Comparison Of The Performance Of Several Data Mining Methods For Bad Debt Recovery In The Healthcare Industry

Jozef Zurada, University of Louisville
Subhash Lonial, University of Louisville

ABSTRACT

The healthcare industry, specifically hospitals and clinical organizations, are often plagued by unpaid bills and collection agency fees. These unpaid bills contribute significantly to the rising cost of healthcare. Unlike financial institutions, health care providers typically do not collect financial information about their patients. This lack of information makes it difficult to evaluate whether a particular patient-debtor is likely to pay his/her bill. In recent years, the industry has started to apply data mining tools to reduce bad-debt balance. This paper compares the effectiveness of five such tools - neural networks, decision trees, logistic regression, memory-based reasoning, and the ensemble model in evaluating whether a debt is likely to be repaid. The data analysis and evaluation of the performance of the models are based on a fairly large unbalanced data sample provided by a healthcare company, in which cases with recovered bad debts are underrepresented. Computer simulation shows that the neural network, logistic regression, and the combined model produced the best classification accuracy. More thorough interpretation of the results is obtained by analyzing the lift and receiver operating characteristic charts. We used the models to score all "unknown" cases, which were not pursued by a company. The best model classified about 34.8% of these cases into "good" cases. To collect bad debts more effectively, we recommend that a company first deploy and use the models, before it refers unrecovered cases to a collection agency.

INTRODUCTION

The cost of healthcare is rising at an alarming rate. Patients who refuse to pay their healthcare bills contribute to this problem. The healthcare industry, specifically hospitals and clinical organizations, are often plagued by bad-debt, collection agency fees and outstanding medical testing costs. Galloro (2003) reported that Nashville-based HCA's provision for bad debt rose to an astounding 10.3% of its net revenue for the third quarter, compared to an already high 8.3% of revenue in the same quarter the previous year. Even when unpaid bills are eventually paid, hospitals typically end up paying 30% to 50% of recovered bad-debt revenue to outside collection agencies (Veletsos, 2003). Pesce (2003) argues that hospitals should secure the funding for and invest in modern information technology to reduce bad-debts, which along with other factors such as billing errors, insurance underpayments, and inability to collect accurate patient and payer information throughout delivery of care, account for 13% of a hospitals' lost revenue each year. The matter of recovering bad debts has become serious and has even involved hospitals suing patients. A review of court records and interviews with hospital trade groups, collection attorneys and consumer advocates shows that hospitals in several states of the U.S. are beginning to use harsh measures to collect debts and secure the arrest and even imprisonment of patients who miss court hearings related to their healthcare debts (Lagnado, 2003).

Predicting whether a particular customer is likely to repay a healthcare debt is an inherently complex and unstructured process. What makes this process especially difficult in the healthcare context is the hospital's inability to obtain detailed financial information concerning the patients. Unlike a financial institution which would collect information and carefully evaluate whether to extend a loan, healthcare institutions must often admit a patient and perform the necessary medical procedures on credit knowing very little about the particular patient. Thus, due to moral and practical constraints, healthcare providers often become unwilling creditors to a multitude of borrowers. A healthcare institution is handicapped by having only a small number of independent attributes of the patient-debtor to evaluate (Pesce, 2003). In addition, some of debt defaults may be attributed to unforeseen events (i.e. divorce, death, loss of employment) or be governed by factors that may be difficult or impossible to see in the attributes of the consumer (i.e. stability of marriage, general health, job stability). Given all of the difficulties described above it is not surprising that the healthcare institution that provided data for this paper recovered bad debts from only about seven percent (7.3%) of the non-paying patients. One can assume that this healthcare institution used some initial scoring model to target "good" patients who were more likely to repay the bad debts. Any improvement in making a reliable distinction between those who are likely to repay the debt and those who are not would allow the healthcare institution to focus preliminary debt recovery efforts on the "good" patients and save on administrative expenses by either writing off the bad debts or turning them over immediately to a collection agency.

The limited raw data available to the healthcare provider about the patient-debtor may seem to have no apparent relationship to the likelihood that a patient-debtor will repay the bad debt. Fortunately, the beleaguered healthcare provider has several knowledge discovery and data mining tools at its disposal. Knowledge discovery is defined as the process of identifying valid, novel, and potentially useful patterns, rules, relationships, rare events, correlations, and deviations in data (Fayyad *et al*., 1996). This process relies on well-established technologies, such as machine learning, pattern recognition, statistics, neural networks, fuzzy logic, evolutionary computing, database theory, artificial intelligence, and high performance computing to find relevant knowledge in very large databases. The knowledge discovery process is typically composed of the following phases: understanding the overall problem domain; obtaining a data set; cleaning, preprocessing, transforming, and reducing data; applying data mining tools; interpreting mined patterns; and consolidating and implementing discovered knowledge.

Data mining, which is an important phase in the knowledge discovery process, uses a number of analytical tools: discriminant analysis, neural networks, decision trees, fuzzy logic and sets, rough sets, genetic algorithms, association rules, and *k*-nearest neighbor (or memory-based reasoning) which are suitable for the tasks of classification, prediction, clustering, summarization, aggregation, and optimization. Classification and prediction are the two tasks that we deal with in this paper. These 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 two classes or outcomes ("debt recovered" or "debt unrecovered"). 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 (debt amount, injury code, patient age etc.) contained in the training set.

This paper examines and compares the effectiveness of four data mining techniques (neural networks, decision trees, logistic regression, memory-based reasoning) and the ensemble model in recovering bad debts. The target/dependent variable in a fairly large data set provided by a healthcare company represents the following three classes: 1 - "good" customers (those who repaid the debt), 2 - "bad" customers (those who defaulted), and 3 - "unknown" customers (those who were not pursued). Due to the low recovery rate, the number of "good" customers is vastly underrepresented in the data set. To build the models, we only used cases representing "good" and "bad" customers, rejecting all "unknown" cases. Computer simulation shows that on the test set, the logistic regression model, neural network model, and the ensemble model (combines logistic regression, neural network, and decision tree) produces the best overall classification accuracy rates, and the decision tree is the best in predicting "good" customers. More subtle and meaningful interpretation of the results are obtained by analyzing the lift and receiver operating characteristic charts. After building and testing the models, we used them to score 2,896 "unknown" cases, which had not been pursued by a company. The neural network model classified 1008 (34.8%) of these cases into "good" cases which could bring the healthcare provider an additional $423,000 in recovered income. In addition, our

models provide the patient financial service department a list of "unknown" patients sorted from the most likely to pay the bill to the least likely to pay the bill.

The paper is organized as follows. Section 2 reviews the recent literature first in general business data mining applications and then in bad debt recovery applications. Section 3 discusses the basic characteristics of the data mining methods used in the study. Section 4 describes the data sample, whereas section 5 provides the theoretical framework for evaluating classification results. Section 6 presents the experiments and simulation results. Finally, section 7 concludes the paper and makes recommendations for future work.

LITERATURE REVIEW

There is ample evidence in the literature that in cases where nonlinearities and complexities among variables are present, data mining tools such as neural networks, genetic algorithms, decision trees, fuzzy logic, rough sets, and memory-based reasoning often provide better classification accuracy rates than common statistical techniques such as regression analysis and discriminant analysis.

In earlier papers describing business applications of data mining, Back *et al.* (1996) designed several neural network models and also employed genetic algorithms to classify the financial performance of Finnish companies. Desai *et al*. (1996) explored the ability of neural networks and traditional techniques, such as discriminant analysis and logistic regression, in building credit scoring models in the credit union environment. They 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.

Jo and Han (1997) used case-based reasoning, neural networks, and discriminant analysis for bankruptcy prediction. Tessmer (1997) examined credits granted to small Belgian firms using a decision tree-based inductive learning approach to refine the credit granting process based on the impact of Type I and Type II credit errors (classifying good loans as bad loans, classifying bad loans as good loans). 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 and recommended the procedure called the dynamic updating process that relocates the boundary between Type I and Type II errors to define a more informed credit granting decision.

Barney *et al.* (1999) analyzed the performance of neural networks and regression analyses in distinguishing between farmers defaulting on farmers Home Administration Loans. 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.

In more recent papers, Jagielska *et al*. (1999) used the credit approval data set to investigate the performance of neural networks, fuzzy logic, genetic algorithms, rule induction software, and rough sets when applied to automated knowledge acquisition for classification problems. Jagielska 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. Neural networks performed significantly better than neurofuzzy systems in terms of classification accuracy, on both training as well as testing data. West (2000) investigated the credit scoring accuracy of five neural network architectures and compared them to traditional statistical methods. Using two real world data sets and testing the models using 10-fold cross-validation, 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. Glorfeld and Hardgrave (2000) 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. Yang *et al*. (2001) examined the application of neural networks to

an early warning system for loan risk assessment. Finally, Zurada (2002) investigated data mining techniques for loan-granting decisions and predicted default rates on consumer loans.

In this paper we apply data mining tools to data-driven marketing applications. In a typical application, solicitations are mailed with the aim of achieving a certain result. Some examples include (1) credit card applications, (2) insurance policies, (3) service contracts, and (4) bad-debt recovery. Regardless of the specific application maximizing the response rate is usually the desired objective.

For example, VISA International's use of neural networks to detect fraudulent credit card transactions provided savings estimated to be $40 million over a six month period (Anonymous, 1995). An English company has applied neural networks in direct marketing to identify the characteristics of people most likely to respond to a direct mailing campaign. The effort was worth 40,000 new customers, equivalent to $500,000 savings in mailing costs. American Express Co. has deployed neural networks in three projects. One involves a character recognition system, another is for direct mail prospects, and the third is for portfolio management trading support system (Berry, 1995). In a pilot study to detect fraud conducted at American Express, neural networks provided an improvement of 3% over the previously used logistic regression models (Punch, 1994).

The healthcare industry has been focused on ways to reduce bad-debt balance for the last several years. In one of the earlier studies, Zollinger *et al*. (1991) identified a sample of 985 patients classified as bad debt and charity cases from 28 Indiana hospitals. They built a multiple regression model and found that several institutional variables such as total hospital charge and the total hospital revenue and patient variables such as marital status, gender, diagnoses, insurance status, employment status, and discharge status were significant factors in recovering unpaid hospital bills. A similar study was performed by Buczko (1994) who analyzed data on charges assigned to charity care and bad debt for 82 short-stay hospitals in Washington. The study confirmed that uncompensated care has become a major issue in hospital finance as the number of uninsured persons has increased and hospital revenues have declined.

In one of the most recent articles, Veletsos (2003) described the predictive modeling software (IBM Intelligent Miner and DB2) used for bad-debt recovery implemented by the IBM Company for the Florida Hospital at Orlando. The final study was completed in 2003 and included approximately 2,400 patients. The model is based on a variety of data variables, including credit factors, demographic information and previous organizational payment patterns. The model provides the patient financial service department a list of patients sorted from the most likely to pay the bill to the least likely to pay the bill. The model yielded approximately $200,000 in savings. In this study, which is a substantial extension of the approach discussed by Veletsos (2003) and Zurada and Lonial (2004), we use a larger and unbalanced sample of patients and fewer variables to test the effectiveness of the five data mining models for recovering bad-debts.

CHARACTERISTICS OF THE TOOLS USED

The tools described in this section and the algorithms they use are relatively well established. As a result, we provide only the basic features of each tool, without giving the mathematical descriptions of the algorithms underlying the operation of each tool. For more details regarding specific algorithms on neural networks, decision trees, regression analysis, and memory-based reasoning the reader is referred to (Hagan *et al*., 1996; Quinlan, 1993; Mitchell, 1997; Fayyad *et al*., 1996, Afifi and Clark, 1990; Christensen, 1997, Pyle, 2003; and Han & Kamber, 2001).

Artificial Neural Networks

Artificial neural networks are one of disciplines of artificial intelligence which attempts to implement some of the powerful characteristics of the human brain on digital computers. Neural networks learn the nonlinear relationships, patterns, and trends in the data when the training data are presented to the network. Once trained, the artificial neural network models make high fidelity predictions for a fresh data set not seen by the network during training. Artificial neural network models are used in a variety of applications, for example, nonlinear mapping, data reduction, pattern recognition, clustering, and classification. The problem considered in this paper is a classification application.

In this study we use a popular three-layer feed-forward neural network with back propagation. The network has three layers, input, hidden, and output layer. Network inputs are fed to the input layer and the output layer and the output layer produces output. The middle layer is called "hidden layer" since it does not communicate with the environment (that is inputs and outputs) directly. All three layers contain a number of neurons that are connected by means of connection weights. The number of neurons in the input layer equals the number of inputs to the network, while the number of neurons in the output layer corresponds to the number of system outputs. Each neuron contains two elements: the summation node and the sigmoid transfer function. The summation node calculates the product of each normalized input value and the weight value. The problem of modeling consists of finding a proper set of connection weights using a suitable optimization algorithm so that the error between predicted and experimental outputs is minimized (Hagan, *et al*., 1996).

Unlike statistical methods, neural networks do not depend on the assumptions about the independence and distribution of residuals or collinearity of input variables. On the other hand, a large volume of data is required for training, and the neural networks parameters (connection weights) provide little insight into the physics of the process. This is a disadvantage because connection weights cannot be easily converted to if-then rules which one can understand.

Logistic Regression Analysis

The purpose of the logistic regression model is to obtain a mathematical equation that predicts the membership of an object among two or more groups. Logistic regression also attempts to predict the probability that a binary or ordinal target will acquire the event of interest as a function of one or more independent variables. For more details see (Afifi and Clark, 1990; 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 which researchers can then analyze to test the usefulness of specific information.

Logistic regression models may be thought of as the simplest feedforward neural networks containing two layers, an input layer and an output layer. Logistic regression is meant for solving classification problems since the computed outputs can be expressed in terms of probabilities. Inputs are fed to the input layer where no computation takes place. Each neuron in an input layer is connected to an output layer neuron by means of a connection weight. The output of an output layer neuron is computed with the logistic sigmoid transfer function. Connection weights are determined by an iterative optimization procedure that seeks to minimize a measure of prediction errors.

Decision Trees

Decision trees are particularly useful for classification tasks. Like neural networks, decision trees learn from data. Using search heuristics, decision trees find explicit and understandable rules-like relationships among the input and output 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 as there are fewer 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 decision trees 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.

One of the most significant 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 cases entering the node, the way those cases would be classified if these were leaf nodes, and the percentage of records classified correctly.

Case- or Memory-based Reasoning

Instance-based methods such as *k*-nearest neighbor share three key principles. First, they defer the decision of how to generalize beyond the training data until a new (unlabeled) sample/case/instance needs to be classified. This contrasts with decision trees and neural networks which construct a generalization model before receiving new samples to classify. Second, instance-based methods classify query instances by analyzing similar instances while ignoring instances that are different from the query. Third, they represent instances as real-valued points in an *n*-dimensional Euclidean space.

Case- or memory-based reasoning classifiers are instanced-based as well. Moreover, they are based on the first two principles, but not the third. Unlike *k*-nearest neighbor classifiers, which store training samples as points in Euclidean space, the sample or cases stored by memory-based reasoner are complex symbolic descriptions. When given a new case to classify, a memory-based reasoner first checks if an identical training case exists. If one is found, then the accompanying solution to that case is returned. If no identical case is found, then the case-based reasoner will search for training cases having components that are similar to those of the new case. Conceptually, these training cases may be considered as neighbors of the new case. If incompatibilities arise with the individual solutions, backtracking to search for other solutions may be necessary (Mitchell, 1997; Han & Kamber, 2001).

Challenges in memory-based reasoning include finding a good similarity metric, developing efficient technique for indexing training cases, and methods for combining solutions.

THE DATA SAMPLE

The healthcare company, which is the subject of this study, relied on only four factors to determine whether the bad debt was recoverable: (1) Patient Age (PA), (2) Patient Gender (PG), (3) Injury Diagnosis Code (IDC), and (4) Dollar Amount of the Claim (DAC). In all likelihood, the four factors constituted all of the information about the patient-debtor that was available to the healthcare company. Furthermore, aside from the amount owed, the information appears to be only tangentially related the probability that a particular bad-debt could be recovered. Over one quarter the company identified 6,319 new cases with an outstanding balance. The dependent variable Status represented groups 1, 2, and 3 containing "good", "bad", and "unknown" cases, respectively. After eliminating cases which had at least one missing value, we obtained a data set containing 6180 observations unequally divided into 449 "good" cases (group 1); 2,835 "bad" cases (group 2); and 2896 "unknown" cases (group 3). We performed the following transformations of the data. We recorded a nominal independent variable IDC into 24 categories (the most frequent occurrences are shown in column 4 of Table 1), which were then converted to 23 dummy variables.

To learn more about the distribution of the variables within the data set and to find out whether any transformation of the variables is needed, we performed a simple bivariate exploratory data analysis. The results are summarized in Table 1. The table reveals that for the DAC variable the average dollar amount of the recovered cases (group 1); not recovered (group 2); and not pursued claims (group 3) are \$1,052; \$417; and \$254; respectively. Furthermore, the table shows that the total amounts for the DAC variable for each of the 3 groups are \$472,461; \$1,182,350; and \$734,188; respectively. Thus it appears that a company used common sense and some procedure which allowed it to target the patients with larger debts and ignore those with smaller debts. Furthermore, the skewness coefficient S_k shows that the distribution of the DAC variable is very positively skewed, especially for group 3, which suggests that small dues were simply not pursued (S_k=19.2). Because for cases belonging to group 3 ("unknown" cases) debt collection was not pursued by the subject company, we did not use this group to build the models. It is possible that the low debt recovery rate could have been caused by the healthcare company using primarily the amount owed to determine which bad debts to target. The purpose of the data mining techniques is to use the seemingly unrelated factors such as the patient's gender, age and type of injury to determine the likelihood that a particular patient-debtor will pay his overdue bill. Instead, each case from group 3 was used to test the final models to find out whether the models would classify them as a "good" or "bad" case. As a result, the data set used to build

and test the models contained 3,284 cases divided unequally between 449 "good" cases (group 1) and 2,835 "bad" cases (group 2). We also performed additional transformation for the two variables. To improve the distribution of the DAC variable and obtain better prediction results, we computed and used log (DAC) instead of DAC. Furthermore, to decrease the dimensionality of the data set (the number of distinct values in the data set), we grouped the PA variable into several bins.

Table 1: Summary of the Exploratory Data Analysis for the Variables Used in the Models

Status	**Patient Gender (PG)**	**Patient Age (PA)**	**Injury Diagnosis Code (IDC)**. (Most frequently represented shown only [>5%].) Code Range	%	**Dollar Amount of the Claim (PAC)** [in US $]
Overall (groups 1, 2, and 3)	Female (N=2884, 52.9%) Male (N=3235, 47.1%)	Mean=34 St. Dev=19 Min=0 Max=88 S_k=0.5	"800"-"829" "840"-"849" "870"-"899" "920"-"924" "958"-"959"	16 21 12 15 14	Mean=$387 St. Dev=$1,470 Min=$1 Max=$40,508 Sum=$2,388,999 S_k=12.9
1 (Good) N=449	Female (N=248, 55.2%) Male (N=201, 44.8%)	Mean=34 St. Dev=18.9 Min=0 Max=88 S_k=0.5	"800"-"829" "840"-"849" "870"-"899" "920"-"924" "958"-"959"	13 23 10 13 16	Mean=$1,052 St. Dev=$3,442 Min=$3 Max=$40,508 Sum=$472,461 S_k=7.2
2 (Bad) N=2835	Female (N=1331, 47%) Male (N=1504, 53%)	Mean=31.6 St. Dev=23 Min=0 Max=100 S_k=0.6	"800"-"829" "840"-"849" "870"-"899" "920"-"924" "958"-"959"	18 17 15 8 19	Mean=$417 St. Dev=$1,248 Min=$1 Max=$19,568 Sum=$1,182,350 S_k=7.8
3 (Unknown) N=2896	Female (N=1335, 46%) Male (N=1561, 54%)	Mean=28 St. Dev=19 Min=0 Max=100 S_k=0.8	"800"-"829" "840"-"849" "870"-"899" "920"-"924" "958"-"959"	14 19 14 10 20	Mean=$254 St. Dev=$1,081 Min=$1 Max=$30,976 Sum=$734,188 S_k=19.2

THE THEORETICAL FRAMEWORK FOR EVALUATING CLASSIFICATION RESULTS

Estimating the classification accuracy of the models allows one to evaluate how accurately a model will classify future data, that is, data on which the model has not been trained. Accuracy estimates also help in the comparison of different classification models. The holdout method, which we used in computer simulation, is one of several techniques for assessing classifier accuracy based on randomly sampled portions of data. In this method, typically two thirds of the data are randomly allocated to the training and validation sets, and the remaining one third is allocated to the test set. If the holdout method is repeated *k* times, the overall accuracy estimate is taken as the average of the accuracy rates on the test set obtained from each iteration. For more details on the holdout method and other methods, the reader is referred to (Han and Kamber, 2001; Kantardzic, 2003).

The classification results can often be summarized in a theoretical confusion matrix presented in Table 2 (Giudici, 2003; Kantardzic, 2003).

Table 2: Theoretical Confusion Matrix

Predicted / **Observed**	**Event (1)**	**Non-event (0)**	**Total**
Event (1)	*a*	*b*	*a+b*
Non-event (0)	*c*	*d*	*c+d*
Total	*a+c*	*b+d*	*a+b+c+d*

The table classifies the observations on a test data set into four possible categories:

- Cases predicted as events and effectively such (with absolute frequency equal to *a*)
- Cases predicted as non-events and effectively events (with absolute frequency equal to *b)*
- Cases predicted as events and effectively non-events (with absolute frequency equal to *c*)
- Cases predicted as non-events and effectively such (with absolute frequency equal to *d*)

The computation of error rate is based on counting of errors in a testing process. These errors are defined as misclassification (wrongly classified examples). In a binary classification problem, where the target variable has only two classes: True (Event or 1) and False (Non-event or 0), two different related error rates are of interest. These are false negative errors [it is expected to be T (Event), but it is classified as F (Non-event)] and false positive errors [it is expected to be F (Non-event), but it is classified as T (event)]. In Table 2, their absolute frequencies are denoted by *b* and *c*, respectively.

Assuming that all errors are of equal importance, and error rate *R* is the number of errors *E* divided by the number of samples *S* in the testing set: (*R=E/S=(b+c)/(a+b+c+d)*). The accuracy of a model is a part of the testing data set that is classified correctly, and it is computed as one minus the error rate: *A=1-R=(S-E)/S=(a+d)/(a+b+c+d)*.

In many applications, it is not adequate to characterize the performance of a model by four single numbers *a*, *b*, *c*, and *d* that measure the correct classification and error rates. More accurate, complex and global measures are necessary to describe the quality of the model. These measures include cumulative and non-cumulative percent response, lift, and receiver operating characteristic (ROC) charts (Giudici, 2003). In section 6 we use these charts to assess more thoroughly the classification performance of the developed models.

The ROC chart, for example, is of special interest because it allows one to analyze both false negative and false positive errors at the same time for different cut-off points. The chart measures the predictive accuracy of a model. It is based on the confusion matrix in Table 2. More precisely, the ROC chart is based on the following conditional probabilities:

- *Sensitivity a/(a+b)* is the proportion of events predicted as such
- *Specificity d/(c+d)* is the proportion of non-events predicted as such
- *False positives c/(c+d)=1-specificity* is the proportion of non-events predicted as events (type II error). For example, in the bad debt recovery case, these are the test cases in which patients/customers are wrongly predicted as payers.
- *False negatives b/(a+b)=1-sensitivity* is the proportion of events predicted as non-events (type I error). In the bad debt recovery example, these are the cases of test patients who are wrongly classified as non-payers.

The curve is obtained by graphing, for any fixed cut-off value, the sensitivity on the vertical axis and the false positives (1-specificity) on the horizontal axis. Each point on the curve corresponds to a particular cut-off. The ROC curve can also be used to select cut-off point, trading off sensitivity and specificity. A classification model can be tuned by setting an appropriate threshold value to operate at a desired value of the false positive rate. If one tries to decrease the false positive rate parameter of the model, however, it would increase the false negative rate and vice versa. In terms of model comparison, the ideal curve coincides with the vertical axis, so the best curve is the leftmost

curve. The curve will always lie above the 45° line. The curve permits one to assess the performance of the model at various operating points (thresholds in a decision process using the available model) and the performance of the model as a whole (using as a parameter the area below the ROC curve). Furthermore, the area between the curve and the 45° line can also be calculated, and gives the Gini index of performance. The higher the area is, the better the model. The ROC curve is especially useful for a comparison of the performances of several models obtained using different data-mining methodologies. For more details, see (Giudici, 2003; Kantardzic, 2003).

A cumulative and non-cumulative percent response and lift charts let one visualize the effectiveness and predictive power of the model. A cumulative response and lift charts show the overall strength of the model, but they obscure the model's performance at each stratam of the score. To see the effectiveness of the model for each level of the score, one needs to analyze non-cumulative charts (www.sas.com). We elaborate on these charts in section 6.

THE EXPERIMENTS AND SIMULATION RESULTS

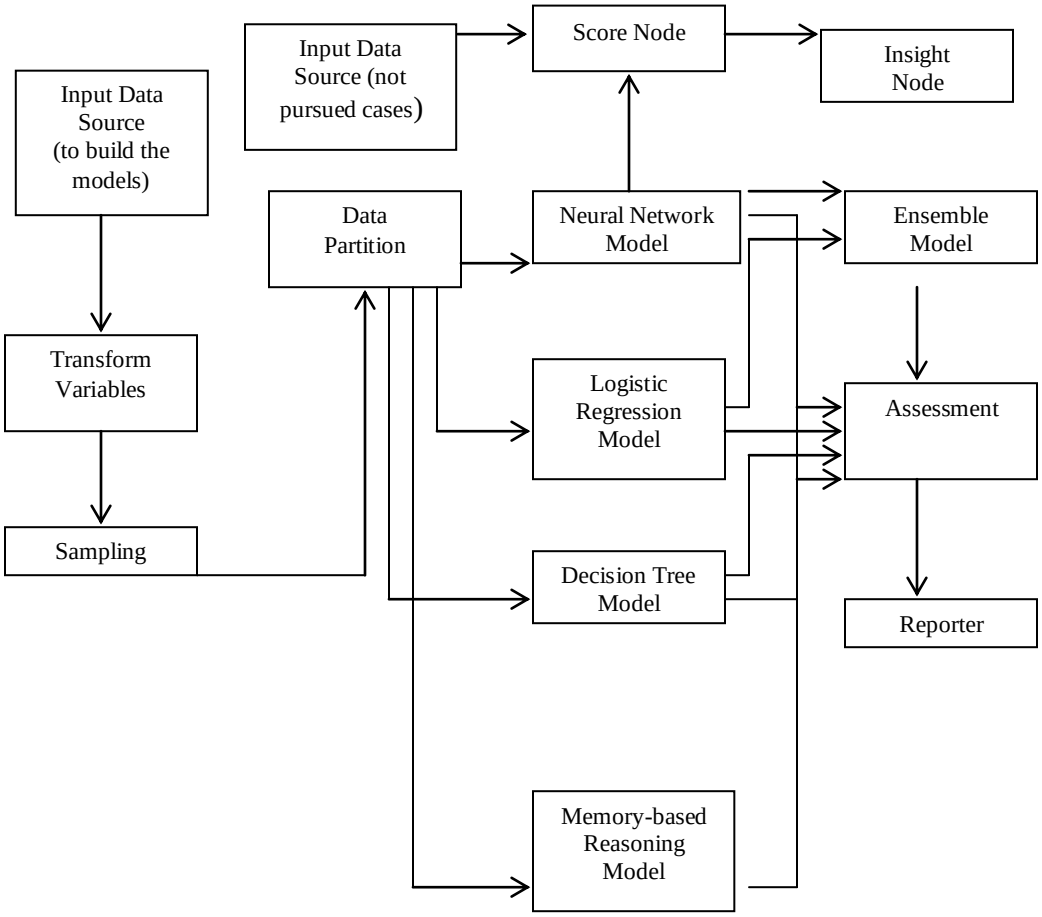
We employed SAS Enterprise Miner (EM) to build an initial model and a final model. The initial model was trained, validated, and tested on an unbalanced sample which contained 3,284 cases. The test data set was used to check the performance of the initial model. The initial model's overall classification accuracy was excellent and the classification rate of "bad" cases was almost perfect (100%). However, the classification rate of "good" cases was poor because "good" cases were vastly underrepresented in the training data sample, and the model classified the majority of "good" cases as "bad" cases. As a result, the initial model was found unacceptable.

In the final model, we performed stratified sampling to balance the data sample, by randomly selecting 449 "bad" cases from 2835 "bad" cases and matching them with all 449 "good" cases. After sampling, the data sample contained 898 cases equally divided between 449 "good" cases and 449 "bad" cases. This data set was divided into training, validation and test sets, each containing 450 (50%), 224 (25%), and 224 cases (25%), respectively, evenly representing the 2 groups. The first two subsets were used to build the model, whereas the test subset was used to check the performance of the model. It was assumed that detecting "good" cases was the target event.

The SAS Enterprise Miner software (release 8.2) simulation workflow, which we used to build, test, and deploy the models, is presented in Figure 1 (SAS Institute, www.sas.com). The Input Data Source node reads a data source and defines the input and output attributes for later processing. One Input Data Source node was used to read 3,284 samples, representing all 2,839 "bad" and 449 "good" cases, to build and test the models. The other was used to read in the 2896 "unknown" cases and pass them on to the Score node, which enables one to generate and manage predicted values from a trained model (i.e., the Neural Network node). The Transform Variables node enables one to transform variables in order to improve the fit of models and, subsequently, the classification and prediction precision of fitted models. The Sampling node enables one to perform random sampling, stratified random sampling, and cluster sampling. We used stratified random sampling to select 449 "bad" cases, which were matched with the same number of all 449 "good" cases. The Data Partition node allows one to randomly partition 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, Memory-based Reasoning, and Regression nodes enable one to build various classification/prediction models. Using the Neural Network node, we constructed, trained, and validated a multiplayer feed-forward network with error back-propagation. Here we used a neural network with 5 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). The standard deviation normalization was used for the variables. (The resulting values for each variable have a mean of 0 and a standard deviation of 1.) 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 (Dhar and Stein, 1997; Mitchell, 1997). The Regression node fitted logistic regression model to the data.

Figure 1: SAS EM Simulation Workflow



The Memory-based Reasoning node is a modeling tool that uses a *k*-nearest neighbor algorithm to categorize or predict observations. The method requires no model to be fitted, or function to be estimated. Instead it requires all observations to be maintained in memory, and when a prediction is required, the method recalls items from memory and predicts the probability of the Status variable for a particular case. Two crucial choices in the nearest neighbor-method are the distance function and the cardinality *k* of the neighborhood. We used the Euclidean and Hamming distance measures for all the independent variables to calculate the similarity between the cases. After performing several experiments, we chose *k*=16 because this value of *k* seemed to give the lowest misclassification rates. This means that 16 most similar cases (neighbors) were used to predict the value of the variable Status. The Ensemble node combined the best three models by averaging the posterior probabilities for the class target variable Status. Finally, the Assessment and Reporter nodes provide a common framework for assessing, comparing, and assembling the results from the models used.

Table 3 compares the test set classification accuracy of decision trees, neural networks, logistic regression, memory-based reasoning, and the ensemble model for one of several computer simulations. Unlike a confusion matrix, Table 3 does not show, however, the number and percentage of false positive and false negative errors. For those, the reader is referred to the ROC chart presented in Figure 6. (We ran computer simulation for a random selection of several different sets of 449 "bad" cases. Each of them was matched with the same 449 "good" cases. We also ran the experiments for several randomly selected training, validation, and test subsets. All the above scenarios yielded very similar classification accuracy rates across the five models. See the holdout method described in section

5.) The ensemble model combines the best three models (neural network, logistic regression, and decision tree) by averaging the posterior probabilities of the response variable Status. It is clear that in the overall classification accuracy, logistic regression, ensemble model, and neural network outperform the other models. The decision tree, however, does the best job in classifying "good" cases. Based on the data in Table 3, we performed several 2-tailed proportional z-tests to determine whether the classification rates across the five models are significantly different from one another. The tests revealed that there are statistically significant differences in classification accuracy rates, especially for "overall" and "bad" cases. The classification accuracy rates obtained seem to be excellent if one considers the fact that the healthcare institution that provided data for this paper successfully recovered bad debts from only about 449 of its 6180 total patients (7.3%).

Table 3: The test set correct classification accuracy rates for the 5 methods used for 0.5 cut-off point. (It shows the percentage and number of test cases classified correctly.)

"Overall": * Decision tree is significantly better than memory-based reasoning at α=0.1. ** Logistic regression's is significantly better than decision tree at α=0.05. ## Neural network, logistic regression, and ensemble model each is significantly better than memory-based reasoning at α=0.01.
"Bad": $ Decision tree is significantly better than memory-based reasoning at α=0.1. $$ Neural network, logistic regression, and ensemble model each is significantly better than memory-based reasoning at α=0.01. && Neural network, logistic regression, and ensemble model each is significantly better than decision tree at α=0.01.

	Decision Tree	Neural Network	Logistic Regression	Memory-based Reasoning	Ensemble Model
"Overall"	67.9% * (152/224)	72.3% ## (162/224)	75.0% ** ## (168/224)	61.2% (137/224)	73.7% ## (165/224)
"Good"	75.0% (84/112)	67.9% (76/112)	71.4% (80/112)	72.3% (81/112)	71.4% (80/112)
"Bad"	60.7% $ (68/112)	76.8% $$ && (86/112)	78.6% $$ && (88/112)	50.0% (56/112)	75.9% $$ && (85/112)

The performance of the five models, however, can be best and more thoroughly evaluated by using a combination of cumulative and non-cumulative percent response charts, lift charts, captured response chart, and the receiver operating characteristic charts. The strength of the models is demonstrated by the degree a cumulative percent response chart curve pushes upward and to the left. To properly interpret a cumulative percent response chart, one needs to understand how the chart is constructed. In the chart (Figure 2), a respondent is defined as an individual from whom payment is recovered ("good", Status=1). For each individual, the fitted models predict the probability that the individual will repay the debt. 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 Status, one can count the percentage of actual responders in each bin. If the model is effective, the proportion of individuals with the event level being modeled (in this example, those who will repay dues) will be relatively high in bins in which the predicted probability of response is high. The chart shows the percentage of respondents in the top 10%, top 20%, and so on. In a chart, the response rate for each decile of the score includes all of the responses for the deciles above it.

One sees (Figure 2) that the 5 models predict that between 65% and 100% of the respondents in the first 10% decile will repay dues. The logistic regression, ensemble, and neural network models outperform the other 2 models with the 100%, 92%, and 83% repay rate, respectively. The regression model needs to be chosen if a company intends to target only 10% of the patients. The performance of the regression model, however, drops significantly between the 1st decile (10% percentile) and the 2nd decile to about 90%. Between the 2nd and 10th deciles, the regression and ensemble models perform almost exactly the same. The ensemble model, however, predicts that 75% of the respondents in the 5 first deciles (50% percentile) will repay dues, and it is slightly better than the regression model (74%) and neural network model (71%). The decision tree and memory-based reasoning (denoted as User) models undoubtedly performs worst in all deciles. The horizontal response line represents the baseline rate (approximately

50%) for comparison purposes, which is an estimate of the percentage of payers that one would expect if they were targeted at random.

Figure 2: The test set cumulative percent response chart for the five methods. Target event: 1 ("good" cases)

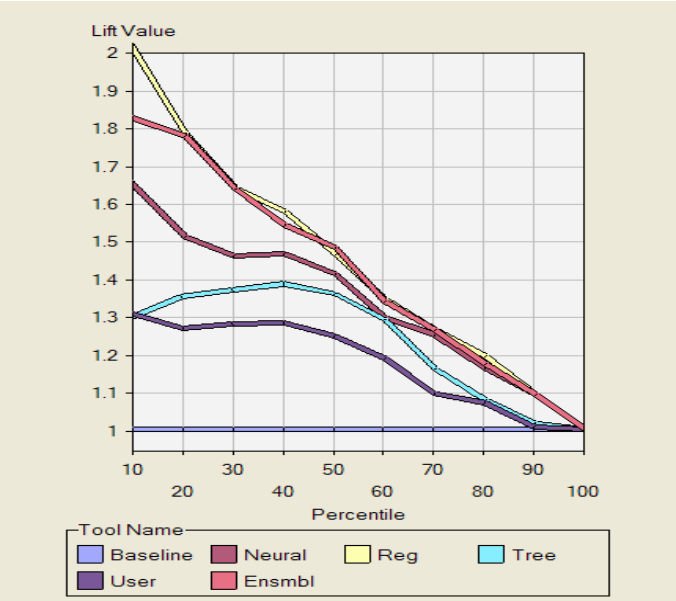
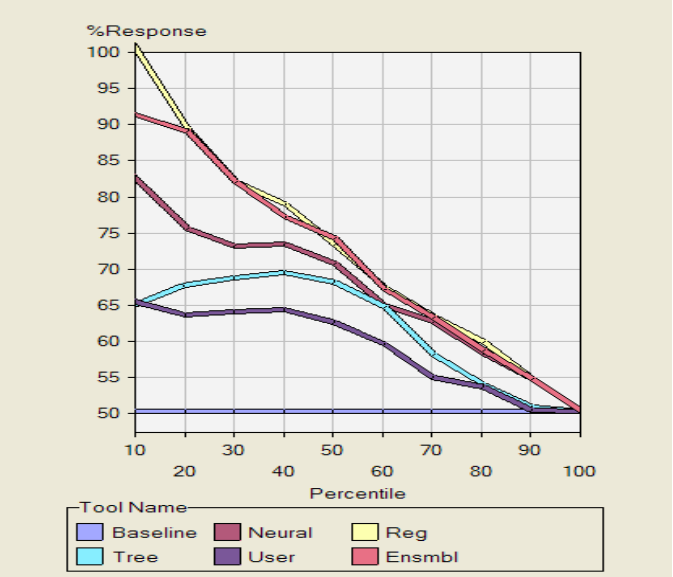


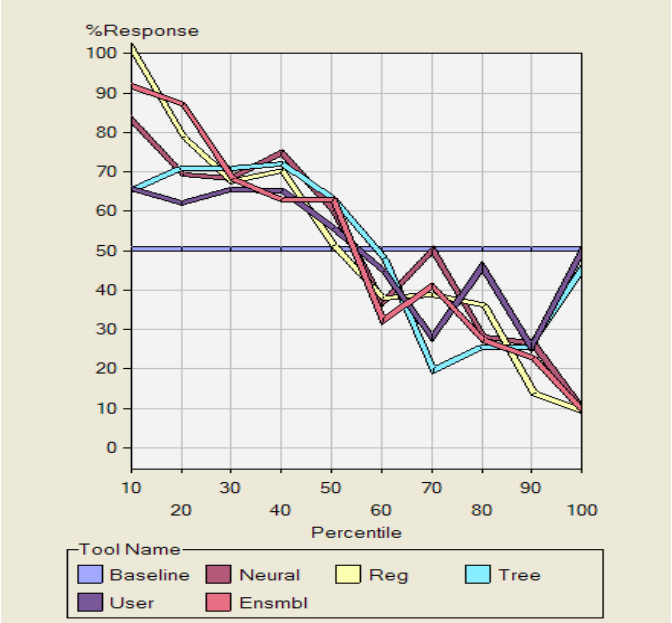
Figure 3: The test set cumulative lift chart for the five methods. Target event: 1 ("good" cases)



The cumulative lift chart (Figure 3) presents basically the same information, but on a different scale. This chart shows that for the regression and ensemble models, the percentage of individuals from whom debts are recovered in the first decile is about 2 times higher than the 50% default rate in the population. Specifically, the

regression, ensemble, and neural network models' lift value is about 2 (100%/50% baseline rate), 1.83, and 1.65, respectively. The lift values for both the regression and ensemble models drop in the second decile to 1.8. Furthermore, the lift value for the neural network model decreases in the second decile to 1.5. Only the lift value for the ensemble model remains approximately the same (1.8) between the 1st and 2nd deciles.

Figure 4: The test set non-cumulative % response chart for the five methods. Target event: 1 ("good" cases)



To analyze the performance of the models at each decile (strata) of the score, it is necessary to examine a non-cumulative percent response chart, which shows the percentage of "good" respondents in each decile. Figure 4 shows that the percentage of "good" respondents for the neural network model and the ensemble model is the highest (76%) and the lowest (62%), respectively, in the 4th decile. The curves for all 5 models decline significantly between the 5th decile and the 6th decile and drop below the baseline for the 6th through the 10th decile. By the 6th decile the models are nearly worthless, performing scarcely better than the baseline. Below the baseline, the models become counterproductive and actually predict "bad" respondents. Crossovers happen often when one compares models on non-cumulative charts. When lines do not cross, the one on top is the better model. But when response rates overlap on the chart, one should consider the target strata before deciding on the best model. For example, if one plans to target 40% of "good" respondents, one should use the neural network model which has the response rate equal to 76%.

To find out the percentage of the total number of "good" respondents in a bin, one can examine a percent captured response curve. The baseline in Figure 5 shows that if one chooses a random sample of 50% of the observations, one would expect to capture 50% of "good" respondents. If the percentage of responders chosen for action were approximately 50%, the neural network, logistic regression, and ensemble models would have identified between 71% and 76% of those who would be "good" responders, which corresponds to a lift value of about 76/50=1.5.

The receiver operating characteristic (ROC) charts display the global measure of the predictive accuracy of the models. They display the sensitivity on the vertical axis against 1-specificity on the horizontal axis 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 curves will always lie above the 45° line. The area between the curves and the line provides a quantitative performance measure called the Gini index. This area will range from 50, for a worthless model, to 100, for a perfect classifier.

Figure 5: The test set captured response chart for the five methods. Target event: 1 ("good" cases)

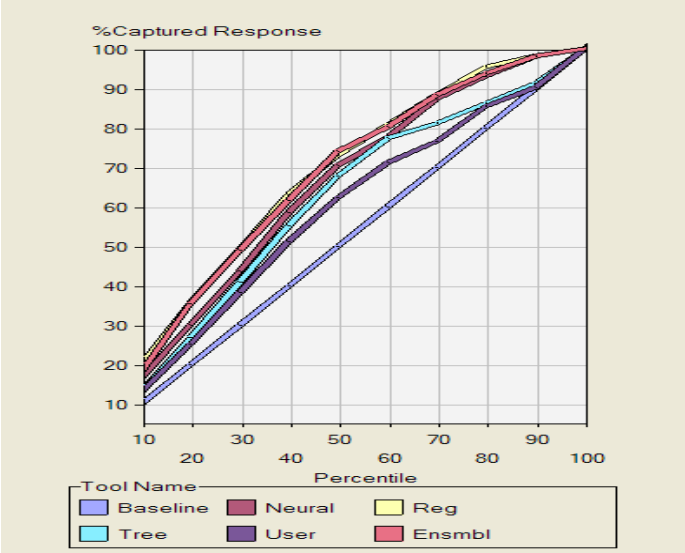
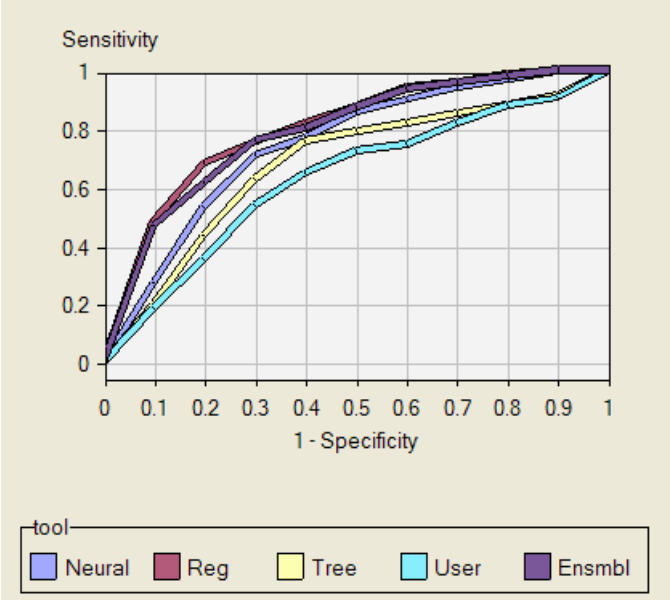


Figure 6: The test set receiving operating characteristics (ROC) chart for the five methods. Target event: 1 ("good" cases)



The ROC chart (Figure 6) confirms our previous findings and indicates that the predictive power all 3 models (neural network, regression analysis, and combined model) stand out and appear to be better at predicting "good" respondents than the decision tree and memory-base reasoning models. For (0.1, 0.5) coordinates representing (1-specificity) and sensitivity, respectively, the regression and ensemble models perform equally well. However, for the coordinates (0.2, 0.7), the regression model appears to be the best. The choice of the model typically may also depend on the costs of the errors. For the problem at hand, type I error (the cases of test patients who are wrongly classified as non-payers) seems to be more costly. As a result, a lower cut-off should be chosen because it will allow one to minimize the type I error. Conversely, in the upper right part of the graph, all 3 mentioned models perform equally well, as they lead to a higher sensitivity (lower type I error).

Finally, the best three models (logistic regression, neural network, and ensemble model) were used to score how many out of the 2896 "unknown" cases would be classified as "good" cases, bringing a company additional revenue. The results summarized in Table 4 are striking. For example, the neural network model classifies 1008 of all "unknown" cases as "good" cases (34.8%), potentially yielding the amount of $422,861 in additional revenue. This amount constitutes about (58%) of the total amount of $734,188 for not pursued cases. The results can be explained by the fact that the models learned how to target patients with larger debts.

Table 4: Results from the Score Node

	Logistic Regression	**Ensemble Model**	**Neural Network**
Number and percentage of cases classified as 1 ("good") - cutoff probability ≥ 0.5	886/2896 (30.6%)	910/2896 (31.4%)	1008/2896 (34.8%)
The additional dollar amount potentially retrieved [in US $]	$374,185	$382,148	$422,861

CONCLUSION

The paper compares the effectiveness of neural networks, decision trees, logistic regression, memory-based reasoning, and the ensemble model in recovering bad debts. The data analysis and evaluation of the performance of the various models is based on a fairly large unbalanced data sample provided by a healthcare company, in which cases with recovered bad debts are underrepresented. Computer simulation shows that the logistic regression model, neural network model, and the ensemble (combined) model produced the best overall classification accuracy, and the decision tree was the best in classifying "good" cases. More subtle and meaningful interpretation of the results is obtained by analyzing the percent response, lift and ROC charts. We used the models to score the "unknown" cases, which were not pursued by a company. The neural network model classified more "unknown" cases into "good" cases than any other remaining models. This may potentially bring a company the additional recovered income. In addition, our models can provide the patient financial service department a list of patients sorted from the most likely to pay the bill to the least likely to pay the bill. To collect bad debts more effectively, we recommend that a company first deploy and use the models, before it turns over unrecovered cases to a collection agency.

The choice of the models may also depend on an interpretational capability of them, and consequently who will be using the results. Decision trees produce if-then rules that are easily interpretable in logical terms. Logistic regression attaches an impact weight to each explanatory variable (measured by a regression coefficient). Neural networks and memory-based reasoning models are non-parametric tools and do not deliver explicit rules. Therefore, they are hard to interpret. If the results are used by trained statisticians or IT experts then it does not matter which model is chosen. However, decision trees and logistic regression are better for business experts because they provide a more user-friendly picture of what is going on.

It seems like the difference in the performance between old method used by the healthcare company and the various date mining techniques is the ability of each method to discover some hidden/not readily apparent relationships between the factors given. While total cost of claim may have been a factor – the others may be important too. Females seem more likely to pay their bills than males (Table 1).

Although the results obtained from this study could be generalized, one of its limitations may be a small number of independent variables used for prediction. In future work, we plan to (1) include for analysis more input variables including current insurance status, employment status, and discharge status if they become available, (2) consider the cost profit matrix, and (3) adjust the results by prior probabilities/conditions to avoid a potential bias in the results. In addition, it would be interesting to explore whether patients with certain injury codes (twisted ankle) are more likely to pay their debts than patients with a serious low back injury who will not go back to their job.

REFERENCES

1. Anonymous, 1995, "Visa Stamps on Fraud", *International Journal of Retail and Distribution Management*, Vol. 23, Iss. 11, pg. 1
2. Afifi, A.A, and Clark, V., 1990, *Computer-Aided Multivariate Analysis*, Van Nostrand Reinhold Co., New York.
3. 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, 99-139, 1999.
4. Back B., Laitinen, T., and Sere, K., "Neural Networks and Genetic Algorithms for Bankruptcy Predictions," *Expert Systems with Applications*, Vol. 11, No. 4, 407-413, 1996.
5. Berry, J., "A Potent New Tool for Selling: Database Marketing", *Business Week*, Iss. 3388, pg. 56, 1995.
6. Berry, M.J.A., and Linoff, G.S., *Data Mining Techniques for Marketing, Sales, and Customer Support*, John Wiley & Sons, Inc, 1997.
7. Buczko, W., "Factors Affecting Charity Care and Bad Debt Charges in Washington Hospitals", *Hospital and Health Services Administration*, Vol. 39, Iss. 2, 179-191, 1994.
8. Christensen, R., *Log-Linear Models and Logistic Regression*, Springer, New York, 1997.
9. 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, 1996.
10. Dhar, V., and Stein, R., *Seven Methods for Transforming Corporate Data Into Business Intelligence*, Prentice Hall, 1997.
11. Fayyad, U.S., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds.), *Advances in Knowledge Discovery and Data Mining*, AAAI Press / The MIT Press, Menlo Park, California, USA, 1996.
12. Galloro, V., "Bad News on Bad Debt", *Modern Healthcare*, Vol. 33, Iss. 43, pg. 8, Oct 2003.
13. Giudici, P., *Applied Data Mining: Statistical Methods for Business and Industry*, John Wiley & Sons, Chichester, West Sussex, England, 2003
14. Glorfeld, L.W., and Hardgrave, B.C., "An Improved Method for Developing Neural Networks: The Case of Evaluating Commercial Loan Credit Worthiness," *Computer & Operations Research*, Vol. 23, No. 10, 933-944, 1996.
15. Hagan, M.T., Demuth, H.B., and Beale, M., *Neural Network Design*, PWS Publishing Company, 1996.
16. Han, J., and Kamber, M., *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers: San Francisco, CA, 2001.
17. 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, 37-54, 1999.
18. Jain, B.A., and Nag, B.N., "Performance Evaluation of Neural Network Models," *Journal of Management Information Systems*, Vol. 14, No. 2, 201-216, 1997.
19. Jo, H., and Han, I., "Bankruptcy Prediction Using Case-Based Reasoning, Neural Networks, and Discriminant Analysis," *Expert Systems with Applications*, Vol. 13, No. 2, 97-108, 1997.
20. Kantardzic, M., *Data Mining: Concepts, Models, Methods, and Algorithms*, IEEE Press/Wiley, 2003.
21. Kumar, N., Krovi, R., and, Rajagopalan, B., "Financial Decision Support with Hybrid Genetic and Neural Based Modeling Tools," *European Journal of Operation Research*, Vol. 103, 339-349, 1997.
22. Lagnado, L., "Hospitals Try Extreme Measures to Collect their Overdue Debts", *The Wall Street Journal*, Oct 30, 2003.
23. Lee, K.C., Han, I., and Kwon, Y., "Hybrid Neural Network for Bankruptcy Predictions," *Decision Support Systems*, 18, 63-72, 1996.
24. Manly, B.F., *Multivariate Statistical Methods: A Primer*, Chapman & Hall, London, 1994.

25. Mitchell, T.M., *Machine Learning*, WCB/McGraw-Hill, Boston, Massachusetts, 1997.
26. Pesce, J., "Stanching Hospitals' Financial Hemorrhage with Information Technology", *Health Management Technology*, Vol. 24, Iss. 8, pg.12, 2003.
27. Piramuthu, S., "Financial Credit-Risk Evaluation with Neural and Neurofuzzy Systems," *European Journal of Operational Research*, Vol. 112, 310-321, 1999.
28. Punch, L., "When Big Brother Goes to Far", *Credit Card Management*, Vol. 7, Iss. 7, pg. 22, 1994.
29. Pyle, D., *Business Modeling and Data Mining*, Morgan Kaufman Publishers (An Imprint of Elsevier Science), San Francisco, California, 2003.
30. Quinlan, J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufman Publishers, San Mateo, California, 1993.
31. SAS Institute, "Data Mining Using Enterprise Miner Software: A Case Study Approach", 1st edition, http://www.sas.com
32. Tessmer, A.C., "What to Learn from Near Misses: An Inductive learning Approach to Credit Risk Assessment," *Decision Sciences*, Vol. 28, No. 1, 105-120, 1997.
33. Thomas, L.C., "A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Consumers," *International Journal of Forecasting*, Vol. 16, 149-172, 2000.
34. Veletsos, A., "Getting to the Bottom of Hospital Finance", *Health Management Technology*, Vol. 24, Iss. 8; pg. 30, 2003.
35. West, D., "Neural Network Credit Scoring Models," *Computers & Operations Research*, Vol. 27, 1131-1152, 2000.
36. 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, 303-306, 2001.
37. Zollinger, Terrell W., Sawyell, Robert M. Jr., Chu, David K. W., Ziegert, Andrea, Woods, John R., LaBov, Dean, "A Determination of Institutional and Patient Factors Affecting Uncompensated Hospital Care", *Hospital & Health Services Administration*, Chicago, 36(2), 243-256, 1991.
38. Zurada, J., and Zurada, M., "Data Mining Techniques in Predicting Default Rates on Customer Loans", *Review of Business Information Systems*, Vol. 6, No. 3, pp. 65-83, 2002.
39. Zurada, J., and Subhash, L., "Application of Data Mining Methods for Bad Debt Recovery in the Healthcare Industry", in Proceedings of the 6th International Conference on Databases and Information Systems, J. Barzdins (Ed.), Vol. 672, pp. 207-217, 2004.

NOTES