining whether a particular consumer should receive a loan is an inherently complex and unstructured process. A financial institution must examine many independent financial attributes of each loan candidate in,

Readers with comments or questions are encouraged to contact the authors via email.

an accurate prompt, and cost effective manner. The financial institution approximates the risk of default by the candidate, and weighs that risk against the benefit of potential earnings on the loan. Due to the difficulty of credit risk assessment, the financial institution that provided data for this paper experienced a default rate of about nine percent (9%), even though the financial institution must have used some credit scoring model to eliminate "bad loans". Some of the defaults can be attributed to unforeseen events (i.e. divorce, death, loss of employment) or governed by factors that may be difficult or impossible to reflect in the financial attributes of the consumer (i.e. stability of marriage, health, job stability). However, some of the "bad loans" could be avoided by using more discriminating credit risk assessment techniques. Any improvement in making a reliable distinction between those who are likely to repay the loan and those who are not would allow the bank to reject the riskiest loans and to adjust the terms of the granted loans according to the risk of default.

The volume and complexity of raw data inherent in credit-risk assessment can be tackled by several knowledge discovery and data mining tools. Knowledge discovery is defined as the process of identifying valid, novel, and potentially useful patterns in data (Fayyad *et al*., 1996). Knowledge discovery relies on many well-established technologies, such as machine learning, pattern recognition, statistics, neural networks, fuzzy logic, evolutionary computing, database theory, artificial intelligence, and high performance computing. All of these technologies are applied to find knowledge in very large databases. The knowledge discovery process is typically composed of the following phases: understanding the overall problem; obtaining a data set; cleaning, preprocessing, and reducing data; data mining; interpreting mined patterns; and consolidating discovered knowledge.

Data mining is an important phase in the knowledge discovery process. The most prominent examples of data mining tools include: regression and discriminant analysis, neural networks, decision trees, fuzzy logic and sets, rough sets, genetic algorithms, association rules, and *k*-nearest neighbor. These tools are suitable for the tasks of classification, prediction, clustering, and optimization. Classification and prediction are the most common and perhaps the most straightforward data mining tasks. Classification consists of examining the features of a newly presented object and assigning it to one of a predefined set of classes or outcomes ("loan defaulted upon" or "loan paid off"). A data mining model employing one of the data mining tools must be trained using pre-classified examples. The goal is to build a model that will be able to accurately classify new data based on the outcomes and the interrelation of many discrete variables (income, length of employment etc.) contained in the training set.

This paper examines and compares the effectiveness of three data mining techniques and the ensemble model used in two scenarios to predict whether a consumer defaulted or paid off a loan. In the first scenario, the data set (comprised of the training, validation, and test subsets) for each technique contained a large ratio of good loans to bad loans. The data mining techniques were outstanding at identifying the good loans and adept at identifying some of the bad loans from the test set. In the second scenario, the data set contained an equal mix of good loans and bad loans. The data mining techniques proved proficient at identifying both good loans and bad loans in the test set. The ensemble model which combined decision trees, neural networks, and logit regression techniques proved to be the most skillful at classifying the good loans and bad loans in both scenarios. The models were impressive at distinguishing good loans from bad loans 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 collection of loans. It then discusses how financial institutions could use the probability of loan default to more precisely match their lending policies to existing credit risk.

The paper is organized as follows: section 2 reviews the current knowledge discovery and data mining applications in the financial domain; section 3 explains the fundamentals of data mining tools -- decision trees, neural network, and logistic regression; section 4 explains the data sample used in this study; section 5 describes the Enterprise Miner nodes used for data analysis; section 6 describes the experiments and simulation results; and section 7 concludes the paper and provides some recommendations for future work.

2. Literature Review

During recent years a plethora of articles have been published showing that data mining tools are very useful in supporting financial decisions. The articles mainly concerned bankruptcy prediction (Wilson and Sharda, 1994; Lee *et al*., 1996; Back *et al*., 1996; Jo and Han, 1997; Olmeda and Fernandez, 1997; Zhang *et al*., 1999; Yang *et al*., 1999; Greenstein and Welsh, 2000), financial distress classification (Coats and Fant, 1993; Altman *et al*., 1994; Lacher *et al*., 1995, Zurada *et al*., 2001), going concern status (Koh and Tan, 1999), business failure predictions (Tam and Kiang, 1992; Boritz and Kennedy, 1995; Dimitras *et al*., 1999), and prediction of the success or failure of new ventures (Jain and Nag, 1997). Data mining techniques have also been successfully applied to credit-risk assessment problems.

The initial research focused on determining the usefulness of data mining tools, such as neural networks and decision trees, and examining how these tools should be applied in a credit-risk assessment context. In one of the early papers, McLeod *et al*. (1993) discussed general features of neural networks and their suitability for the credit-granting process. Glorfeld and Hardgrave (1996) presented a comprehensive and systematic approach to developing an optimal architecture of a neural network model for evaluating the creditworthiness of commercial loan applications. The neural network developed using their architecture was capable of correctly classifying 75% of loan applicants and was superior to neural networks developed using simple heuristics.

Tessmer (1997) examined credits granted to small Belgian businesses using a decision tree-based learning approach. Tessmer focused on the impact of Type I credit errors (classifying good loans as bad loans), and Type II credit errors (classifying bad loans as good loans), on the accuracy, stability and conceptual validity of the learning process. Tessmer argued that near misses have the ability to nudge the learning process towards a more accurate definition of the boundary between positive and negative examples. Tessmer recommended the procedure called "the dynamic updating process", which relocates the boundary between Type I and Type II errors, to define a more informed credit granting decision.

Subsequent authors built on the existing research by comparing the performance of various data mining techniques in various credit risk assessment contexts. Desai *et al*. (1996) analyzed the usefulness of neural networks and traditional techniques, such as discriminant analysis and logistic regression, in building credit scoring models for credit unions. Desai studied data samples containing 18 variables collected from three credit unions and showed that neural networks were particularly useful in detecting bad loans, whereas logistic regression outperformed neural networks in the overall (bad and good loans) classification accuracy.

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, rule induction software, and rough sets and concluded that the genetic/fuzzy approach compared more favorably with the neuro/fuzzy and rough set approaches.

Piramuthu (1999) analyzed the beneficial aspects of using both neural networks and nuerofuzzy 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 neurofuzzy 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 neurofuzzy systems that explain decisions using simple if-then rules.

In a very comprehensive study, 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.

In other recent publications focusing on loan risk assessment, 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 Zurada (2001) reported some preliminary results comparing the performance of data mining techniques in predicting the credit worthiness of customers.

3. Decision Trees, Neural Network, And Logit Regression Fundamentals

This section gives a very brief overview of decision trees, neural networks, and logistic regression and discusses their fundamentals. Decision trees and logistic regression are well-established traditional statistical techniques, whereas neural networks are relatively new data mining tools that have been successfully used for classification and prediction.

3.1. Decision Trees

Decision trees are particularly useful for classification tasks. Like neural networks, decision trees learn from data. Using search heuristics, decision trees are able to find explicit and understandable rules-like relationships among independent and dependent variables. Search heuristics use recursive partitioning algorithms to split the original data into finer and finer subsets, or clusters. The algorithms have to find the optimum number of splits and determine where to partition the data to maximize the information gain. The fewer the splits, the more explainable the output is (there are less rules to understand).

Decision trees are built of nodes, branches and leaves that indicate the variables, conditions, and outcomes, respectively. The most predictive variable is placed at the top node of the tree. The operation of the decision tree is based on the ID3 or C4.5 algorithms (Quinlan, 1993; Mitchell, 1997). The algorithms make the clusters at the node gradually purer by progressively reducing disorder (impurity) in the original data set. Disorder and impurity can be measured by the well-established measures of entropy and information gain borrowed from information theory. We briefly introduce these measures below.

Given a collection *S*, containing the positive (*yes*) and negative examples (*no*) of some target concept, the entropy of *S* relative to this Boolean classification is

$$Entropy(S) \equiv -p_{yes}\log_2 p_{yes} - p_{no}\log_2 p_{no}$$

where p_{yes} is the proportion of positive examples in *S* and p_{no} is the proportion of negative examples in *S*. If the target attribute can take on *k* different values, then the entropy of *S* relative to this *k*-wise classification is defined as

$$Entropy(S) \equiv \sum_{i=1}^{k} - p_i \log_2 p_i$$

The information gain, *Gain(S,A)* of an attribute *A*, relative to a collection of examples *S*, is defined as

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{S_v}{S} Entropy(S_v)$$

where *Values(A)* is the set of all possible values for attribute *A*, and S_v is the subset of *S* for which attribute *A* has the value *v* (i.e., $S_v = \{ s \in S \mid A(s) = v \}$.

One of the greatest advantages of decision trees is the fact that knowledge can be extracted and represented in the form of classification if-then rules. Each rule represents a unique path from the root to each leaf. In addition, at each node one can measure the number of records entering the node, the way those records would be classified if these were leaf nodes, and the percentage of records classified correctly.

3.2. Artificial Neural Networks

Artificial neural networks are one of the most common data mining tools. Neural networks have attracted the attention of researchers because they are particularly useful for the tasks of classification, prediction, and clustering in business applications. Neural networks try to emulate biological neurological systems. In other words, they try to mimic the way the human brain functions and processes information. Neural network models are characterized by three properties: the computational property, the architecture of the network, and the learning property (Hagan *et al*., 1996).

Computational property. Neural networks are built of neurons or nodes, which are simple processing elements. Each neuron contains a summation node and often a nonlinear sigmoidal activation function of the form

$$f(n) = \frac{1}{1+e^{-\lambda n}}$$

where n=**Wp** is the output from a summation node; λ is the steepness of the activation function; **W** is a weight matrix and **p** is an input vector. Because a single neuron has a limited capability, neurons (sometimes hundreds) are organized in layers and are interconnected between layers using connections called weights. Each weight carries a numerical value that represents the strength of connection or expresses the relative importance of each input to the neuron.

Architecture. Neural networks come in several architectures. One of the most common architectures used in financial applications is a two-layer feed-forward network with error back-propagation. This network typically has only two layers, a hidden layer and an output layer. Signals propagate through these layers from input to output.

Learning. Neural networks use a variety of learning modes. These are supervised learning, unsupervised learning, and reinforcement learning. During supervised learning, which is the most common for the mentioned feed-forward networks, weights are initialized at small random values and training patterns are presented to the network one pattern at a time. The output produced by the training pattern is compared with the actual response provided by a teacher. The differences modify the weights of the network to make them closer to the actual output. This process is repeated for all training patterns contained in a training set until the cumulative error between the actual outputs and the network's output is reduced to a small value. Weights are crucial to the operation of the neural network because through their repeated adjustment the neuron (or network) learns. Knowledge of the network is encoded in its weights.

Neural networks can be implemented as simple computer programs that build models or relationships from data by trial and error. Software that implements different learning architectures and learning algorithms is widely available on the market and became an integral part of the software packages used for data mining. The most attractive features of these networks are their ability to adapt, generalize, and learn from training patterns. These features are not present in modern conventional computers whose processing is based on precise algorithms converted to computer programs. One of the main drawbacks of neural networks is the fact that they produce black box outcomes. In other words, the results generated by a network cannot easily be explained or converted to if-then rules. Although neural networks can approximate very complex nonlinear functions, the explicit equation of these functions that the network learns to classify data is unknown.

3.3. Logit Regression Model

The purpose of the logistic regression model is to obtain a regression equation that could predict in which of two or more groups an object could be placed (i.e. whether a loan should be classified as a good loan or a bad loan). The logistic regression also attempts to predict the probability that a binary or ordinal target will acquire the event of interest (e.g. loan payoff or loan default) as a function of one or more independent variables (i.e. amount of loan, borrower job category, reason of loan). The logit model is represented by the logistic response function *P(y)* of the form:

$$P(y) = \frac{1}{1+e^{-z}}, \text{ where } \quad z = b_0 + \sum_{i=1}^{m} b_i x_i .$$

The function *P(y)* describes a dependent variable *y* containing two or more qualitative outcomes. *z* is the function of *m* independent variables *x* called predictors, and *b* represents the parameters. The *x* variables can be categorical or continuous variables of any distribution. The value of *P*(*y*) that varies from 0 to 1 denotes the probability that a dependent variable *y* belongs to one of two or more groups. The principal of maximum likelihood can commonly be used to compute estimates of the *b* parameters. This means that the calculations involve an iterative process of improving approximations for the estimates until no further changes can be made. (For a more detailed explanation see Afifi and Clark, 1990; Manly, 1994; Christensen, 1997).

Unlike neural networks, logistic regression models are designed to predict one dependent variable at a time. On the positive side, one can note that logistic regression output provides statistics on each variable included in the model. Researchers then can analyze these statistics to test the usefulness of specific information.

4. Data Set Used In The Study

Because the amount and quality of useful credit assessment data is limited, researchers have used data derived from various loan-granting contexts. Desai *et al*. (1996) used 18 predictor variables and three data sets of about 900 observations describing the ordinary customers of three credit unions. West (2000) used German credit scoring data that contained 24 predictor variables and consisted of 700 examples of creditworthy applicants and 300 examples of non credit worthy applicants. The Australian scoring data (Quinlan, 1987) were similar but more balanced with 307 and 383 examples of each outcome.

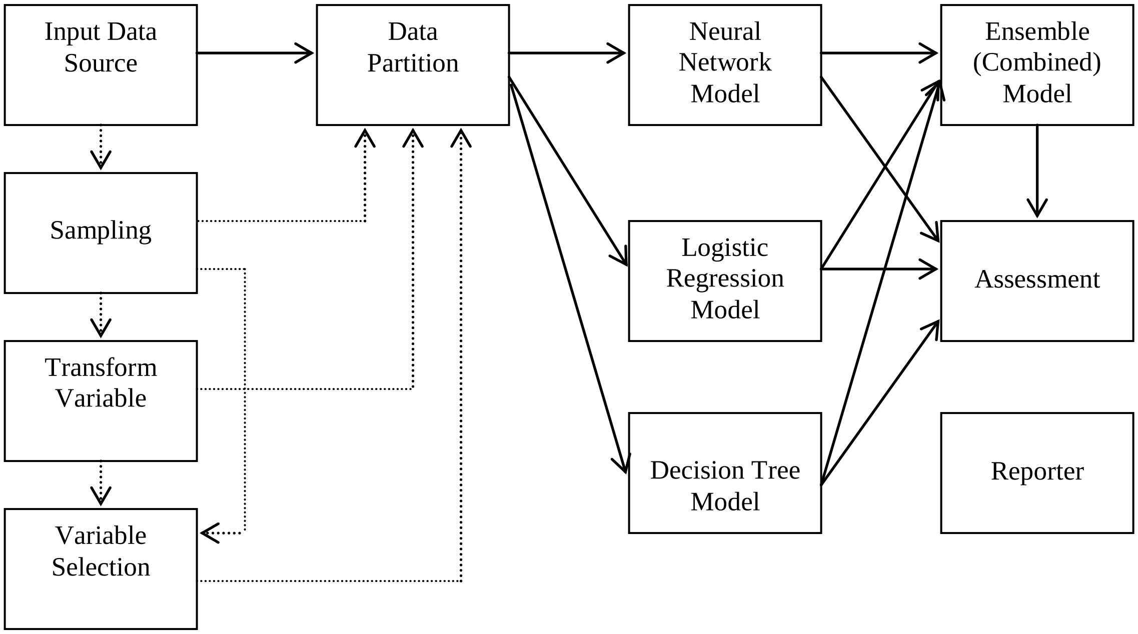
In our study we used qualitative variables that are similar to the ones used in the previous studies. We used a sample data provided by a money lending institution. The data set contains financial information about 3364 consumers allocated among 13 variables. The financial institution extended loans to all of the applicants in the data set. Out of the 3364 applicants, 3064 had paid off their loans and 300 defaulted on the loans, resulting in a default rate of approximately 9%. There were no missing values in the data set. Out of these 13 variables, there were 12 independent variables (loan and consumer characteristics) and one dependent/target variable (loan default or loan repaid) that we were going to predict. Using this data set, we built a decision tree model, neural network model, logistic regression model, and combined (ensemble) model to predict whether a future applicant will default on a loan. The data set contains the following variables:

1. Loan_Status takes values of 1 (client defaulted on loan or seriously delinquent) or 0 (loan repaid) - target variable.
2. Loan_Amt - Amount of the current loan request.
3. Mort_Due - Amount due on existing mortgage.
4. Prop_Value - Value of current property.
5. Debt_To_Inc - Debt-to-income ratio.
6. Years_On_Job - Years on current job.
7. Derog_Rep - Number of major derogatory credit reports.
8. Crd_Ln - Number of credit lines.
9. del_crd_ln - Number of delinquent credit lines.
10. Crd_Inq - Number of recent credit inquiries.
11. Age_Trd_Ln - Age (in months) of oldest trade line.
12. Reason_For_Ln - Reason for loan (debt consolidation or home improvement).
13. Job_Cat - Applicants' job categories.

5. Description Of The Nodes Used For Data Analysis

For data analysis we used the release 8.2 of the SAS Enterprise Miner software. Flowchart 1 presents the major nodes used in the workflow. In the Input Data Source node, we specified the name of a data set and details about the variables in the data set to be used as input for later processing. For one of the simulations, the Sampling node performed stratified random sampling to balance the data set so that it contained an equal number of observations representing good loans and bad loans. The Variable Selection node identified variables that are useful for predicting the target variable. We used an R-square selection criterion. The Transform Variable transformed variables and allowed us to create new ones. Using the Data Partition node, we partitioned the data set into training, validation, and test data subsets. The training subset was used for preliminary model fitting; the validation subset was used to tune the model parameters (i.e., neural networks weights) during estimation; the test subset was used for model final assessment.

The Neural Network, Decision Tree, and Regression nodes enabled us to build various classification/prediction models. Using the Neural Network node, we constructed, trained, and validated a multilayer feed-forward network with error back-propagation. Here we used a neural network with 3 neurons in a hidden layer. The number of neurons in the hidden layer was determined experimentally from the number of observations in the data set and the number of weights in the network (Berry and Linoff, 1997). We used the Decision Tree node to classify observations by segmenting the data created according to a series of simple rules. We used the entropy gain reduction method to build the tree. The Regression node fitted the logistic regression model to the data. The Ensemble node combined the three models by averaging the posterior probabilities for the class target variable (BAD). We used the Assessment and Reporter nodes to provide a common framework for assessing, comparing, and assembling the results from the four models used.



Flowchart 1. Simplified workflow of the nodes used for data analysis. The dotted lines indicate optional nodes used in experiments two and three only.

6. Experiments And Results

We performed an extensive computer simulation divided into two scenarios. In the first scenario, we used the original, unbalanced data set containing a total of 3364 customers. This first data set contained 3064 good loans and 300 bad loans. In the second scenario, we created a balanced data set by randomly selecting 300 loans from the 3064 good loans and matching them with the 300 bad loans. This random sampling produced the second data set consisting of 600 cases divided evenly among good and bad loans.

In each scenario, we performed three different experiments. In each experiment, we allocated the cases as follows: 60% for training, 20% for validation, and 20% for testing. In experiment one, we used the data set without any variable transformations or variable reduction. In experiment two, we used the original data set without any variable transformations, but we eliminated the variables that were weakly correlated with the target variable. In experiment three, we used the original number of 13 variables, but performed variable transformations-- such as bucketing for four variables that had highly skewed distributions-- in order to improve the model. Furthermore, in each experiment, we employed three different data mining tools: neural networks, decision trees, and logistic regression. Finally, we combined the three tools into an ensemble model to increase the reliability of the classification accuracy by improving the stability of the three disparate non-linear models. The ensemble model averages the posterior probabilities for class target variable BAD from the three tools. Given the posterior probabilities, each case can be classified into the most probable class.

Table 1

	Experiment One	**Experiment Two**	**Experiment Three**
Decision tree “Overall”	93.3% (628/673)	93.6% (630/673)	93.2% (627/673)
“Good”	99.0% (608/614)	98.5% (605/614)	99.5% (611/614)
“Bad”	33.9% (20/59)	42.4% (25/59)$^{\#\# \$\$}$	27.1% (16/59)$^{\$\$}$
Neural network "Overall"	94.0% (633/673)	93.3% (628/673)	93.8% (631/673)
"Good"	99.0% (608/614)	99.3% (610/614)	99.7% (612/614)
"Bad"	42.4% (25/59)$^{** \$}$	30.5% (18/59)$^{\$}$	32.2% (19/59)*
Logistic regression "Overall"	93.2% (627/673)	93.3% (628/673)	99.6% (623/673)
"Good"	99.7% (612/614)	99.8% (613/614)	99.5 (611/614)
"Bad"	25.4% (15/59)**	25.4% (15/59)$^{\#\#}$	20.3 (12/59)*
Ensemble model "Overall"	93.9% (632/673)	93.6% (630/673)	93.3% (628/673)
"Good"	99.8% (613/614)	100.0% (614/614)	100.0% (614/614)
"Bad"	32.2% (19/59)	27.1% (16/59)$^{\#\#}$	23.7% (14/59)

Table I. Classification rates for the four models and three experiments used in scenario one. It shows the percentage and number of test cases classified correctly. ** Neural network used in experiment one classifies bad loans significantly better than logistic regression at α=0.05. $^{\#\#}$ Decision tree used in experiment two classifies bad loans significantly better than logistic regression at α=0.05. * Neural network used in experiment three classifies bad loans significantly better than logistic regression at α=0.1. $^{\$}$ Neural network used in experiment one classifies bad loans significantly better than the one used in experiment two at α=0.1. $^{\$\$}$ Decision tree used in experiment classifies bad loans significantly better than the one used in experiment three at α=0.05.

Table I presents the classification accuracy for the unbalanced test data set containing 673 observations (614 good loans and 59 bad loans). This test set represents 20% of the cases which were extracted from the unbalanced data set composed of 3064 good loans and 300 bad loans (first scenario). Because the training sample in the data set was overwhelmingly dominated by good loans, (1) the classification accuracy of good loans was almost perfect (close to 100%), (2) some bad loans were recognized as bad loans, and (3) the three methods and the ensemble method tended to classify a substantial amount of bad loans as good loans. Although the overall classification rate is excellent and averaged about 93%, the average classification accuracy of bad loans, which were underrepresented in the training sample, is relatively low and amounts to an average of about 30% for all the methods across all experiments. The neural network used in experiment one is the most efficient in the correct classification of bad loans (42.4%) and logistic regression seems to be the worst (25.4%). Based on the data in Table I, we performed several 2-tailed proportional z-tests to determine whether the classification rates across the four models and the three experiments are significantly different from one another. The test revealed that there are statistically significant differences in classification accuracy rates, especially for bad loans (see Table I for details).

The performance of the four methods can be best evaluated using a combination of the percent cumulative and non-cumulative response lift charts, and the receiver characteristics response charts. All of these charts reflect the performance of the four data mining tools on the test data sets.

To properly interpret the cumulative percent response chart (Figure 1), one needs to consider how the chart is constructed. If the response of interest is loan defaults, a respondent is defined as an individual who defaults on a loan (BAD=1). For each individual, the fitted models predict the probability that the individual will default. The observations are sorted by the predicted probability of response from the highest probability of response to the lowest probability of response, and then grouped into ordered bins, each containing approximately 10% of the data. Using the target variable BAD, one can count the percentage of actual respondents in each bin. If the model is effective, the proportion of individuals with the event level that one is modeling (here those who defaulted on a loan) will be relatively high in bins in which the predicted probability of response is high. One sees that the first 10% decile contains an enormous amount of defaulters -- the four models predict that between 37% and 47% of customers in the first decile will default on their loans (Figure 1). The ensemble model seems to be slightly better than the neural network, with 48% and 45% defaults in the top 10%, respectively. The horizontal 9% response line represents the baseline rate of about 9% of defaulters that one would expect if one were to take a random sample. In a cumulative % response chart, the response rate for each decile of the score includes all of the responses for the deciles above it.

Figure 1. The test set cumulative % response chart for the four methods for experiment one for scenario one. Target event: 1 (bad loans).

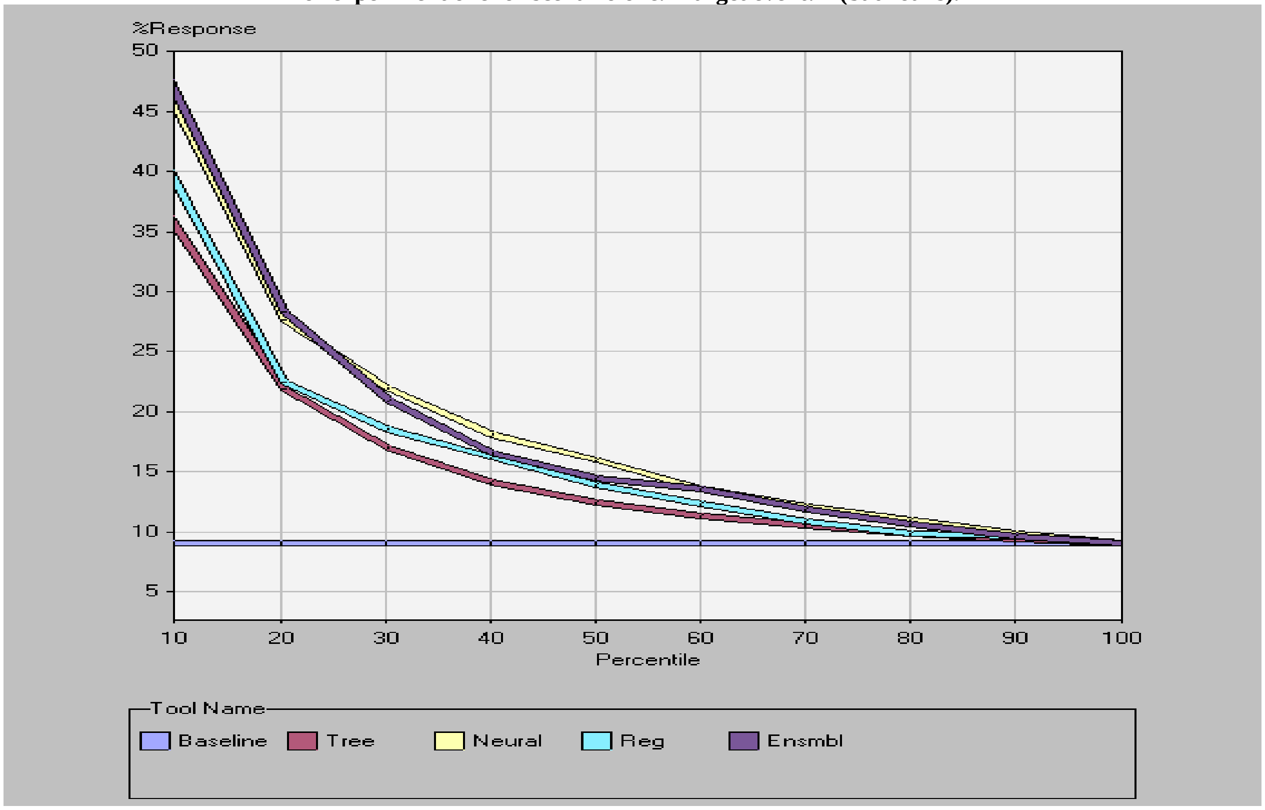


Figure 2. The test set non-cumulative % response chart for the four methods for experiment one for scenario one. Target event: 1 (bad loans).

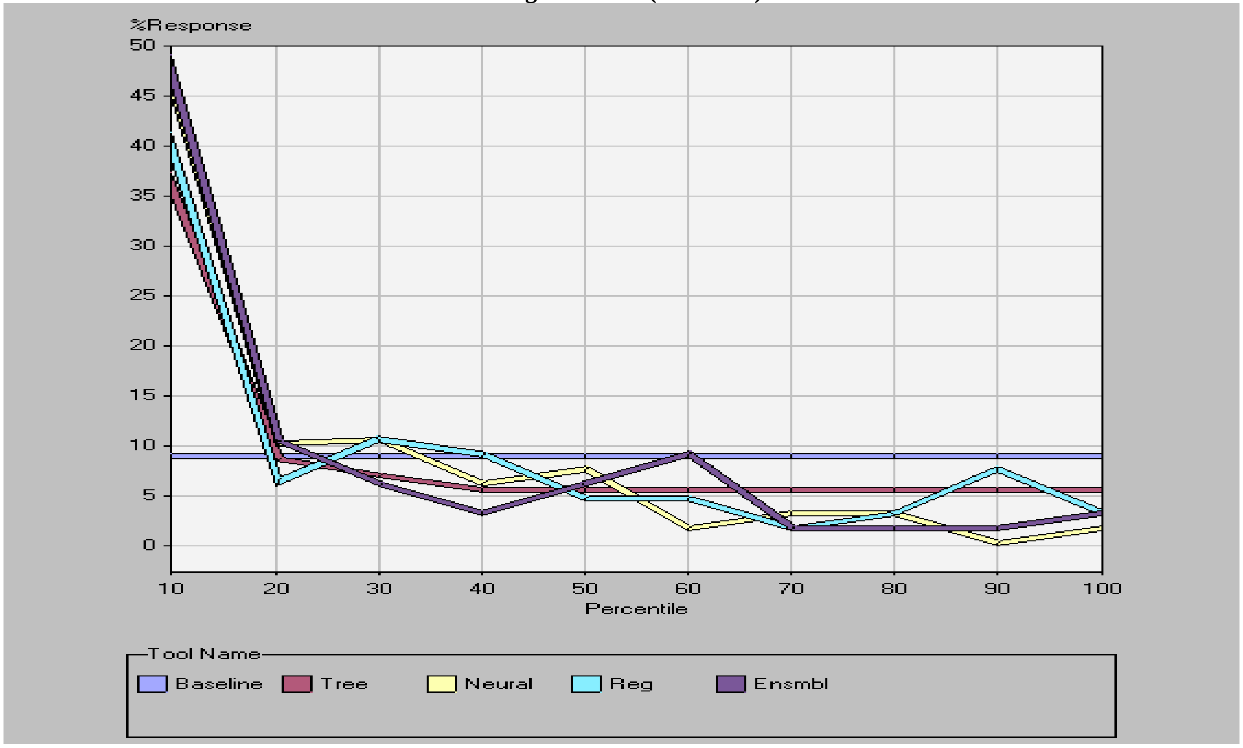


Figure 3. The cumulative lift chart for the four methods for experiment one for scenario one. Target event: 1 (bad loans).

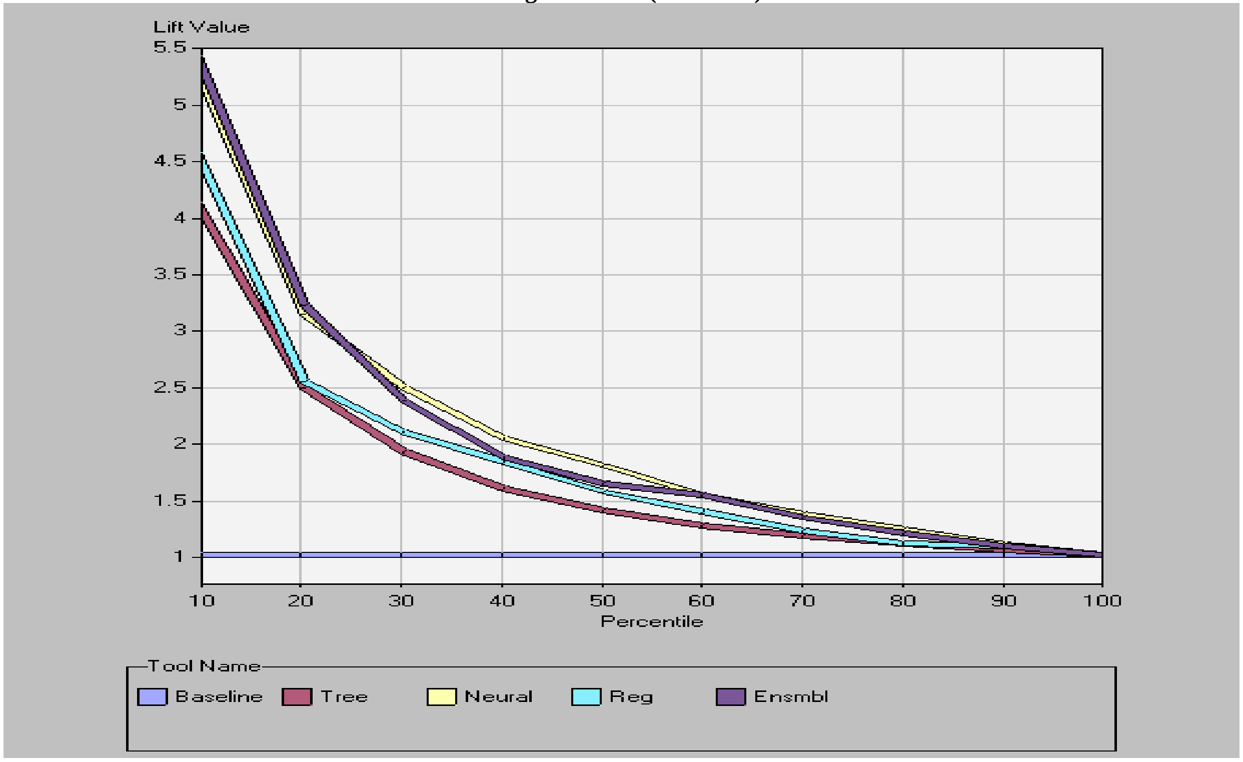
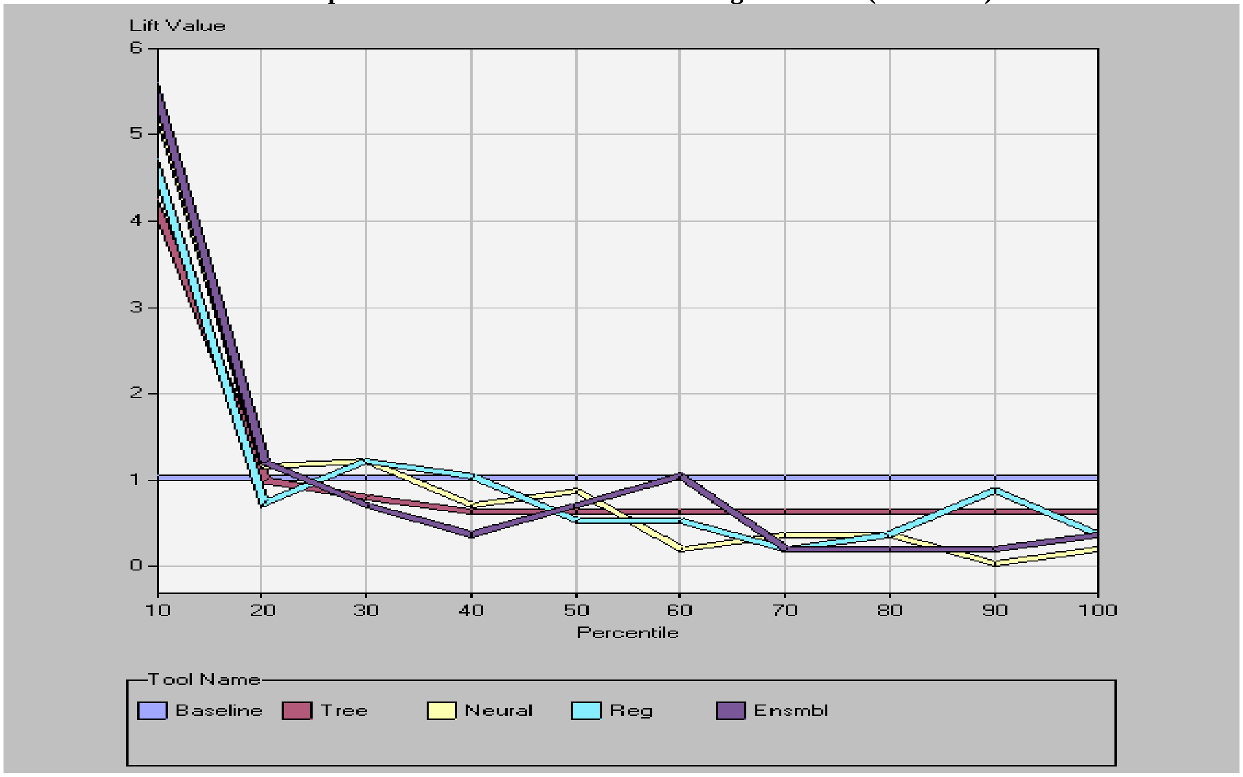


Figure 4. The non-cumulative lift chart for the four methods for experiment one for scenario one. Target event: 1 (bad loans).



A cumulative response chart shows the overall strength of the model, but it obscures the model's performance at each strata of the score. The test set non-cumulative percent response chart (Figure 2) shows the percentage of defaulters in each decile. The chart reveals that the percentage of defaulters declines very significantly in the second decile to about 6% to 11%, continues to decrease slightly between the second and seventh decile to about 3% and 6%, and remains at that level between the seventh and tenth decile. According to the model, the customers between the fourth and tenth deciles are good because their projected default rate is below the baseline rate, while the last three deciles contain the very best customers with the lowest default rate.

Figures 3 and 4 depict the cumulative and non-cumulative lift charts that plot the same loan default information on a different scale. The cumulative lift chart (Figure 3) shows that for the neural network and ensemble models, the percentage of defaulters in the first decile is over five times higher than the 9% default rate in the population. For example, for the ensemble model the lift value is about 5.3 (48% of defaulters / 9% baseline rate). The lift values drop significantly from the second decile on.

The receiver operating characteristic (ROC) charts are graphical displays that give the global measure of the predictive accuracy of the models (Figures 5 and 6). They display the sensitivity against 1-specificity of a classifier for a range of cutoffs. Sensitivity is a measure of accuracy for predicting events that is equal to the true positive divided by total actual positive. 1-specificity is a measure of accuracy for predicting nonevents that is equal to the true negative divided by total actual negative. Each point on the curves represents a cutoff probability. Points closer to the upper-right corner correspond to low cutoff probabilities. Points in the lower left correspond to higher cutoff probabilities. The extreme points (1,1) and (0,0) represent no-data rules where all cases are classified into class 1 or class 0, respectively. The performance quality of the models is demonstrated by the degree to which the ROC curves push upward and to the left. The area under the curves can provide a quantitative performance measure. The area will range from 50, for a worthless model, to 100, for a perfect classifier. The shapes of the ROC charts indicate that the predictive power all four models (decision tree, neural network, regression analysis, and combined model) for predicting bad and good loans is fairly good (Figures 5 and 6). Both the ensemble model and the neural network model stand out and appear to be better than the decision tree model and regression model at predicting both bad and good loans.

Figure 5. Receiver operating characteristics (ROC) chart: Target event = 1 (bad loans).

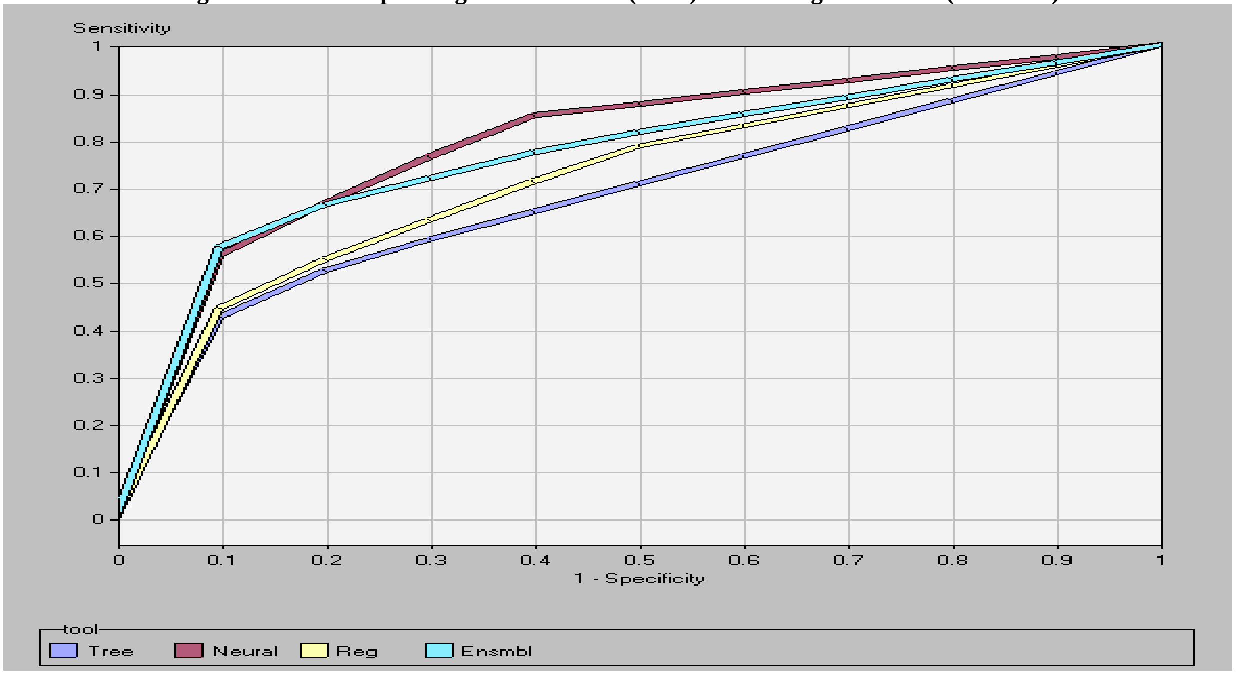


Figure 6. Receiver operating characteristics (ROC) chart: Target event = 0 (good loans).

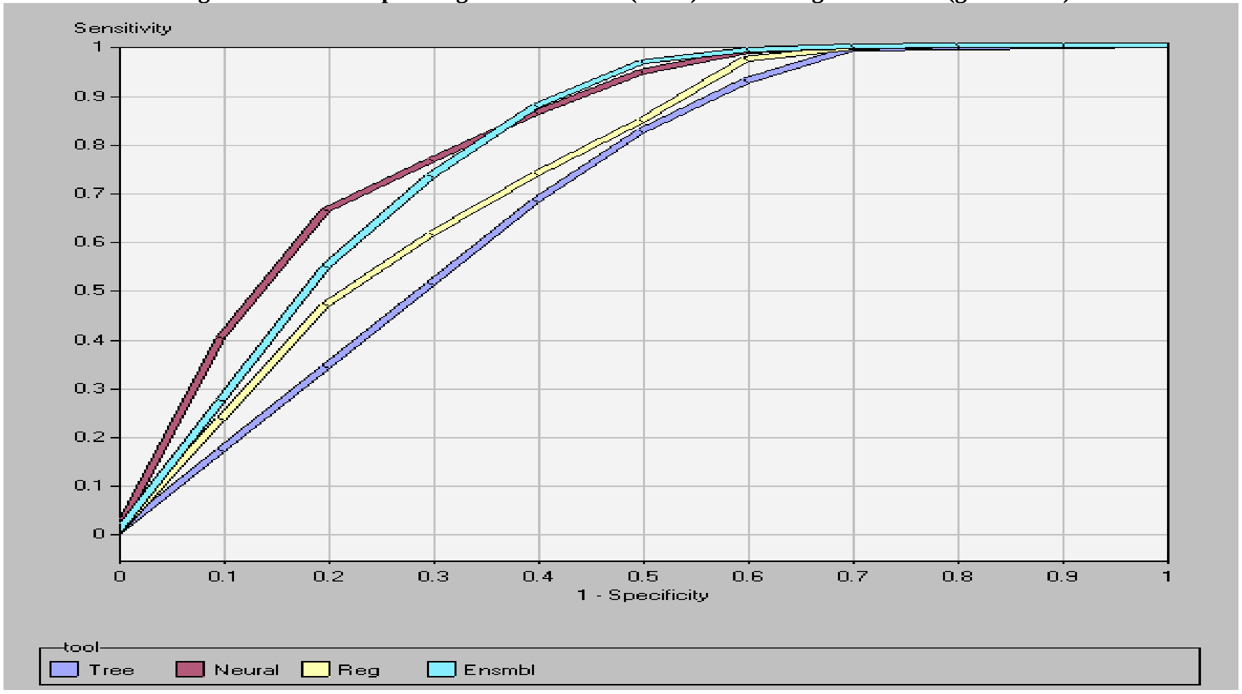


Table II. Classification rates for the four models and three experiments used in scenario two. It shows the percentage and number of test cases classified correctly. ** In experiment two ensemble model classified good loans significantly better than logistic regression at α=0.05. * In experiments one and two ensemble model's overall classification rate was significantly better than logistic regression at α=0.1. ## In experiment one decision tree classified good loans significantly better than logistic regression at α=0.05.

	Experiment One	**Experiment Two**	**Experiment Three**
Decision tree			
“Overall”	76.7% (92/120)	76.7% (92/120)	76.7% (92/120)
“Good”	83.6% (51/61)##	83.6% (51/61)	83.6% (51/61)
“Bad”	69.5% (41/59)	69.5% (41/59)	69.5% (41/59)
Neural network			
"Overall"	72.5% (87/120)	75.8% (91/120)	71.7% (86/120)
"Good"	73.8% (45/61)	77.1% (47/61)	72.1% (44/61)
"Bad"	71.2% (42/59)	74.6% (44/59)	71.2% (42/59)
Logistic regression			
"Overall"	70.8% (85/120)*	72.5% (87/120)*	70.0% (84/120)
"Good"	68.9% (42/61)##	72.1% (44/61)**	72.1 (44/61)
"Bad"	72.9% (43/59)	72.9% (43/59)	67.8 (40/59)
Ensemble model			
"Overall"	79.2% (95/120)*	80.0% (96/120)*	75.0% (90/120)
"Good"	78.7% (48/61)	85.2% (52/61)**	80.3% (49/61)
"Bad"	79.7% (47/59)	74.6% (44/59)	72.9% (41/59)

Table II shows the classification rates for the test set containing 120 observations of the second scenario in which the good and bad loans are almost equally represented. Although the average overall classification accuracy, and the classification accuracy of good loans, dropped to about 74.8% and 77.6%, respectively, across all experiments and methods, the average classification accuracy of bad loans improved significantly to about 72.2% compared to Table I. It is worth noting that the ensemble model for experiment one yields the best classification accuracy for bad loans (79.7%). Generally, the regression model performed the worst. Two-tailed proportional z-tests found some significant differences in the classification accuracy rates between the regression and ensemble models. The data in Table II reveal that the combined tool is effective in classifying bad loans as well as good loans. For example, in experiment two, the combined tool correctly classified 85.2% of good loans. Because one can extract simple and understandable if-then rules from a decision tree and a money lending institution has to explain the reasons for which the loan was denied to the customer, the output from the combined tool may be complemented by the rules generated by the decision tree created for experiment one. Some of the rules based on the training data set generated by this decision tree are as follows:

1. IF (Debt-to-income ratio ≥ 42.625) THEN (The loan good and bad 6.2% and 93.8% of the time, respectively).
2. IF (Value of current property < $47,546.50) AND (Number of delinquent trade lines < 0.5) AND (Debt-to-income ratio < 42.625) THEN (The loan was good and bad 20.8% and 79.2% of the time, respectively).
3. IF (Years on Current Job < 14.5) AND (Number of delinquent trade lines < 0.5) AND (Debt-to-income ratio < 42.625) THEN (The loan was good and bad 21.2% and 78.8% of the time, respectively).
4. IF (Age of oldest trade line in months > 86.51) AND (Value of current property > $47,546.5) AND (Number of delinquent trade lines < 0.5) AND (Debt-to-income ratio < 42.625) THEN (The loan was good and bad 80.6% and 19.4% of the time, respectively).

However, it may be especially difficult to explain a denial of a loan if the decision tree model identifies the loan as a good loan, but the other methods identify it as a bad loan.

The cumulative lift charts contained in Figure 7 present Table II figures in a more useful format. The chart reveals that for the neural network, ensemble, and regression models, the percentage of defaulters in the first decile is over two times as high as the 50% default rate in the general population of the second scenario. For example, for the three models, the lift value is about 2.05 and it does not drop rapidly. The test set non-cumulative lift chart (Figure 8) shows the lift value for each decile. The chart reveals that the predicted power of four models drops to the 50% baseline (lift 1) between the fourth and the fifth deciles. By the fifth decile the models permanently drop below the baseline's 50% default baseline and remain at roughly 0.5 lift value for the last four deciles. The percent captured response chart (Figure 9) reveals the total number of defaulters in a particular bin. For example, if one wanted to eliminate the first and second deciles of customers, the total number of defaulters would drop by about 35-40%. The straight response line represents the response rate that one would obtain by not using a model (or the rate of defaulters that one would expect if one were to take a random sample). For example, if one takes a random sample of 20% of the data, one would expect to capture 20% of the defaulters. The choice of the final model(s) may depend on the proportion of individuals that one has chosen for action. When comparing several models on the same proportion of the data, the model with the higher lift is often preferred. The shapes of the ROC charts for the first scenario (Figures 5 and 6) and second scenario (Figures 10 and 11) indicate that the predictive power all four models (decision tree, neural network, regression analysis, and combined model), for differentiating between loans that will be paid off and loans that customer will default on, is good.

Figure 7. The cumulative lift chart for the four methods for experiment two for scenario two. Target event: 1 (bad loans).

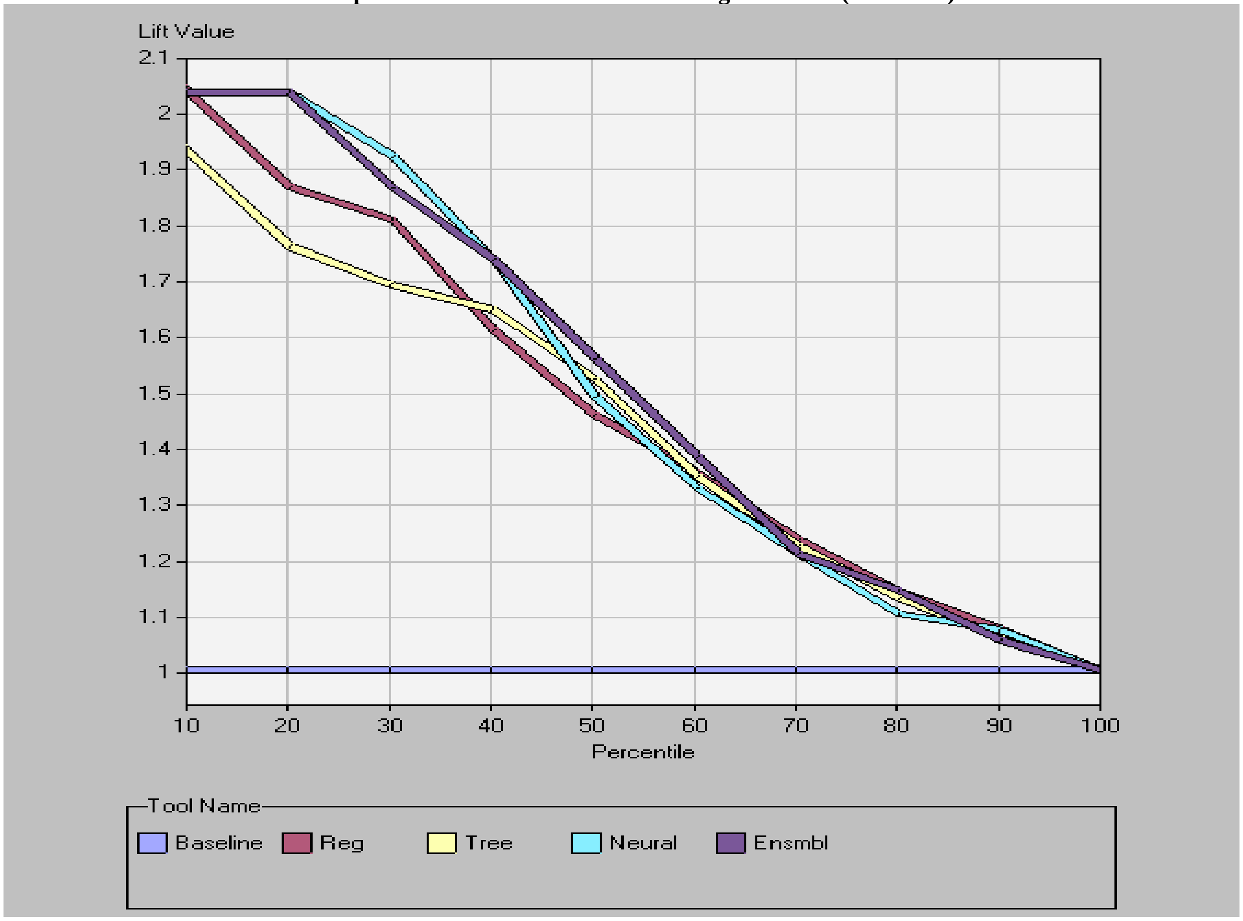


Figure 8. The non-cumulative lift chart for the four methods for experiment two for scenario two. Target event: 1 (bad loans).

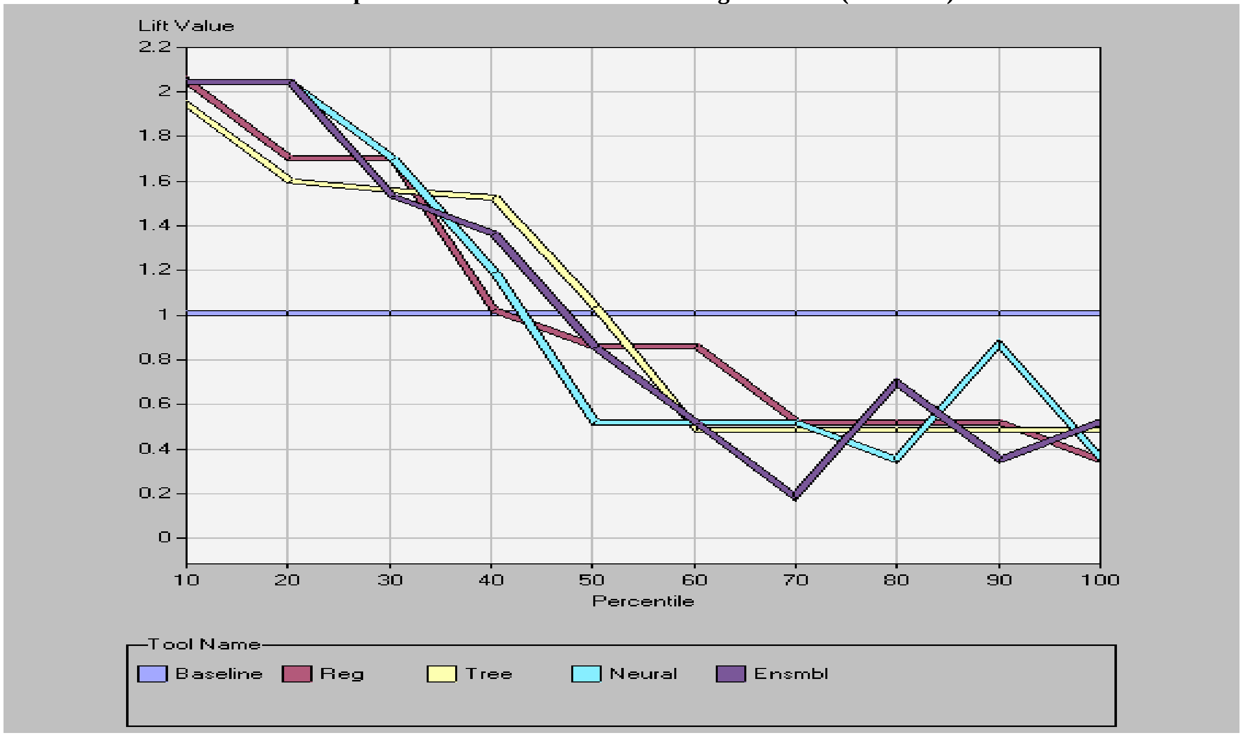


Figure 9. The % captured response chart for the four methods for experiment one for scenario one. Target event: 1 (bad loans).

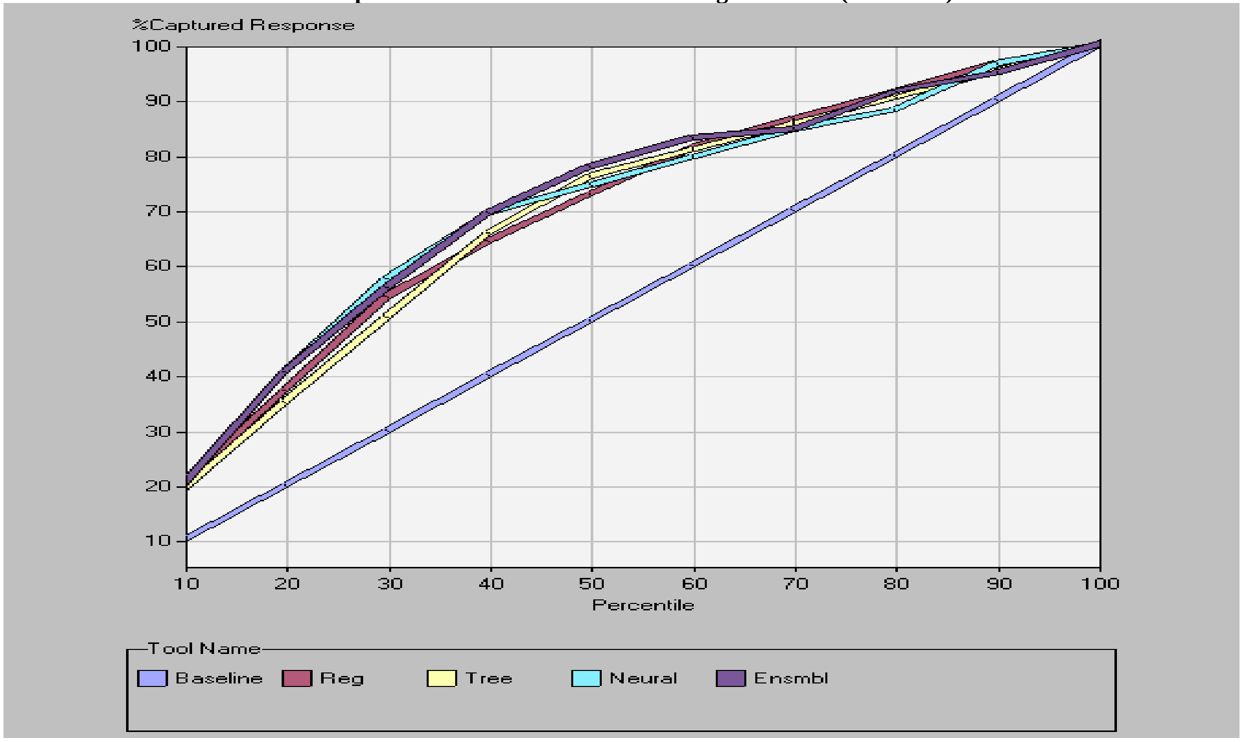


Figure 10. Sensitivity as a function of 1-specificity for the four methods for experiment two for scenario two. Target event - 1 (bad loans).

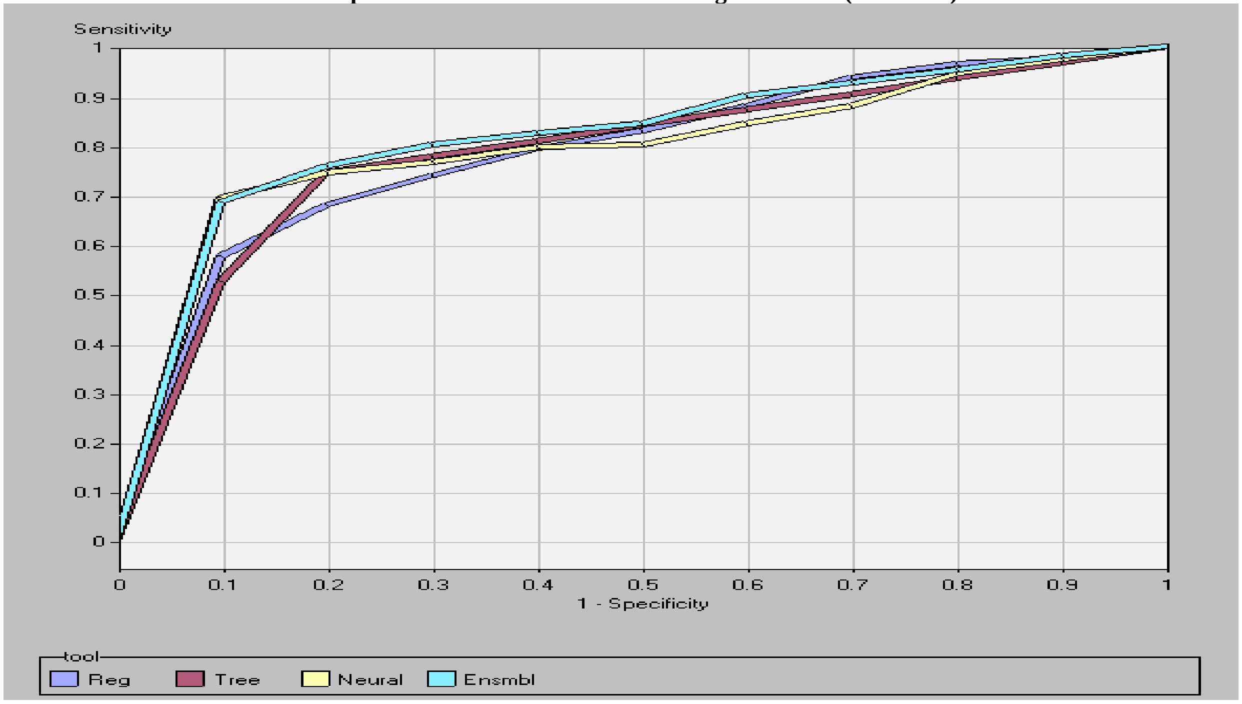
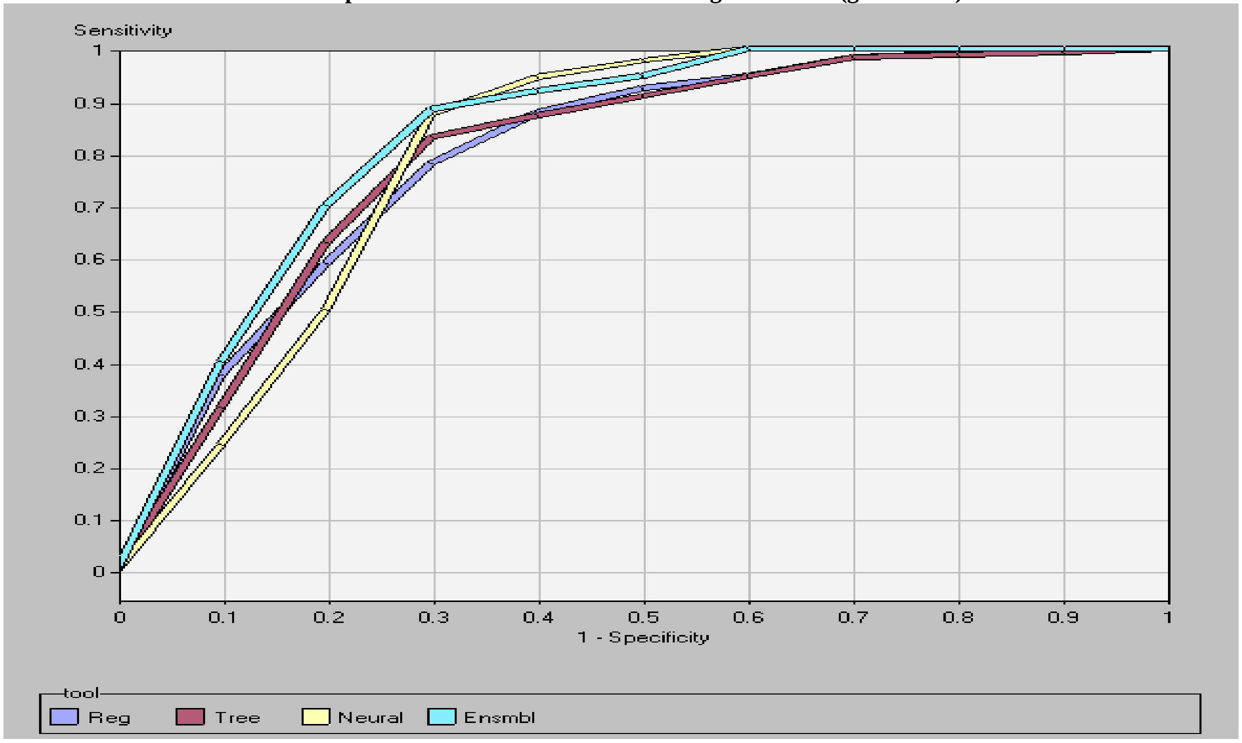


Figure 11. Sensitivity as a function of 1-specificity for the four methods for experiment two for scenario two. Target event - 0 (good loans).



7. Conclusion

Of the four models used in the two scenarios, the neural network and the ensemble model were the best overall in classifying loans into good and bad loans. In the second scenario, the four models proved efficient at identifying bad loans because they were trained and tested on a balanced data set containing an equal mix of good and bad loans. Conversely, the four models in the first scenario were trained and tested on a more realistic unbalanced data set containing mostly good loans. As a result, in the first scenario, the four models were superb at identifying the good loans but not nearly as good at identifying the bad loans. However, one must keep in mind that all of the bad loans in the data set were close cases that were initially misclassified as good loans by the financial institution providing the data. In practical application, faced with a mixture of clear-cut cases and close cases, the four models are likely to identify a much higher percentage of bad loans. If the financial institution providing the data were to use these data mining models developed in the first scenario, the financial institution would be able to identify at least 30% of the bad loans and deny credit to those applicants, thus reducing the default rate from 9% to less than 7% without a noticeable reduction in the number of good loans granted. However, a group of consumers characterized by a high default rate may contain many close calls between good and bad loans making the classification process very difficult and ultimately insufficient to protect against defaults. In addition to predicting whether a loan is good or bad, the data mining techniques also provided valuable information that would allow the financial institution to adjust its lending policies to avoid groups of high risk consumers. If the results of the first stimulation were to be used, the financial institution would be wise to refuse loans to the bottom 10% of its customers because roughly one-third to one-half of these customers will default. This first decile contains between one-third and five-ninths of the total number of defaulters in the data set. By refusing credit to the bottom 10% of the customers, the financial institution could reduce the default rate by 3 to 5 percentage points at the cost of losing between 5-7% of the total number of non-defaulting customers. The customers in the 20 through 69 percentile have a default rate at or below the 9% average and appear to be credit worthy. The financial institution could extend loans to these customers on standard terms. The customers in the 70 to 99 percentiles constitute the financial institution's prime customers because of their very low default rate. The financial institution could reward these customers by issuing loans to them on more favorable terms. The classification properties of the data mining tools presented in this article are very useful in identifying good and bad loans. Statistical information derived from these data mining techniques compliments the classification process and could allow financial institutions to fine tune their lending policies to avoid high risk groups that cannot be classified with great accuracy. Further research should focus on refining the training and testing of the ensemble model and the neural network using various balanced and unbalanced data sets to improve the classification performance. Once the data mining techniques are fine tuned, one could attempt 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. 🕮

8. References

1. Adrianns, P, and Zantinge, D., 1996, *Data Mining*, Addison-Wesley Longman Limited, Harlow, England.
2. Adya, M., and Collopy, F., 1998, How Effective are Neural Networks at Forecasting and Prediction? A Review and Evaluation, *Journal of Forecasting*, Vol. 17, 481-495.
3. Afifi, A.A, and Clark, V., 1990, *Computer-Aided Multivariate Analysis*, Van Nostrand Reinhold Co., New York.
4. Altman, E.I., Marco, G., and Varetto, F., 1994, Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks (the Italian Experience), *Journal of Banking and Finance*, Vol. 18, No. 3, pp. 505-529.
5. Barney, D.K., Graves, O.F., and Johnson, J.D., 1999, The Farmers Home Administration and Farm Debt Failure Prediction, *Journal of Accounting and Public Policy*, Vol. 18, pp. 99-139.
6. Back B., Laitinen, T., and Sere, K., 1996, Neural Networks and Genetic Algorithms for Bankruptcy Predictions, *Expert Systems with Applications*, Vol. 11, No. 4, pp. 407-413.
7. Berry, M.J.A., and Linoff, G.S., 1997, *Data Mining Techniques for Marketing, Sales, and Customer Support*, John Wiley & Sons, Inc.
8. Boritz, J.E., and Kennedy, D.B, 1995, Effectiveness of Neural Network Types for Prediction of Business

Failure, *Expert Systems with Applications*, Vol. 9, No. 4, pp. 503-512.

9. Christensen, R., 1997, *Log-Linear Models and Logistic Regression*, Springer, New York.
10. Coats, P.K., and Fant, F.L., 1993, Recognizing Financial Distress Patterns Using a Neural Network Tool, *Financial Management*, Vol. 22, September, 142-150.
11. 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, 24-37.
12. Dimitras, A.I., Slowinski, R., Susmaga, R., and Zopounidis, C., 1999, Business failure Prediction Using Rough Sets, *European Journal of Operation Research*, Vol. 114, pp. 263-280.
13. Fayyad, U.S., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds.), 1996, *Advances in Knowledge Discovery and Data Mining*, AAAI Press / The MIT Press, Menlo Park, California, USA.
14. Glorfeld, L.W., and Hardgrave, B.C., 1996, An Improved Method for Developing Neural Networks: The Case of Evaluating Commercial Loan Credit Worthiness, *Computer & Operations Research*, Vol. 23, No. 10, pp. 933-944.
15. Greenstein, M.M., and Welsh, M.J., 2000, Bankruptcy Prediction Using *Ex Ante* Neural Network and Realistically Proportioned Testing Sets, *Artificial Intelligence in Accounting and Auditing*, Vol. 6 (forth-coming).
16. Hagan, M.T., Demuth, H.B., and Beale, M., 1996, *Neural Network Design*, PWS Publishing Company.
17. Jagielska, I., Matthews, C., and Whitfort, T., 1999, 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.
18. Jain, B.A., and Nag, B.N., 1997, Performance Evaluation of Neural Network Models, *Journal of Management Information Systems*, Vol. 14, No. 2, pp. 201-216.
19. Jo, H., and Han, I., 1997, Bankruptcy Prediction Using Case-Based Reasoning, Neural Networks, and Discriminant Analysis, *Expert Systems with Applications*, Vol. 13, No. 2, pp. 97-108.
20. Koh, H.C., and Tan, S.S., 1999, A Neural Network Approach to the Prediction of Going Concern Status, *Accounting and Business Research*, Vol. 29, No. 3, pp. 211-216.
21. Kumar, N., Krovi, R., and, Rajagopalan, B., 1997, Financial Decision Support with Hybrid Genetic and Neural Based Modeling Tools, *European Journal of Operation Research*, Vol. 103, 339-349.
22. Kumar A., and Olmeda, I., 1999, A Study of Composite or Hybrid Classifiers for Knowledge Discovery, INFORMS Journal on Computing, Vol. 11, No. 3, pp. 267-277.
23. Lacher, R.C., Coats, P.K, Sharma, S.C, and Fant, L.F, 1995, A Neural Network for Classifying the Financial Health of a Firm, *European Journal of Operational Research*, 85, pp. 53-65.
24. Lee, K.C., Han, I., and Kwon, Y., 1996, Hybrid Neural Network for Bankruptcy Predictions, *Decision Support Systems*, 18, pp. 63-72.
25. Manly, B.F.,1994, *Multivariate Statistical Methods: A Primer*, Chapman & Hall, London.
26. Mitchell, T.M., 1997, *Machine Learning*, WCB/McGraw-Hill, Boston, Massachusetts.
27. McLeod, R.W., Malhotra, D.K., and Malhotra, R., 1993, Predicting Credit Risk: A Neural Network Approach, *Journal of Retail Banking*, Vol. XV, No. 3, pp. 37-40.
28. Olmeda, I., and Fernandez, E., 1997, Hybrid Classifiers for Financial Multicriteria Decision Making: The Case of Bankruptcy Prediction, *Computational Economics*, Vol. 10, pp. 317-335
29. Piramuthu, S., 1999, Financial Credit-Risk Evaluation with Neural and Neurofuzzy Systems, *European Journal of Operational Research*, Vol. 112, 310-321.
30. Piramuthu, S., 1999, Feature Selection for Financial Credit-Risk Evaluation Decisions, INFORMS Journal on Computing, Vol. 11, No. 3, pp. 258-266.
31. Rosenberg, E., and Gleit, A., 1994, Quantitative Methods in Credit Management: A Survey", *Operations Research*, 42(4), pp. 589-613.
32. Quinlan, J.R., 1987, Simplifying Decision Trees, *International Journal of Man-Machine Studies*, Vol. 27, pp. 221-234.
33. Quinlan, J.R., 1993, *C4.5: Programs for Machine Learning*, Morgan Kaufman Publishers, San Mateo, California.
34. Tam, K.Y., and Kiang, M.Y., 1992, Managerial Applications of Neural Networks: The Case of Bank Failure Predictions, *Management Science*, Vol. 38. No. 7, pp. 926-947.
35. Tessmer, A.C., 1997, What to Learn from Near Misses: An Inductive learning Approach to Credit Risk

Assessment, *Decision Sciences*, Vol. 28, No. 1, pp. 105-120.

36. Thomas, L.C., 2000, A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Consumers, *International Journal of Forecasting*, Vol. 16, 149-172.
37. West, D., 2000, Neural Network Credit Scoring Models, *Computers & Operations Research*, Vol. 27, 1131-1152.
38. Wilson, R.L., and Sharda, R., 1994, Bankruptcy Prediction Using Neural Networks, *Decision Support Systems*, Vol. 11, pp. 545-557.
39. Yang, B., Li, L.X., Ji, H., and Xu, J., 2001, An Early Warning System for Loan Risk Assessment Using Artificial Neural Networks, *Knowledge-Based Systems*, Vol. 14, pp. 303-306.
40. Young, Z.R., Platt, M.B., and Platt, H.D., 1999, Probabilistic Neural Networks in bankruptcy Prediction, *Journal of Business Research*, Vol. 44., pp. 67-74
41. Zhang, G., Hu, M.Y., Patuwo, B.E., and Indro, D.C., 1999, Artificial Neural Networks in Bankruptcy Prediction: General Framework and Cross-Validation Analysis, *European Journal of Operation Research*, Vol. 116, pp. 16-32.
42. Zurada, J., Foster, B.P., and Ward, T.J, 2001, An Investigation of Artificial Neural Networks for Classifying Levels of Financial Distress of Firms: The Case of an Unbalanced Training Sample, in *Knowledge Discovery for Business Information S*ystems, W. Abramowicz and J. Zurada (Eds.), Kluwer Academic Publishers, Boston, USA, pp. 397-424.
43. Zurada, J., 2001, Comparison of the Performance of Several Data Mining Techniques for Loan-Granting Decisions, The Proceedings of the Tenths International Conference on Information Systems Development, England, London, September 2001, (in press).

Notes

Notes