Distinguishing the Forest from the TREES: A Comparison of Tree Based Data Mining Methods

Richard Derrig, Ph.D. and Louise Francis, FCAS, MAAA

Richard Derrig, PhD,
OPAL Consulting LLC
41 Fosdyke Street
Providence
Rhode Island, 02906, U.S.A.
Phone: 001-401-861-2855
email: richard@derrig.com

Louise Francis, FCAS, MAAA
Francis Analytics & Actuarial Data Mining
706 Lombard Street
Philadelphia
Pennsylvania, 19147, U.S.A.
Phone: 001-215-923-1567
email: louise_francis@msn.com

Abstract

In recent years a number of "data mining" approaches for modeling data containing nonlinear and other complex dependencies have appeared in the literature. One of the key data mining techniques is decision trees, also referred to as classification and regression trees or CART (Breiman et al, 1993). That method results in relatively easy to apply decision rules that partition data and model many of the complexities in insurance data. In recent years considerable effort has been expended to improve the quality of the fit of regression trees. These new methods are based on ensembles or networks of trees and carry names like TREENET and Random Forest. Viaene et al (2002) compared several data mining procedures, including tree methods and logistic regression, for prediction accuracy on a small fixed data set of fraud indicators or "red flags". They found simple logistic regression did as well at predicting expert opinion as the more sophisticated procedures. In this paper we will introduce some available regression tree approaches and explain how they are used to model nonlinear dependencies in insurance claim data. We investigate the relative performance of several software products in predicting the key claim variables for the decision to investigate for excessive and/or fraudulent practices, and the expectation of favorable results from the investigation, in a large claim database. Among the software programs we will investigate are CART, S-PLUS, TREENET, Random Forest and Insightful Miner Tree procedures. The data used for this analysis are the approximately 500,000 auto injury claims reported to the Detailed Claim Database (DCD) of the Automobile Insurers Bureau of Massachusetts from accident years 1995 through 1997. The decision to order an independent medical examination or a special investigation for fraud, and the favorable outcomes of such decisions, are the modeling targets. We find that the methods all provide some predictive value or lift from the available DCD variables with significant differences among the methods and the four targets. All modeling outcomes are compared to logistic regression as in Viaene et al. with some model/software combinations doing significantly better than the logistic model.

Keywords: Fraud, Data Mining, ROC Curve, Variable Importance, Decision Trees

INTRODUCTION

In recent years a number of approaches for modeling data containing nonlinear and other complex dependencies have appeared in the literature. Many of the methods were developed by researchers from the computer science, artificial intelligence and statistics disciplines[1]. The methods have been widely characterized as *data mining* techniques. These procedures include several that should be of interest to actuaries dealing with large and complex data sets. The procedures of interest for the purposes of this paper are various varieties of classification and regression trees or CART. Viaene et al (2002) applied a wider set of procedures, including neural networks, support vector machines, and a classical general linear model, logistic regression, on a small single data set of insurance claim fraud indicators or "red flags" as predictors of suspicion of fraud. They found simple logistic regression did as well at predicting expert opinion on the presence of fraud as the more sophisticated procedures. Stated differently, the logistic model performed well enough in modeling the expert opinion of fraud that there was little need for the more sophisticated procedures[2].

A wide variety of statistical software is now available for implementing fraud and other predictive models through clustering and data mining. In this paper we will introduce a variety of Regression Tree data mining approaches[3] and explain how they are used to model nonlinear dependencies in insurance claim data. We also investigate the relative performance of several software products that implement these models. As an example of relative performance, we test for the key claim variables in the decision to investigate for excessive and/or fraudulent practices in a large claim database. The software programs we will investigate are CART, S-PLUS, TREENET, Random Forests, and Insightful Tree and Ensemble from the Insightful Miner package. Naïve Bayes and Logistic models are used as benchmarks. The data used for this analysis are the auto bodily injury liability claims reported to the Detailed Claim Database (DCD) of the Automobile Insurers Bureau of Massachusetts from accident years 1995 through 1997[4]. Three types of variables are employed. Several variables thought to be related to the decision to investigate are included here as reported to the DCD, such as outpatient provider medical bill amounts. A few variables are included that are derived from publicly available demographic data sources, such as income per household for each claimant's zip code. Additional variables are derived by accumulating proportional statistics from the DCD; e.g., the distance from the claimant's zip code to the zip code of the first medical provider or claimant's zip code rank for the number of plaintiff attorneys per zip code. The decision to order an independent medical examination or a special investigation for fraud, and a favorable outcome for each, are the modeling target.

Eight modeling software results will be compared for effectiveness based on a standard procedure, the area under the receiver operating characteristic curve (AUROC). We find that the methods all provide some predictive value or lift from the DCD variables we make available, with significant differences among the eight methods and four targets. Modeling outcomes can be compared to logistic regression as in Viaene et al. but the results here are different. They show some software/methods can improve significantly on the predictive

ability of the logistic model. That result may be due to the relative richness of this data set and/or the types of independent variables at hand compared to the Viaene data. We show how "important" each variable is within each software/model tested[5] and note the type of data that are important for this analysis. This entire exercise should provide practicing actuaries with guidance on regression tree software and market methods to analyze complex nonlinear relationship commonly found in all types of insurance data.

The paper is organized as follows. Section 1 introduces the general notion of non-linear dependencies in insurance data. Section 2 describes the data set of Massachusetts auto bodily injury liability claims and variables used for illustrating the models and software implementations. Descriptions and illustrations of the data mining methods applied in the paper appear in Section 3 while the specific software procedures are covered in Section 4. Comparative outcomes for the variables ("importance") and software ("AUROC") are reported in Sections 5 and 6. We provide some interpretation of the results in terms of the decision to investigate within the Massachusetts data as an illustration of the usefulness of the modeling effort in Section 7. Implications for the use of the software models are discussed in section 8. Conclusions are shown in Section 9.

SECTION 1. NONLINEARITY IN INSURANCE DATA

Actuaries are nearly inseparable from data and data manipulation techniques. Data come in all forms as a matter of course. Numeric (loss ratios), categorical (injury types), and text (accident description) data all flood insurers on a daily basis. Reserving and pricing are two major functions of casualty actuaries. Reserving involves compiling and understanding through mathematical techniques historical patterns of a portfolio of insurance claims in order to predict an ultimate value. Pricing involves taking the best estimates of historical cost data on claims and expenses, combining that data with financial asset pricing models that include projecting future values in order to arrive at best estimates of all costs of accepting underwriting risk. Of course, actuaries continually look back at both analytic exercises to determine the accuracy of those estimates as the real accounting data develops over time.

Traditionally, actuarial models were confined to linear, multiplicative or mixed algebraic equations in the absence of the powerful computing environment we enjoy today. Those mostly manual methods provided crude approximations that sufficed when alternative methods were unavailable or non-existent. Simple deviations from linear relationships, such as escalating inflation, could be handled by simple transformations of the data (log transform) that allowed linear techniques to be applied to the data. Gradually, over time these transformation techniques became more sophisticated and could be applied to many problems with a variety of non-linear data[6].

Trend lines of time series data, such as claim severity or frequency, are generally amenable to linear techniques. However, data where interactions and cross correlations are essential to the modeling of the dynamics of the process underlying the data, require more

comprehensive techniques that yield more precision on more types of data complexities. Figure 1-1 shows a particular non-linear relationship between two insurance variables that would be difficult, if not impossible, to model with simple techniques. One purpose of this paper is to demonstrate a range of so-called artificial intelligence or statistical learning techniques that have been developed to handle complicated relationships within data sets.

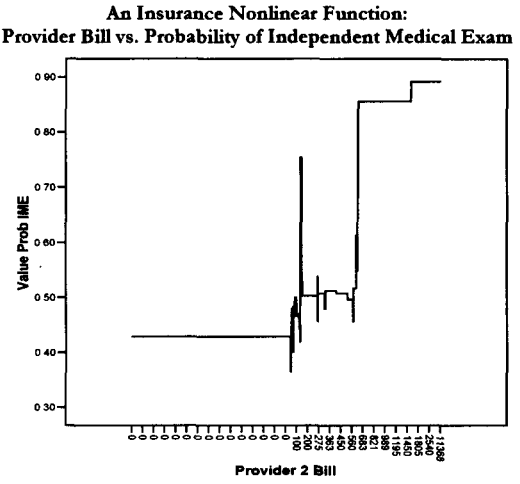

Figure 1 -1

Nearly all regression and econometric academic courses address the topic of nonlinearity, at least briefly. Students are instructed in methods to detect nonlinearity and how to model it. Detection generally involves using scatter plots of independent versus dependent variables or evaluating plots of residuals. Two methods of modeling nonlinearity that are generally taught: are 1) transformation of variables and 2) polynomial regression (Miller and Wichern[7], 1977, and Neter et al, 1985). For instance, if an examination of residual plots indicates that the magnitude of the residuals increases with the size of an independent variable, the log transformation is recommended. Polynomial regressions are considered useful approximations when a curvilinear relationship exists but its exact form is unknown.

A generalization of linear models known as Generalized Linear Models or GLM (McCullagh and Nelder, 1989) enabled the modeling of multivariate relationships in the presence of certain kinds of non-normality (i.e. where the random component is from the exponential family of distribution). The link function of GLMs formalizes the incorporation of certain nonlinear relationships into the modeling procedure: The transformations incorporated into the common GLMs are:

The identity link: $h(Y) = Y$

$$\text{The log link: } h(Y) = \ln(Y)$$
$$\text{The inverse link: } h(Y) = \frac{1}{Y} \qquad (1)$$
$$\text{The logit link: } h(Y) = \ln\left(\frac{Y}{1-Y}\right)$$
$$\text{The probit link: } h(Y) = \Phi(Y), \Phi \text{ denotes the normal CDF}$$

Of these transformations, the log and logit transformation appear frequently in the insurance literature. Because many insurance variables are right skewed, the log transformation is applied to attained approximate normality and homogeneity of variance. In addition, apriori or domain considerations (e.g., the relationship between the independent variables and the dependent variable is believed to be multiplicative) sometimes suggest the log transformation. The logit transform is commonly used when the dependent variable is binary.

Unfortunately, while the techniques cited above add significantly to the analyst's ability to model nonlinearity, they are not sufficient for many situations encountered in practice. In actual insurance data, complex nonlinear relationships are the rule rather than the exception. Some of the reasons the traditional approaches often do not provide a satisfactory approximation to nonlinear functions are:

- The form of the nonlinearity may be other than one of those permitted by the known transformations which produce linearity. Figure 1-1 displays one such non-linear function based on the insurance database used in this analysis.
- While a polynomial of adequate degree can approximate many complex functions, extrapolation beyond the data, or interpolation within the data, may be problematic, particularly for higher order polynomials.
- Determining the appropriate transformation (or polynomial) can be difficult if not impossible when there are many independent variables, and the appropriate relation between the target and each independent variable must be found.
- The relationship between a dependent variable and an independent variable may be confounded by a third variable due to interaction or correlations that are not simple to approximate.

To remedy these problems requires methods where:

- Any nonlinear relationship can be approximated.
- The analyst does not need to know the form of the nonlinearity.
- The effect of interactions can be easily determined and incorporated into the model.
- The method generalizes well on out-of-sample data for interpolation or extrapolation purposes.

The regression tree methods included in our analysis meet these conditions. Section 3 of this paper describes how each of our methods models nonlinearity. We now turn to a description of the data set we will use in this analysis.

SECTION 2. DESCRIPTION OF THE MASSACHUSETTS AUTO BODILY INJURY DATA

The database we will use for our analysis is a subset of the Automobile Insurers Bureau of Massachusetts Detail Claim Database (DCD); namely, those claims from accident years 1995-1997 that had closed by June 30, 2003 (AIB, 2004). All auto claims[8] arising from injury coverages: Personal Injury Protection (PIP)/ Medical payments excess of PIP[9], Bodily Injury Liability (BIL), Uninsured and Underinsured Motorist. While there are more than 500,000 claims in this subset of DCD data, we will restrict our analysis to the 162,761 third party BIL coverage claims. This will allow us to divide the sample into training, test, and holdout sub samples, each containing in excess of 50,000 claims[10]. The dataset contains fifty-four variables relating to the insured, claimant, accident, injury, medical treatment, outpatient medical providers (2 maximum), attorney presence, and three claims handling techniques for mitigating claims cost for their presence, outcome, and formulaic savings amounts.

The claims handling techniques tracked are: Independent Medical Examination (IME), Medical Audit (MA) and Special Investigation (SIU). IMEs are performed by licensed physicians of the same type as the treating physician[11]. They cost approximately $350 per exam with a charge of $75 for no shows. They are designed to verify claimed injuries and to evaluate treatment modalities. One sign of a weak or bogus claim is the failure to submit to an IME and, thus, an IME can serve as a screening device for detecting fraud and build-up claims. MAs are peer reviews of the injury, treatment and billing. They are typically done by physicians without a claimant examination, by nurses on insurers' staff or by third party organizations, but also from expert systems that review the billing and treatment patterns[12]. Favorable outcomes are reported by insurers when the damages are mitigated, the billing and treatment are curtailed, and when the claimant refuses to undergo the IME or does not show. In the latter two situations the insurer is on solid ground to reduce or deny payments under the failure-to-cooperate clause in the policy.[13]

Special Investigation (SIU) is reported when claims are handled through non-routine investigative techniques (accident reconstruction, examinations under oath and surveillance are examples), possibly including an IME or Medical Audit, on suspicion of fraud. For the most part, these claims are handled by Special Investigative Units (SIU) within the claim department or by some third party investigative service. Occasionally, companies will be organized so that additional adjusters, not specifically a part of the company SIU, may also conduct special investigations on suspicion of fraud. Both types are reported to DCD and we refer to both by the shorthand SIU in subsequent tables and figures. Favorable outcomes are reported for SIU if the claim is denied or compromised based on the SIU investigation.

For purposes of this analysis and demonstration of non-linear models and software, we employ twenty-one potentially predicting variables and four target variables. Thirteen predicting variables are numeric, two from DCD fields (F), eight derived from internal demographic type data (DV), and three variables derived from external demographic data (DM) as shown in Table 2-1.

Auto Injury Liability Claim Numeric Variables						
Variable	**N**	**Type**	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
Provider 1_BILL	162,761	F	0	1,861,399	2,671.92	6,640.98
Provider 2_BILL	162,761	F	0	360,000	544.78	1,805.93
Age	155,438	DV	0	104	34.15	15.55
Report Lag	162,709	DV	0	2,793	47.94	144.44
Treatlag	147,296	DV	1	9	3.29	1.89
HouseholdsPerZipcode	118,976	DM	0	69,449	10,868.87	5,975.44
AverageHouseValue Per Zip	118,976	DM	0	1,000,001	166,816.75	77,314.11
IncomePerHousehold Per Zip	118,976	DM	0	185,466	43,160.69	17,364.45
Distance (MP1 Zip to CLT. Zip)	72,786	DV	0	769	38.85	76.44
Rankatt1 (rank att/zip)	129,174	DV	1	3,314	150.34	343.07
Rankdoc2 (rank prov/zip)	109,387	DV	1	2,598	110.85	253.58
Rankcity (rank claimant city)	118,976	DV	1	1,874	77.37	172.76
Rnkpcity (rank provider city)	162,761	DV	0	1,305	30.84	91.65
Valid N (listwise)	70,397					
N = Number of non missing records; F=DCD Field, DV = Internal derived variable, DM = External derived variable						

Source: Automobile Insurers Bureau of Massachusetts, Detail Claim Database, AY 1995-1997 and Authors' Calculations.

Table 2-1

Eight predicting variables, and four target variables (IME and SIU, Decision and Favorable Outcome for each), are categorical variables, all taken as reported from DCD and as shown in Table 2-2.

Auto Injury Liability Claim Categorical Variables			
Variable	**N Type**	**Type**	**Description**
Policy Type	162,761	F	Personal 92%, Commercial 8%
Emergency Treatment	162,761	F	None 9%, Only 22%, w Outpatient 68%
Health Insurance	162,756	F	Yes, 15%, No 26%, Unknown 60%
Provider 1 – Type	162,761	F	Chiro 41%, Physical Th. 19%, Medical 30%, None 10%
Provider 2 – Type	162,761	F	Chiro 6%, Physical Th. 6%, Medical 36%, None 52%
2001 Territory	162,298	F	Rating Territories 1 (2.2%) Through 26 (1.3%); Territory 1-16 by increasing risk, 17-26 is Boston
Attorney	162,761	F	Attorney present (89%), no attorney (11%)
Susp1 (SIU Done)	162,761	F	Special Investigation Done (7%), No SIU (93%)
Susp2 (IME Done)	162,761	F	Independent Medical Examination Done (8%), No IME (92%)
Susp3 (SIU Favorable)	162,761	F	Special Investigation Favorable (3.4%), Not Favorable/Not Done (95.6%)
Susp4 (IME Favorable)	162,761	F	Independent Medical Exam Favorable (4.4%), Not Favorable/Not Done (96.6%)
Injury Type	162,298	F	Injury Types (24) including minor visible (4%), strain or sprain, back and/or neck (81%), fatality (0.4%), disk herniation (1%) and others
N = Number of non missing records F= DCD Field			
Note: Descriptive percentages may not add to 100% due to rounding			

Source: Automobile Insurers Bureau of Massachusetts, Detail Claim Database, AY 1995-1997 and Authors' Calculations.

Table 2-2

Similar claim investigation variables are now being collected by the Insurance Research Council in their periodic sampling of countrywide injury claims (IRC, 2004a, pp 89-104)[14]. Nationally, about 4% and 2% of BI claims involved IMEs and SIU respectively, only one-half to one-quarter of the Massachusetts rate. Most likely, this is because (1) a majority of other states have a full tort system and so BIL contains all injury claims and (2) Massachusetts is a fairly urban state with high claim frequencies and more dubious claims[15]. In fact, the most recent IRC study shows (IRC, 2004b, p25) Massachusetts has the highest percentage of BI claims in no-fault states that are suspected of fraud (23%) and/or buildup (41%). It is therefore, entirely consistent for the Massachusetts claims to exhibit more non-routine claim handling techniques. Favorable outcomes average about 67% when an IME is done or a claim is referred to SIU. We now turn to descriptions of the types of models, and the software that implements them, in the next two sections before we describe how they are applied to model the IME and SIU target variables.

SECTION 3. MODELS FOR NON-LINEAR DEPENDENCIES

How models handle nonlinearity

Traditional actuarial and statistical techniques often assume that the functional relationship between the independent variables and the dependent variable is linear or that some transformation of the data exists that can be treated as linear. Insurance data often contain

variables where the relationship among variables is nonlinear. Typically when nonlinear relationships exist, the exact nature of the nonlinearity (i.e., where some transformation can be used to establish linearity) is not known. In the field of data mining, a number of nonparametric techniques have been developed which can model nonlinear relations without any assumption being made about the nature of the nonlinearity. We cover how each of our methods reviewed in this paper models nonlinearities in the following two examples. The variables in this example were selected because of a known nonlinear relationship between independent and dependent variables.

Ex. 1 The dependent variable, a numeric variable, is total paid losses and the independent variable is provider 2 bill. Table 3-1 displays average paid losses at various bands of provider 2 bill[16].

Ex. 2 The dependent variable, a binary categorical variable, is whether or not an independent medical exam is requested and the independent variable again is provider 2 bill.

Nonlinear Example Data

Provider 2 Bill (Banded)	Avg Provider 2 Bill	Avg Total Paid	Percent IME
Zero	-	9,063	6%
1 – 250	154	8,761	8%
251 – 500	375	9,726	9%
501 - 1,000	731	11,469	10%
1,001 - 1,500	1,243	14,998	13%
1,501 - 2,500	1,915	17,289	14%
2,501 - 5,000	3,300	23,994	15%
5,001 - 10,000	6,720	47,728	15%
10,001+	21,350	83,261	15%
All Claims	545	11,224	8%

Table 3-1

Trees

Trees, also known as classification and regression trees (CART) fit a model by recursively partitioning the data into two groups, one group with a higher value on the dependent variable and the other group with a lower value on the dependent variable. Each partition of the tree is referred to as a node. When a parent node is split, the two children nodes, or "leaves" of the tree, are each more homogenous (i.e., less variable) with respect the dependent variable[17]. A goodness of fit statistic is used to select the split which maximizes the difference between the two nodes. When the independent variable is numeric, such as provider 2 bill, the split takes the form of a cutpoint, or threshold: $x \geq c$ and $x < c$ as in Figure 3-1.

CART Example of Parent and Children Nodes
Total Paid as a Function of Provider 2 Bill

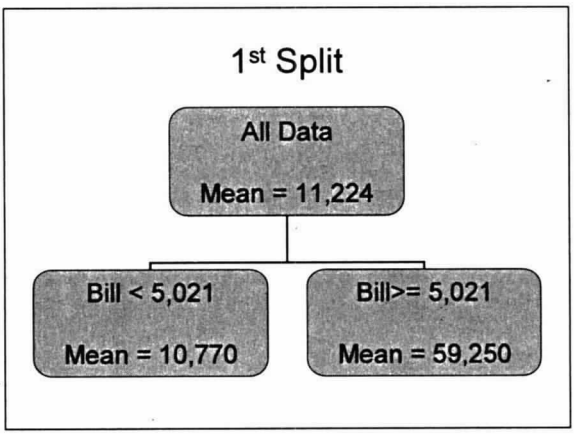

Figure 3-1

The cutpoint c is found by evaluating all possible values for splitting the numeric variable into higher and lower groups, and selecting the value that optimizes the split in some manner. When the dependent variable is numeric, the split is typically based on the value which results in the greatest reduction in residual sum of squares. For this example, all values of provider 2 bill are searched and a split is made at the value $5,021. All claims with provider 2 bill less than $5,021 go to the left node and "predict" a total paid of $10,770 and all claims with provider 2 bill greater than $5,021 go to the right node, and "predict" a total paid of $59,250. This is depicted in Figure 3-1. The tree graph shows that the total paid mean is significantly lower for the claims with provider 2 bills less than $5,021.

One statistic often used as a goodness of fit measure to optimize tree splits is sum squared error or the total squared deviation of actual values around the predicted values. The selected cutpoint is the one which produces the largest reduction in total sum squared errors (SSE). That is, for the entire database the total squared deviation of paid losses around the predicted value (i.e., the mean) of paid losses is 4.95×10^{13}. The SSE declines to 4.66×10^{13} after the data are partitioned using $5,021 as the cutpoint. Any other partition of the provider bill produces a larger SSE than 4.66×10^{13}. For instance, if a cutpoint of $10,000 is selected, the SSE is 4.76×10^{13}.

The two nodes in Figure 3-1 can each be split into to children nodes and these can then be further split. The sequential splitting continues until no improvement in the goodness of fit statistic occurs. The nodes containing the result of all the splits resulting from applying a sequence of decision rules are the final nodes often referred to as terminal nodes. The terminal nodes provide the predicted values of the dependent variables. When the dependent

variable is numeric, the mean of the dependent variable at the terminal nodes is the prediction.

The curve of the predicted value resulting from a tree fit to total paid losses is a step function. As shown in Figure 3-2A, with only two terminal nodes, the fitted function is flat until $5,021, steps up to a higher value and then remains flat. Figure 3-2B displays the predicted values of a tree with 7 terminal nodes. The steps or increases are more gradual for this function.

CART Example with Two and Seven Nodes
Total Paid as a Function of Provider 2 Bill

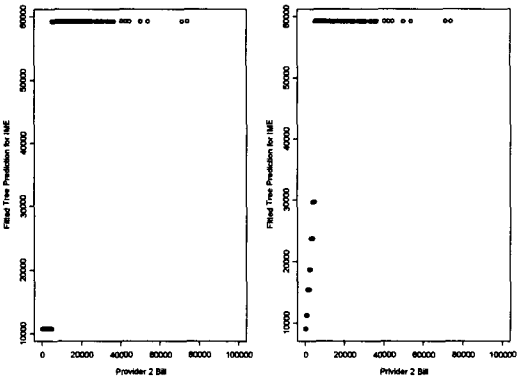

Figure 3-2A **Figure 3-2B**

The procedure for modeling data where the dependent variable is categorical (binary in our example) is similar to that of a numeric variable. For instance, one of the fraud surrogates is independent medical exam (IME) requested. The target class is claimants for whom an IME was requested and the non-target group of (presumably) legitimate claims is that where an IME was not requested. At each step, the tree procedure selects the split that best improves or lowers node impurity. That is, it attempts to partition the data into two groups so that one partition has a significantly higher proportion of the target category, IME requested, than the other node. A number of statistical goodness of fit statistics measures is used in different products to select the optimal split. These include entropy/deviance and Gini index (which is described later in this paper). Kantardzic (2003), Breiman et al (1993) and Venibles and Ripley (1999) describe the computation and application of the Gini index and entropy/deviance measures[18]. A score or probability can be computed for each node after a split is performed. This is generally estimated based on the number of observations in the target groups versus the total number of observations at the node. The score or probability

is frequently used to assign records to one of the two classes. Typically, if the model score exceeds a threshold such as 0.5, the record is assigned to the target class; otherwise it is assigned to the non-target class.

Figure 3-3A displays the result of using a tree procedure to predict a categorical variable from the AIB data. The graph shows that each time the data is split on provider 2 bill; one child node has a lower proportion and the other a higher proportion of claimants receiving IMEs. The fitted tree function is used to model a nonlinear relationship between provider bill and the probability that a claim receives an IME as shown in Figure 3-3B.

CART Example with Seven Nodes
IME Proportion as a Function of Provider 2 Bill

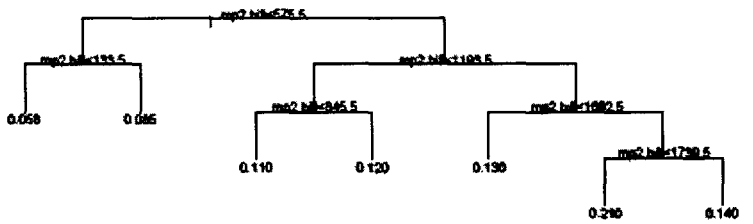

Figure 3-3A

CART Example with Seven Step Functions
IME Proportion as a Function of Provider 2 Bill

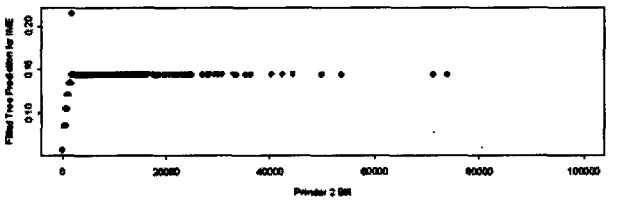

Figure 3-3B

Tree models use categorical as well as numeric independent variables in modeling complex data. However, because the levels on categorical data may not be ordered, all possible two-way splits of categorical variables must be considered before the data are partitioned.

Ensemble Models-Boosting

Ensemble models are composite tree models. A series of trees is fit and each tree improves the overall fit of the model. In the data mining literature the technique is often referred to as

"boosting" (Hastie et al 2001, Freidman, 2001). The method initially fits a small tree of say 5 to 10 terminal nodes on a training dataset. Typically, the user specifies the number of terminal nodes, and every tree fit has the same number of terminal nodes. The error, or difference between the actual and fitted values, is computed and used in another round of fitting as a dependent variable. The error is also used in the computation of the weight in subsequent rounds of fitting, with records containing larger errors receiving higher weighting in the next round of estimation.

One algorithm for computing the weight is described by Hastie et al[19]. Consider an ensemble of trees 1, 2, ...,M. The error for the m^{th} tree measures the departure of the actual from the fitted value on the test data after the m^{th} model has been fit. When the dependent variable is categorical, as it is in the fraud application in this paper, a common error measure used in boosting is:

$$err_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq F_m(\mathbf{x}_i))}{\sum_{i=1}^{N} w_i} \qquad (2)$$

where N is the total number of records, w_i is a weight (which is initialized to 1/N in the first round of fitting), I is an indicator function equal to zero if the category is correctly predicted and one if the class assigned is incorrect, y_i is the dependent variable, $\mathbf{x}$ is a matrix of predictors and $F_m(\mathbf{x})$ is the prediction for the i^{th} record of the m^{th} tree.

Then, the coefficient alpha is a function of the weight:

$$\alpha_m = \log(\frac{1-err_m}{err_n})$$

and the new weight is: (3)

$$w_{i,m+1} = w_m \exp(\alpha_m I(y_i \neq F_m(\mathbf{x})))$$

The process is performed many times until no further statistical improvement in the fit is obtained.

The specific boosting procedures implemented differ among different software products. For instance, TREENET (Freidman, 2001) uses stochastic gradient boosting. Stochastic gradient boosting incorporates a number of procedures which attempt to build a more robust model by controlling the tendency of large complex models to overfit the data. A key technique used is resampling. A new sample is randomly drawn from the training data each time a new tree is fit to the residuals from the prior round of model estimation. The goodness of fit of the model is assessed on data not included in the sample, the test data. Another procedure used by TREENET to control overfitting is *shrinkage* or *regularization*. A simple way to implement shrinkage is to apply a weight which is greater than zero and less than one to the contribution of each tree as it is added to the weighted average estimate.

Alternatively, the Insightful Miner Ensemble model employs a simpler implementation of boosting which applies non-stochastic boosting and uses all the training data in each round of fitting.

The final estimate resulting from an ensemble approach will be a weighted average of all the trees fit. Using a large collection of trees allows:

- Many different variables to be used. Some of these would not be used in smaller models[20].
- Many different models are used. The predictive modeling literature (Hastie et al., 2001, Francis, 2003a, 2003c) indicates that composites of multiple models perform better than the prediction of a single model[21].
- Different training and test records are used (with stochastic gradient boosting). This makes the procedure more robust to the influence of a few extreme observations.

The method of fitting many (often 100 or more) small trees results in fitted curves which are almost smooth. Figures 3-4A and 3-4B display two nonlinear functions fit to total paid and IME variables by the TREENET ensemble model.

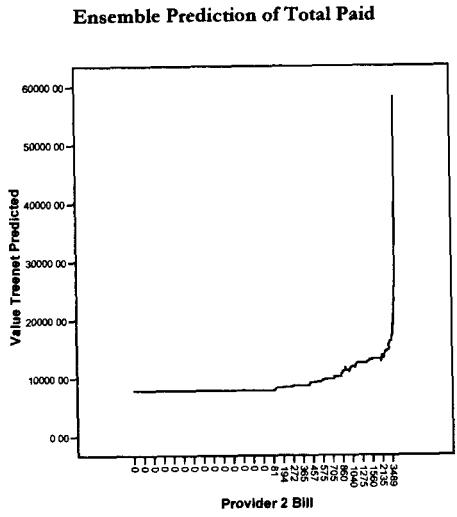

Figure 3-4A

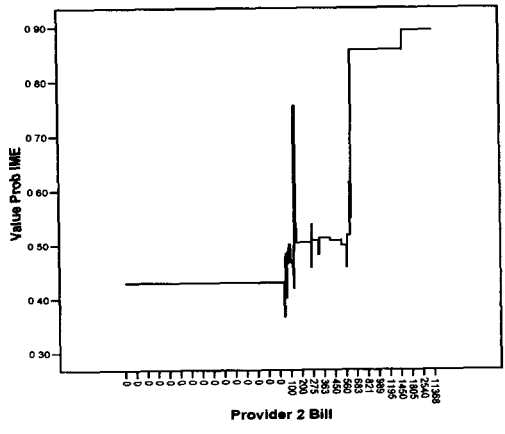

Figure 3-4B

Ensemble Models-Bagging

Bagging is an ensemble approach based on resampling or bootstrapping. Bagging is an acronym for "bootstrap aggregation" (Hastie et al., 2000). Bagging does not use the error from the prior round of fitting as a dependent variable or weight in subsequent rounds of fitting. Bagging uses recursive sampling of records in the data to fit many trees. For instance an analyst may decide to take a 50% of the data as a training set each time a model

is fit. Under bagging, 100 or more models may be fit, each one to a different sample. The trees fit are not necessarily small trees with 5 to 10 terminal nodes as with boosting and each tree may have a different number of terminal nodes. By averaging the predictions of a number of bootstrap samples, bagging reduces the prediction variance. The implementation of bagging used in this paper is known as Random Forest. In addition to using only a sample of the data each time a tree model is fit, Random Forest also samples from the variables. For the analysis in this paper, one third of the variables were sampled for each tree fit.

Figures 3-5A displays an ensemble Random Forest tree fit to total paid losses and Figure 3-5B displays a tree fit to IME.

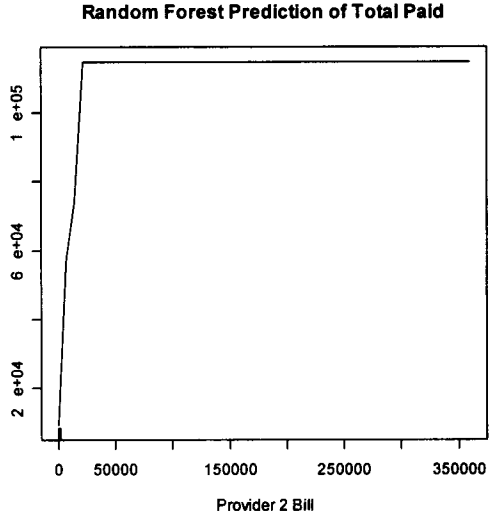

Figure 3-5 A

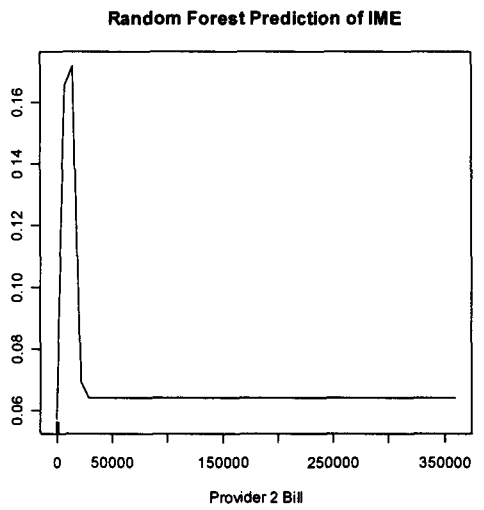

Figure 3-5 B

Naive Bayes

The Naïve Bayes method is a relatively simple and easy to implement method. In our comparison, we view it as a benchmark data mining method. That is, we are interested in how more complex methods improve performance (or not) against an approach where simplifying assumptions are made in order to make the computations more tractable. We also use logistic regression models as a second benchmark.

The Naïve Bayes method was developed for categorical data. Specifically, both dependent and independent variables are categorical. Therefore, its application to fitting nonlinear functions will be illustrated only for the categorical target variable IME. In order to utilize numeric predictor variables it was necessary to derive new categorical variables based on discretizing, or "binning", the distribution of data for the numeric variables[22].

The key simplifying assumption of the Naïve Bayes method is the assumption of independence. All predictor variables are assumed to act independently in influencing the target variable. Interactions and correlations among the predictor variables are not considered.

Bayes rule is used to estimate the probability that a record with given independent variable vector $X = \{x_i\}$ is in category $C = \{c_j\}$ of the dependent variable.

$$P(c_j \mid x_i) = P(x_i \mid c_j)P(c_j) / P(x_i) \qquad (4a)$$

Because of the Naive Bayes assumption of conditional independence, the probability that an observation will have a specific set of values for the independent variables is the product of the conditional probabilities of observing each of the values given category c_j

$$P(X \mid c_j) = \prod_I P(x_i \mid c_j) \qquad (4b)$$

The method is described in more detail in Kantardzic (2003). To illustrate the use of Naive Bayes in predicting discrete variables, the provider 2 bill data was binned into groups based on the quintiles of the distribution. Because about 50 percent of the claims have a value of zero for provider 2 bill, only four categories are created by the binning procedure. The new variable was used to estimate the IME targets. Figure 3-6 displays a bar plot of the predicted probability of an IME for each of the groups. Figure 3-7 displays the fitted function. This function is a step function which changes value at each boundary of a provider 2 bill bin.

Bayes Predicted Probability IME Requested vs. Quintile of Provider 2 Bill

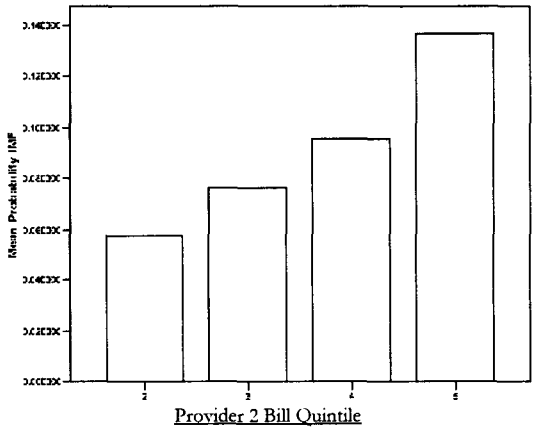

Figure 3-6

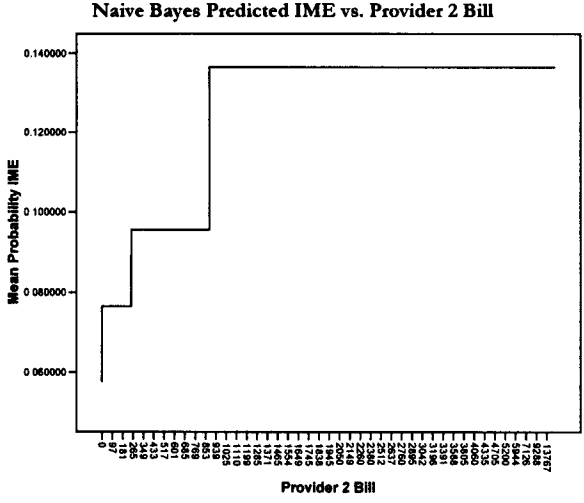

Figure 3-7

SECTION 4. SOFTWARE FOR MODELING NON-LINEAR DEPENDENCIES

Nonadditivity: interactions

Conventional statistical models such as regression and logistic regression assume not only linearity, but also additivity of the predictor variables. Under additivity, the effect of each variable can be can be added to the model one at a time. When the exact form of the relationship between a dependent and independent variable depends on the value of one or more other variables, the effects are not additive and one or more interactions exist. For instance, the relationship between provider 2 bill and IME may vary by type of injury (i.e. traumatic injuries versus sprains and strains). Interactions are common in insurance data (Weisberg and Derrig, 1998, Francis, 2003c).

With conventional linear statistical models, interactions are incorporated with multiplicative terms:

$$Y = a + b_1X_1 + b_2X_2 + b_3{*}X_1{*}X_2 \qquad (5)$$

In the case of a two-way interaction, the interaction terms appear as products of two variables. If one of the two variables is categorical, the interaction terms allow the slope of the fitted line to vary with the levels of the categorical variable. If both variables are continuous the interaction is a bilinear interaction (Jicard and Turrisi, 2003) and the slope of one variable changes as a linear function of the other variable. If both variables are categorical the model is equivalent to a two factor ANOVA with interactions.

The conventional approach to handling interactions has some limitations.

- Only a limited number of types of interactions can be modeled easily.
- If many predictor variables are included in the model, as is often the case in many predictive modeling applications, it can be tedious, if not impossible, to find all the significant interactions. Including all possible interactions in the model without regard to their significance likely results in a model which is over-parameterized.

The tree-based data mining techniques used in this paper each have efficient methods for handling interactions.

- Interactions are inherent in the method used by trees to partition data. Once data have been partitioned, different partitions can and typically do split on different variables and capture different interactions among the predictor variables. When the decision rules used by a tree to reach a terminal node involve more than one variable, in general, an interaction is being modeled.
- Ensemble methods incorporate interactions because they are based on the tree approach.
- Naïve Bayes, because it assumes conditional independence of the predictors, ignores interactions.
- Logistic regression incorporates interactions in the same way ordinary least squares regression does, with product interaction terms. In this fraud comparison study, no attempt was made to incorporate interaction terms as this procedure lacks an efficient way to search for the significant interactions.

Multiple predictors

Thus far, the discussion of the tree-based models concerned only simple one or two variable models. Extending the tree methods to incorporate many potential predictors is straightforward. For each tree fit, the method proceeds as follows:

- For each variable determine the best two-way partition of the data.
- Select the variable which produces the best improvement in the goodness of fit statistic to split the data at a particular node.
- Repeat the process until no further improvement in fit can be obtained.

Software for modeling nonlinear dependencies and testing the models

Four software products were included in our fraud comparison: They are CART, TREENET,
S-PLUS (R) and Insightful Miner[23].

CART and TREENET are Salford Systems stand-alone software products that each performs one technique. CART (Classification and Regression Trees) does tree analysis and TREENET applies stochastic gradient boosting using the method described by Freidman (2001). All the software tested produce SAS code[24] that can be used to implement the model

in a production stage. All the products contain a procedure for handling missing values using surrogate variables. At any given split point, CART and TREENET find the variable that is next in importance in influencing the target variable and they use this variable to replace the missing data. The specific statistic used to rank the variables and find the surrogates is described in Brieman et. al. (1993). Different versions of CART and TREENET handle different size databases. The number of levels of categorical variables affects how much memory is needed, as more levels necessitate more memory. The 128k version of each product was used for this analysis. With approximately 100,000 records in the training data, occasional memory problems were experienced and it became necessary to sample fewer records. One of the very useful features of the Salford Systems software is that all the products rank variables in importance[25].

S-PLUS and R are comprehensive statistical languages used to perform a range of statistical analyses including exploratory data analysis, regression, ANOVA, generalized linear models, trees and neural networks. Both S-PLUS and R are derived from S, a statistical programming language originally developed at Bell Labs. The S progeny, S-PLUS and R, are popular among academic statisticians. S-PLUS is a commercial product sold by Insightful which has a true GUI interface that facilitates easier handling of some functions. Insightful also supplies technical support. The S-PLUS programming language is widely used by analysts who do serious number crunching. They find it more effective, especially for processes that are frequently repeated. R is free open source statistical software that is supported largely by academic statisticians and computer science faculty. It has only limited GUI functionality and the data mining functions must be accessed through the language. Most code written for S-PLUS will also work for R. One notable difference is that data must be converted to text mode to be read by R (a bit of an inconvenience, but usually not an insurmountable one). Fox (2002) points out some of the differences between the two languages, where they exist. The S-PLUS procedures used here in the fraud comparison are found in both S-PLUS and R. However one ensemble tree method used, Random Forest, appears only to be available in R. The S-PLUS (R) procedures used were: the tree function for decision trees and the glm (generalized linear models) for logistic regression. S-PLUS (R) incorporates relatively crude methods for handling missing values. These include eliminating all records with a missing value on any variable, an approach which is generally not recommended (Francis 2005, Allsion 2002). S-PLUS also creates a new category for missing values (on categorical variables) and allows aborting the analysis if a missing value is found. In general, it is necessary to preprocess the data (at least the numeric variables where there is no missing value method[26]) to make a provision for the missing values. In the fraud comparison, a constant not in the range of the data was substituted into the variable and an indicator dummy variable for missing was created for each numeric variable with missing values. S-PLUS and R are generally not considered optimal choices for analyzing large databases. After experiencing some difficulty reading training data of about 100,000 records into S-PLUS, the database was reduced to contain only the variables used in the analysis. Once the data was read into S-PLUS, few problems were experienced. Another eccentricity is that the S-PLUS tree function can only handle 32 levels on any given categorical variable, so in the preprocessing the number of levels may need to be reduced[27]. The R Random Forest function incorporates a procedure that can be used to rank variables in importance. The

procedure produces an *impurity* statistic which can be used to rank the variables. The impurity is based on the Gini index for classification applications and mean squared error for numeric dependent variables. The S-PLUS tree function contains no built-in capability for ranking variables in importance. Therefore using the S-PLUS language, an algorithm was coded into S-PLUS to rank the variables. The method is described in Francis (2001) and Potts (2000). The procedure quantifies how much the error increases when a variable is removed from the model; the larger the increase in errors, the more important the variable.

The Insightful Miner is a data mining suite that contains the most common data mining tools: regression, logistic regression, trees, ensemble trees, neural networks and Naïve Bayes[28]. As mentioned earlier, Insightful also markets S-PLUS. However, the Insightful Miner has been optimized for large databases and contains methods (Naïve Bayes) which are not part of S-PLUS (R). The Naïve Bayes, Tree and Ensemble Tree procedures from Insightful Miner are used here in the fraud comparison. The insightful Miner has several procedures for automatically handling missing values. These are 1) drop records with missing values, 2) randomly generate a value, 3) replace with the mean, 4) replace with a constant and 5) carry forward the last observation. Each missing value was replaced with a constant. In theory, the data mining methods used, such as trees, should be able to partition records coded for missing from the other observations with legitimate categorical or numeric values and separately estimate their impact on the target variable (possible after allowing for interactions with other variables). Server versions of the Insightful Miner generate C code that can be used in deploying the model, but the version used in this analysis did not have that capability. As mentioned above some preprocessing was necessary for the Naïve Bayes procedure. Since Insightful Miner contains no procedure for ranking variables in importance, no rankings were provided for the Iminer methods.

Validating and Testing

It is common in data mining circles to partition the data into three groups (Hastie et al., 2001). One group is used for "training", or fitting the model. Another group, referred to as the validation set, is used for "testing" the fit of the model and re-estimating parameters in order to obtain a better model. It is common for a number of iterations of testing and fitting to occur before a final model is selected. The third group of data, the "holdout" sample, is used to obtain an unbiased test of the model's accuracy. An alternative approach to a validation sample that is especially appropriate when the sample size used in the analysis is relatively modest, is cross-validation. Cross-validation is a method involving holding out a portion of the training sample, say one fifth of the data, fitting a model to the remainder of the data and testing it on the held out data. In the case of 5-fold cross validation, the process is repeated five times and the average goodness of fit of the five validations is computed. The various software products and procedures have different methods for validating the models. Some (Insightful Miner Tree) only allow cross-validation. Others (TREENET) use a validation sample. S-PLUS (R) allows either approach[29] to be used (so a test sample of about 20% of the training data was used as we had a relatively large database). Neither validation sample nor cross-validation was used with Naïve Bayes, Logistic Regression or the Ensemble Tree.

In this analysis, approximately a third of the data, about 50,000 records, was used as the holdout sample for the final testing and comparison of the models. Two key statistics often used to compare models accuracy are sensitivity and specificity. *Sensitivity* is the percentage of events (i.e., claims with an IME or referred to a special investigation unit) that were predicted to be events. The *specificity* is the percentage of nonevents (in our applications claims believed to be legitimate) that were predicted to be nonevents. Both of these statistics should be high for a good model. Table 4-1, often referred to as a confusion matrix (Hastie et. al., 2001), presents an example of the calculation.

Sample Confusion Matrix: Sensitivity and Specificity

	True Class		
Prediction	**No**	**Yes**	**Row Total**
No	800	200	1,000
Yes	200	400	600
Column Total	1,000	600	

	Correct	**Total**	**Percent Correct**
Sensitivity	800	1,000	80%
Specificity	400	600	67%

Table 4-1

In the example confusion matrix, 800 of 1,000 non-events are predicted to be non-events so the sensitivity is 80%. The specificity is 67% since 400 of 600 true positives are accurately predicted.

SECTION 5. SOFTWARE RANKINGS OF "IMPORTANT" VARIABLES IN THE DECISION TO INVESTIGATE: IME AND SIU

The remainder of this paper is devoted to illustrating the usefulness and effectiveness of eight model/software combinations applied to our Example 2, the decision to investigate via IMEs or referral to SIU. We model the presence and proportion of favorable outcomes, of each investigative technique for the DCD subset of automobile bodily injury liability (third party) claims from 1995-1997 accident years.[30] We employ twenty-one potentially predicting variables of three types: (1) eleven typical claim variable fields informative of injury claims as reported, both categorical and numeric, (2) three external demographic variables that may play a role in capturing variations in investigative claim types by geographic region of Massachusetts, and (3) seven internal "demographic" variables derived from informative pattern variables in the database. Variables of type 3 are commonly used in predictive modeling for marketing purposes. The variables used for these illustrations are by no means optimal choices for all three types of variables. Optimization can be approached by other procedures (beyond the scope of this paper) that maximize information and minimize cross correlations and by variable construction and selection by domain experts.

The eight model/software combinations we will use here are of two categories: six tree models, and two benchmark models (Naïve Bayes and Logistic). They are:

1) TREENET
2) Iminer Tree
3) SPLUS Tree
4) CART
5) Iminer Ensemble
6) Random Forest
7) Naïve Bayes
8) Logistic

As described in Section 4, CART and TREENET are Salford Systems stand-alone software products that each performs one technique. CART (Classification and Regression Trees) does tree analysis, and TREENET applies stochastic gradient boosting to an ensemble of trees using the method described by Freidman (2001). The S-PLUS procedures used here in the fraud comparison are found in both S-PLUS and in a freeware version in R. These were: the tree function for decision trees, and the GLM (generalized linear models) for logistic regression.

Insightful Miner is a data mining suite. The Naïve Bayes, Tree and Ensemble Tree procedures, from Insightful Miner are used here in the fraud comparison.

Model performance is covered in the next section, section 6, as we first cover the ranking of variables by "importance" in relation to the target variables: the decision to perform an IME or a Special Investigation (SIU) and the favorable outcomes of each investigative technique. The training data of approximately 75,000 records was used in the ranking evaluations.

Data mining models are typically complex models where it is difficult to determine the relevance of predictors to the model result. One of the handy tasks that some of the data

mining software products perform is to rank the predictor variables by their importance to the model in predicting the dependent variable. Where the software did not supply a ranking, we omitted an importance ranking leaving five model/software determinations of importance for the twenty-one variables. Different procedures are used for different methods and different products.

Two software products, CART and TREENET supply importance rankings. The procedures used are:

CART: CART uses a goodness of fit measure, also referred to in the literature as an impurity measure, and computed over the entire tree, to determine a variable's importance. In this study the goodness of fit measure was the Gini Index defined below (Hastie, et al., p.271-272):

$$i(t) = 1 - \sum_i p_i \quad \text{i=the categories of the dependent variable and } p_i \text{ is the probability of class} \quad (6)$$

Each split of the tree lowers the overall value for the statistic. CART keeps track of the impurity improvement at each node for both the variable used in the split and for surrogate variables used as a replacement in the case of missing values. A consequence of this is that a variable not used for splitting may rank higher in importance than a variable that is.

TREENET: Because it is composed of many small CART trees, TREENET uses the same method as CART to compute importance rankings.

S-PLUS (R) does not supply an importance ranking, but the programming language can be used to program a procedure to compute rankings. A sensitivity value was computed for each variable in the model. The sensitivity is a measure of how much the predicted value's error increases when the variables are excluded from the model one at a time. However, instead of actually excluding variables and refitting the model, their values are fixed at a constant value. (See Francis, 2001 for a detailed recipe for applying the approach). The sensitivity statistic was used to rank the variables from the tree function. For the logistic regression, information about the variables contribution to sum of squared variation explained by the model was used to rank it. Like CART and TREENET, Random Forest uses an impurity measure (i.e., Gini Index) to produce an importance ranking.

Insightful Miner does not supply importance rankings. Unlike S-PLUS (R), the analytical methods are not accessed through the language but through a series of icons placed on a palate. Thus, we were not able to custom program a ranking procedure for application with the Iminer's modeling methods. The resulting importance rankings were used in Tables 5-1A & 5-2A for the decisions to investigate and 5-1B and 5-2B for the favorable outcomes.

Each of five model/software combination outputs allowed for the evaluation of the predicting variables in rank order of importance, when significant, together with a measure of the relative value of importance on a scale of zero (insignificant) to 100 (most significant

variable). Table 5-1A displays the importance results for predicting an IME using the five tree models while Table 5-1B displays those results for the remaining five model/software combinations, including the benchmark Naïve Bayes and Logistic. The predicting variables are listed in the order of importance in the TREENET model, where all variables are significant. The number of significant variables found ranges from a low of twelve variables (S-PLUS Tree) to all twenty one (TREENET).

Software Ranking of Variables for IME Decision By Importance Rank and Value					
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
Variable	**TREENET**	**S Plus Tree**	**CART**	**Random Forest**	**Logistic**
Provider 2 Bill	1 (100)	2 (91)	1 (100)	1 (100)	10 (1)
Attorneys Per Zip	2 (80)	5 (26)	13 (9)	6 (34)	11 (1)
Territory	3 (71)	4 (32)	11 (11)	3 (59)	*
Health Insurance	4 (61)	1 (100)	3 (68)	2 (84)	1 (100)
Injury Type	5 (50)	6 (24)	5 (47)	10 (18)	2 (51)
Provider 1 Bill	6 (47)	3 (51)	4 (58)	4 (59)	*
Provider 1 Type	7 (31)	9 (7)	*	12 (15)	6 (8)
Report Lag	8 (31)	7 (16)	8 (18)	8 (27)	13 (1)
Attorney	9 (25)	12 (3)	*	19 (5)	5 (18)
Age	10 (23)	*	17 (2)	17 (8)	*
Provider 2 Type	11 (19)	8 (9)	*	5 (42)	3 (47)
Income Household/Zip	12 (18)	*	10 (13)	11 (16)	9 (2)
Avg. Household Price/Zip	13 (17)	*	15 (5)	*	*
Providers per City	14 (17)	*	9 (15)	16 (9)	*
Claimants per City	15 (16)	*	*	7 (32)	12 (1)
Providers/Zip	16 (16)	*	*	15 (13)	8 (2)
Households/Zip	17 (16)	11 (3)	*	13 (15)	7 (2)
Treatment Lag	18 (14)	10 (4)	18 (2)	9 (24)	4 (24)
Distance MP1 Zip to Clt Zip	19 (13)	*	20 (0.1)	14 (14)	*
Emergency Treatment	20 (4)	*	7 (20)	18 (6)	*
Policy Type	21 (3)	*	19 (2)	20 (0)	*

Note: * represents insignificance of variable in the model.

Table 5-1A

The same set of model/software combinations was used with the same set of twenty-one predicting variables to predict the favorable outcome of the IME. Table 5-1B shows the importance of each of the 21 predictors for modeling favorable outcomes of IMEs.

Software Ranking of Variables for IME Favorable By Importance Rank and Value					
	(1)	(2)	(3)	(4)	(5)
Variable	**TREENET**	**S Plus Tree**	**CART**	**Random Forest**	**Logistic**
Provider 2 Bill	5 (64)	3 (22)	4 (37)	5 (49)	2 (13)
Attorneys Per Zip	11 (28)	*	11 (6)	13 (28)	11 (1)
Territory	2 (98)	2 (43)	12 (5)	1 (100)	4 (9)
Health Insurance	1 (100)	1 (100)	1 (100)	2 (71)	1 (100)
Injury Type	4 (76)	5 (10)	9 (15)	4 (67)	3 (13)
Provider 1 Bill	7 (45)	4 (15)	2 (51)	3 (70)	*
Provider 1 Type	8 (38)	9 (16)	5 (36)	10 (32)	5 (9)
Report Lag	6 (53)	8 (7)	18 (0)	6 (45)	8 (6)
Attorney	12 (25)	*	*	18 (3)	7 (8)
Age	13 (24)	*	19 (0)	9 (33)	*
Provider 2 Type	10 (29)	*	6 (30)	12 (31)	*
Income Household/Zip	20 (7)	11 (4)	17 (0)	8 (33)	10 (2)
Avg. Household Price/Zip	15 (16)	*	15 (0)	*	*
Providers per City	19 (8)	*	8 (17)	15 (23)	*
Claimants per City	9 (36)	12 (3)	13 (2)	16 (22)	13 (1)
Providers/Zip	17 (12)	13 (2)	7 (20)	11 (31)	*
Households/Zip	16 (15)	7 (7)	16 (0)	7 (37)	9 (2)
Treatment Lag	14 (22)	14 (1)	10 (6)	14 (28)	6 (8)
Distance MP1 Zip to Clt Zip	3 (78)	6 (8)	14 (1)	*	*
Emergency Treatment	18 (9)	10 (6)	3 (44)	17 (5)	12 (1)
Policy Type	21 (5)	*	*	*	*

Note: * represents insignificance of variable in the model.

Table 5-1B

The same set of five model/software combinations was used with the same set of twenty-one predicting variables to predict the use of special investigation or SIU. Tables 5-2A and 5-2B show the corresponding ranking of variables by importance for each of the five model combinations and two target variables, decision and favorable.

Software Ranking of Variables for SIU Decision By Importance Rank and Value					
	(1)	(2)	(3)	(4)	(5)
Variable	**TREENET**	**S Plus Tree**	**CART**	**Random Forest**	**Logistic**
Providers/Zip	1 (100)	1 (100)	8 (37)	3 (74)	*
Provider 2 Type	2 (98)	10 (3)	15 (34)	10 (30)	6 (39)
Territory	3 (92)	5 (18)	3 (84)	1 (100)	*
Health Insurance	4 (64)	3 (33)	7 (52)	6 (50)	7(28)
Provider 1 Bill	5 (59)	2 (51)	2 (85)	2 (89)	14 (2)
Injury Type	6 (52)	7 (6)	5 (59)	16 (5)	2 (71)
Attorney	7 (47)	8 (4.5)	17 (13)	18 (4)	3 (63)
Provider 1 Type	8 (38)	4 (29)	4 (69)	5 (51)	1 (100)
Age	9 (31)	*	*	17 (5)	*
Provider 2 Bill	10 (30)	*	1 (100)	4 (74)	13 (5)
Report lag	11 (28)	*	6 (54)	8 (10)	11 (17)
Average House Price	12 (28)	*	15 (18)	*	*
Attorneys/zip	13 (22)	6 (8)	14 (20)	9 (30)	12 (7)
Distance to Provider	14 (20)	*	19 (4)	15 (18)	4 (58)
Emergency Treatment	15 (19)	*	13 (27)	19 (4)	5 (49)
Income/Cap Household	16 (18)	11 (3)	9 (4.5)	13 (21)	9 (27)
Claimants per City	17 (17)	*	12 (30)	11 (26)	*
Treatment Lag	18 (16)	9 (34)	18 (12)	14 (20)	15 (2)
Households/Zip	19 (16)	*	16 (16)	12 (21)	8 (28)
Policy Type	20 (8)	*	*	20 (1)	*
Providers per City	21 (6)	12 (1)	11 (30)	7 (44)	10 (22)

Note: * represents insignificance of variable in the model.

Table 5-2A

Software Ranking of Variables for SIU Favorable By Importance Rank and Value					
	(6)	(7)	(8)	(9)	(10)
Variable	TREENET	S Plus Tree	CART	Random Forest	Logistic
Providers/Zip	10 (20)	10 (6)	12 (25)	7 (24)	13 (2)
Provider 2 Type	4 (41)	*	7 (35)	9 (18)	5 (21)
Territory	1 (100)	2 (94)	1 (100)	1 (100)	1 (100)
Health Insurance	13 (18)	6 (16)	*	15 (10)	7 (19)
Provider 1 Bill	6 (30)	13 (4)	15 (9)	5 (29)	14 (1)
Injury Type	3 (58)	5 (16)	6 (39)	16 (8)	3 (41)
Attorney	14 (16)	12 (4)	9 (27)	18 (6)	6 (20)
Provider 1 Type	5 (40)	1 (100)	3 (50)	3 (33)	2 (45)
Age	8 (22)	*	17 (7)	13 (13)	11 (2)
Provider 2 Bill	2 (66)	4 (18)	8 (32)	6 (26)	9 (3)
Report lag	7 (25)	7 (14)	19 (2)	2 (36)	12 (2)
Average House Price	15 (16)	*	13 (24)	*	*
Attorneys/zip	11 (19)	8 (14)	4 (45)	10 (17)	*
Distance to Provider	16 (15)	9 (14)	5 (39)	*	*
Emergency Treatment	21 (9)	3 (72)	14 (17)	14 (11)	4 (25)
Income/Cap Household	17 (14)	11 (5)	2 (61)	11 (16)	*
Claimants per City	12 (19)	*	11 (25)	12 (13)	15 (1)
Treatment Lag	19 (13)	*	18 (4)	17 (6)	*
Households/Zip	18 (13)	*	16 (9)	8 (19)	8 (5)
Policy Type	20 (10)	*	*	*	*
Providers per City	9 (21)	14 (3)	10 (26)	4 (31)	10 (2)

Note: * represents insignificance of variable in the model.

Table 5-2B

Clearly, in both instances of target variables the specific model and software implementation determines how to unwind the cross correlations to extract the most information for prediction purposes. For example, the distance between the claimant's zip code and the first outpatient provider (Distance) ranks low in importance (19/21) in the TREENET application for the IME decision target but it is quite important in the TREENET model for favorable IME outcome (3/21). Note, however, provider 2 bill is deemed highly important in all IME non-benchmark applications. One way to isolate the importance of each predicting variable is to tally a summary importance score across models. We will use a score of (21-rank)*(importance), with all insignificant variables assigned zero importance, summed over all relevant model combinations. For example, the variable provider 2 type would have a summary score relating to the IME target across the five tree models for a total importance score of 2,268. This scoring formula is typical of the ad hoc methods common to data mining analytics. The multiplicative form gives emphasis to both the categorical rank and the importance score in a dual monotone way. The numeric value of the score is less important than the final rankings of the variables. Tables 5-3A&B and 5-4A&B show the range of variable importance summary scores for all variables relative to the two targets, IME and SIU, respectively. The ranks of the variables according to the two summary scores are highly (Pearson) correlated as, for example, the decision summary ranks and favorable summary ranks have correlation coefficients of 0.65 for IME and 0.57 for SIU. The tables

also indicate the variable category of original DCD field (F), an internally derived variable (DV) and an external demographic variable (DM). The external demographic variables do not seem to be very informative in the presence of the field and derived variables chosen.

Important Variable Summarizations for IME Tree Models Applied to Decision and Favorable Targets					
			Total Score	Decision Score	Favorable Score
Variable	Variable type	Total Score	Rank	Rank	Rank
Health Insurance	F	17,206	1	2	1
Provider 2 Bill	F	10,820	2	1	4
Territory	F	7,871	3	5	2
Provider 1 Bill	F	6,726	4	4	3
Injury Type	F	6,084	5	6	5
Attorneys Per Zip	DV	3,102	6	3	15
Provider 2 Type	F	2,873	7	8	9
Report Lag	DV	2,859	8	16	7
Provider 1 Type	F	2,531	9	10	6
Distance MP1 Zip to Clt Zip	DV	1,655	10	11	8
Treatment Lag	DV	1,331	11	17	16
Emergency Treatment	F	1,216	12	7	10
Claimants per City	DV	1,146	13	14	13
Income Household/Zip	DM	987	14	13	17
Attorney	F	971	15	9	19
Households/Zip	DM	957	16	19	11
Age	F	881	17	12	14
Providers/Zip	DV	838	18	18	12
Providers per City	DV	719	19	20	18
Avg. Household Price/Zip	DM	262	20	15	20
Policy Type	F	4	21	21	21

Table 5-3

Important Variable Summarizations for SIU Tree Models Applied to Decision and Favorable Targets					
			Total Score	Decision Score	Favorable Score
Variable	**Variable Type**	**Total Score**	**Rank**	**Rank**	**Rank**
Territory	F	15,242	1	2	1
Provider 1 Type	F	9,965	2	4	2
Providers/Zip	DV	6,676	3	1	13
Provider 1 Bill	F	6,240	4	3	10
Provider 2 Bill	F	6,030	5	5	4
Injury Type	F	5,845	6	7	3
Provider 2 Type	F	4,753	7	8	6
Health Insurance	F	4,262	8	6	15
Emergency Treatment	F	3,039	9	13	5
Attorney	F	2,705	10	9	14
Report lag	DV	2,642	11	10	9
Providers per City	DV	2,275	12	12	10
Attorneys/zip	DV	2,183	13	14	8
Distance to Provider	DV	2,109	14	11	14
Income/Cap Household	DM	2,091	15	15	7
Claimants per City	DV	1,142	16	18	16
Households/Zip	DM	1,061	17	16	18
Age	F	830	18	19	17
Treatment Lag	DV	706	19	17	20
Average House Price	DM	648	20	20	9
Policy Type	F	19	21	21	21

Table 5-4

Additional Analyses

Most software allow for additional diagnostic tools that focus on the importance of individual variable levels in the predictive model. We focus on two such features: partial dependency plots and pruning of trees. Both features are designed to illustrate the contribution of each *level* of categorical variable and each *interval* of continuous variables created by the cut points. We illustrate the additional analyses using the Random Forest and S-PLUS's tree software.

Partial Dependence

The partial dependence plot is a useful way to visualize the effect of the values of a specific variable on a dependent variable when a complex modeling method such as Random Forest is used. The partial dependence plot is a graph of the marginal effect of a variable on the class probability. For a classification application (in Random Forest), the partial plot uses the logit or log of the odds ratio (the odds of being in the target category versus its compliment) rather than the actual probability.

$$f(x) = \log p_k(x) - \sum_{j=1}^{K} \log(p_j) \qquad (7)$$

Figures 5-1 and 5-2 show the partial dependence plot for the two IME targets for the most important variable in Table 5-4, territory.

Random Forest: IME Requested

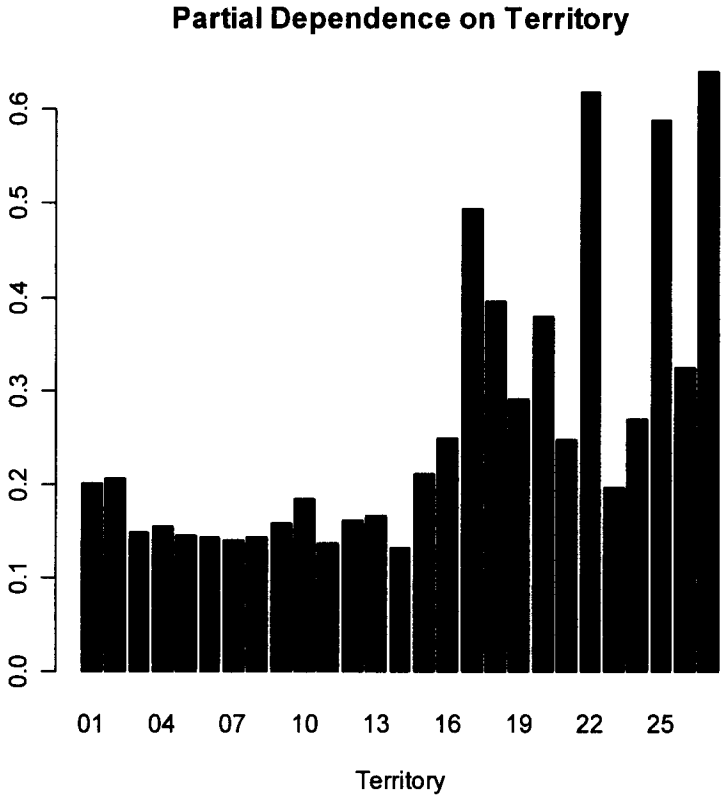

Figure 5-1

Random Forest: IME Favorable

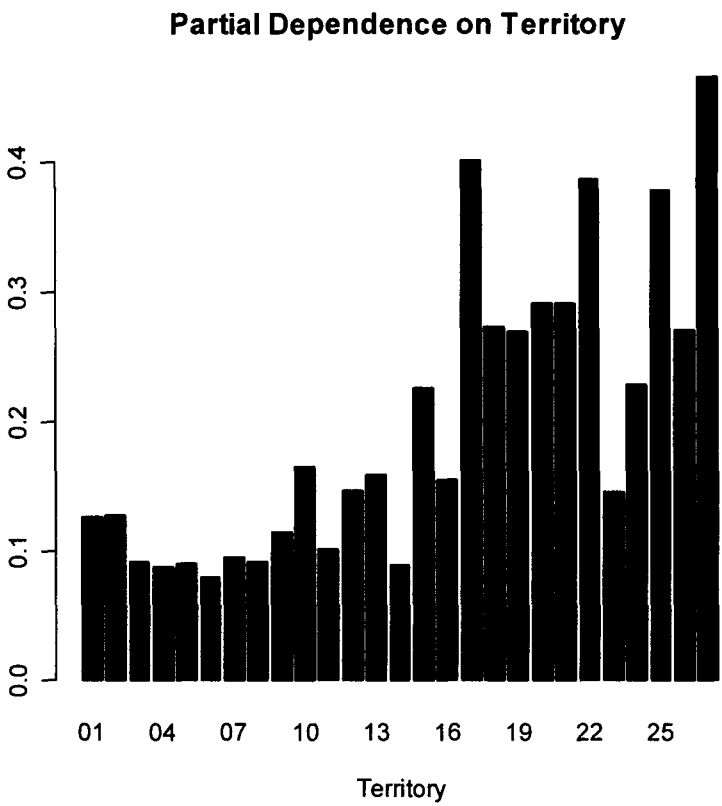

Figure 5-2

Both bar graphs have a distinctive right shift in the size of the partial dependency on the territory variable. This result is not surprising given that Massachusetts automobile territories are set every two years based upon the calculation of a single 5-coverage pure premium index for each of 350 towns. Towns are then grouped into 16 nearly homogenous territories with the index generally rising from territory 1 (lowest) to territory 16 (highest). Territories 17-26 are 10 individual parts of Boston that vary widely in this calculated pure premium index (Conger, 1987). Figure 5-3 shows a bar graph of the pure premium indices for the 26 territories used in this analysis for comparison purposes.

Massachusetts Rating Territories

Five Coverage Pure Premiums

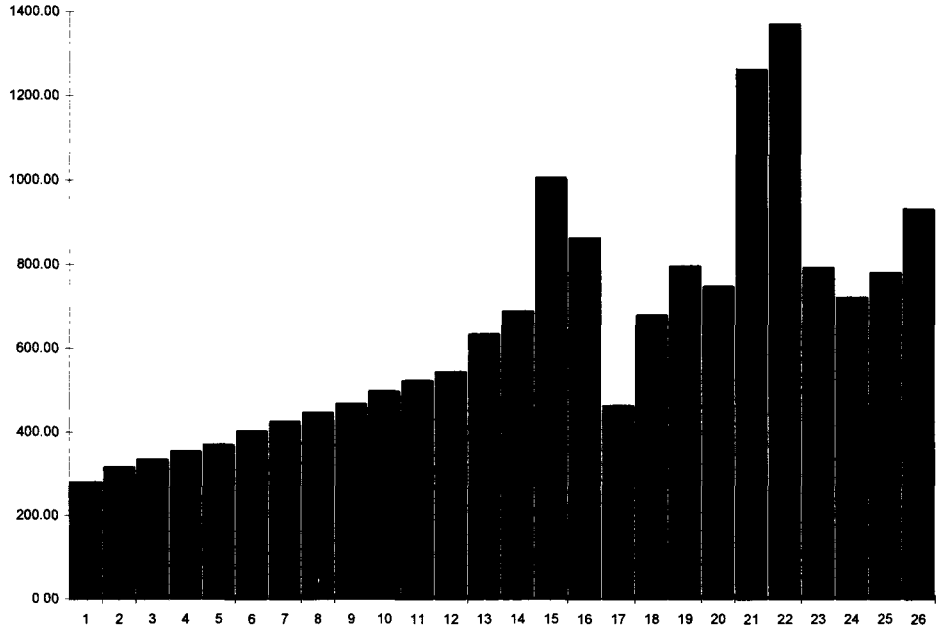

Figure 5-3

Figure 5-4 displays the proportion of claims with an IME requested (not marginal effects) by territory, superimposed on the pure premium territory levels. In contrast to the similarity of the marginal importance of the IME territory variable to the territory pure premiums, the proportions of claims with IME requested shown in Figure 5-4 show more uniformity across territories, indicating a real dependence on other important variables.

Massachusetts Rating Territories

Five Coverage Pure Premium vs IME Request Ratios

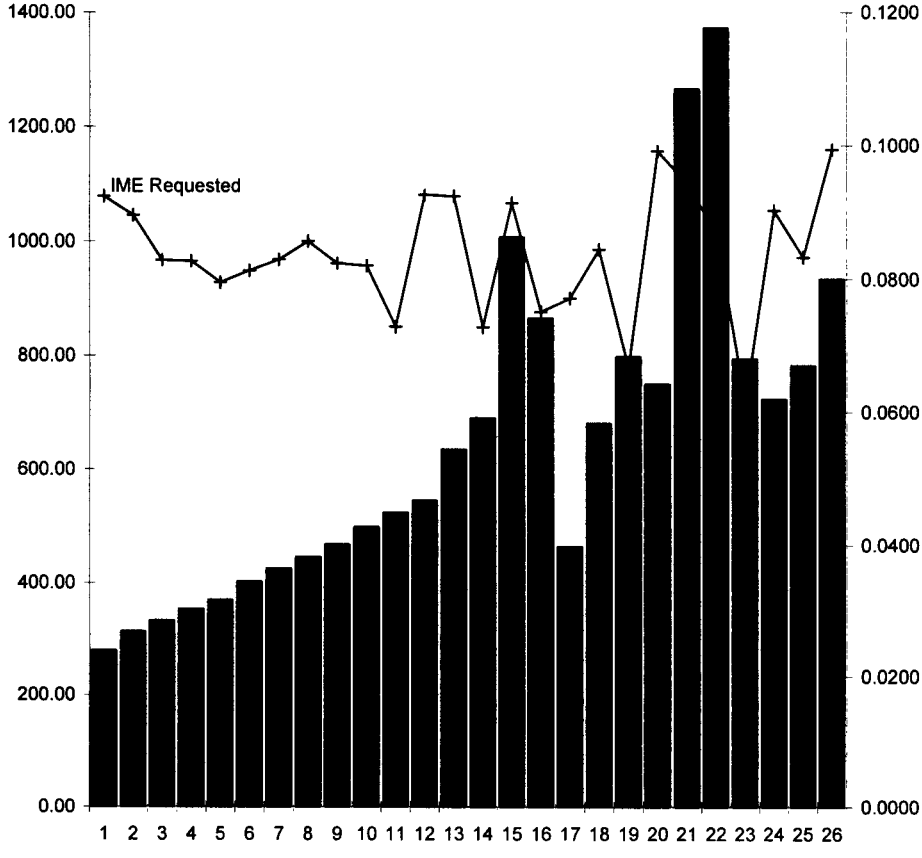

Figure 5-4

Pruning the Trees

Simple trees[31] that extend to a large number of terminal nodes are difficult to assess the full importance of individual variable levels because (1) later node splits may or may not be statistically significant depending on the software algorithms employed and (2) terminal nodes on the order of fifty plus may obscure the precise contribution of each variable level despite the importance value described above for the overall variable.

The full tree produced by the software can be *pruned* back to the "best" tree with a pre-determined number of nodes. For example, Figure 5-5 shows a best 10 node pruned tree from S-PLUS. It begins with the health insurance variable as the "root" node (Y/N to the left and U to the right)[32] and proceeds to make general node splits based only on the provider 2 bill amount. The universe of records is then classified by terminal node IME requested ratios ranging from 0.019 to 0.170. A similar pruned tree can be produced for the other three targets.

S-PLUS TREE: IME Requested
Best Ten Node Pruned Tree

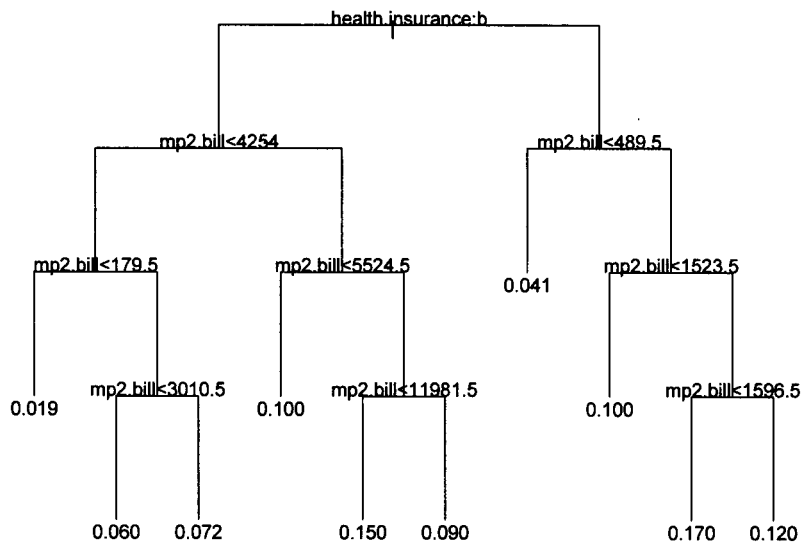

Figure 5-5

We next turn to consideration of model performance as a whole in section 6 with an interpretation of the models and variables relative to the problem at hand (example 2) in Section 7.

SECTION 6. ROC CURVES AND LIFT FOR SOFTWARE: TREES, NAIVE BAYES AND LOGISTIC MODELS

The sensitivity and specificity measures discussed in Section 4 are dependent on the choice of a cutoff value for the prediction. Many models score each record with a value between zero and one, though some other scoring scale can be used. This score is sometimes treated like a probability, although the concept is much closer in spirit to a fuzzy set measurement function[33]. A common cutoff point is .5 and records with scores greater than .5 are classified as events and records below that value are classified as non-events[34]. However, other cutoff

values can be used. Thus, if a cutoff lower than 50% were selected, more events would be accurately predicted and fewer non-events would be accurately predicted.

Because the accuracy of a prediction depends on the selected cutoff point, techniques for assessing the accuracy of models over a range of cutoff points have been developed. A common procedure for visualizing the accuracy of models used for classification is the receiver operating characteristic (ROC) curve[35]. This is a curve of sensitivity versus specificity (or more accurately 1.0 minus the specificity) over a range of cutoff points. It illustrates graphically the sensitivity or true positive rate compared to 1- specificity or false alarm rate. When the cutoff point is very high (i.e. 1.0) all claims are classified as legitimate. The specificity is 100% (1.0 minus the specificity is 0), but the sensitivity is 0%. As the cutoff point is lowered, the sensitivity increases, but so does 1.0 minus the specificity. Ultimately a point is reached where all claims are predicted to be events, and the specificity declines to zero (1.0 - specificity = 1.0). The baseline ROC curve (where no model is used) can be thought of as a straight line from the origin with a 45-degree angle. If the model's sensitivity increases faster than the specificity decreases, the curve "lifts" or rises above a 45-degree line quickly. The higher the "lift" or "gain"; the more accurate the model[36]. ROC curves have been used in prior studies of insurance claims and fraud detection regression models (Derrig and Weisberg, 1998 and Viaene et al., 2002). The use of ROC curves in building models as well as comparing performance of competing models is a well established procedure (Flach et al (2003)).

A statistic that provides a one-dimensional summary of the predictive accuracy of a model as measured by an ROC curve is the area under the ROC curve (AUROC). In general, AUROC values can distinguish good models from bad models but may not be able to distinguish among good models (Marzban, 2004). A curve that rises quickly has more area under the ROC curve. A model with an area of .50 demonstrates no predictive ability, while a model with an area of 1.0 is a perfect predictor (on the sample the test is performed on). For this analysis, SPSS was used to produce the ROC curves and area under the ROC curves. SPSS generates cutoff values midway between each unique score in the data and uses the trapezoidal rule to compute the AUROC. A non-parametric method was used to compute the standard error of the AUROC. The formula for the standard error[37] is:

$$SE(A) = \sqrt{\frac{A(1-A) + (n_+ - 1)(Q_1 - A^2) + (n_- - 1)(Q_2 - A^2)}{n_+ N}} \qquad (8)$$

Where n_+ is the number of events, n_- is the number of non-events, N is the sample size
A is the AUROC and scores are denoted as x

$$Q_1 = \frac{1}{n_- n_+^2} \sum_x n_- = j \times [n_{+>/}^2 + n_{+>/} + n_{+=/} + \frac{n_{+=/}^2}{3}]$$

$$Q_2 = \frac{1}{n_-^2 n_+} \sum_x n_+ = j \times [n_{->/}^2 + n_{->/} + n_{-=/} + \frac{n_{-=/}^2}{3}]$$

Tables 6-1A&B show the values of AUROC for each of eight model/software combinations in predicting a decision to investigate with an IME (6-1A) and an SIU (6-1B). for the Massachusetts auto bodily injury liability claims that comprise the holdout sample, about 50,000 claims. Upper and lower bounds for the "true" AUROC value are shown as the AUROC value ± two standard deviation determined by equation (7). TREENET, Random Forest both do well with AUROC values about 0.7, significantly better than the logistic model. The Iminer models (Tree, Ensemble and Naïve Bayes) generally have AUROC values significantly below the top two performers, with two (Tree and Ensemble) significantly below the Logistic and the Iminer Naïve Bayes benchmarks. CART also scores at or below the benchmarks and significantly below TREENET and Random Forest. On the other hand, S-PLUS (R) tree scores at or somewhat above the benchmarks.

Area Under the ROC Curve – IME Decision				
	CART Tree	**S-PLUS Tree**	**Iminer Tree**	**TREENET**
AUROC	0.669	0.688	0.629	0.701
Lower Bound	0.661	0.680	0.620	0.693
Upper Bound	0.678	0.696	0.637	0.708
	Iminer Ensemble	**Random Forest**	**Iminer Naïve Bayes**	**Logistic**
AUROC	0.649	703	0.676	0.677
Lower Bound	0.641	695	0.669	0.669
Upper Bound	0.657	711	0.684	0.685

Table 6-1A

Area Under the ROC Curve – IME Favorable				
	CART Tree	**S-PLUS Tree**	**Iminer Tree**	**TREENET**
AUROC	0.651	0.664	0.591	0.683
Lower Bound	0.641	0.653	0.578	0.673
Upper Bound	0.662	0.675	0.603	0.693
	Iminer Ensemble	**Random Forest**	**Iminer Naïve Bayes**	**Logistic**
AUROC	0.654	0.692	0.670	0.677
Lower Bound	0.643	0.681	0.660	0.667
Upper Bound	0.665	0.702	0.681	0.687

Table 6-1B

Tables 6-2A&B show the values of AUROC for the model/software combinations tested for the SIU dependent variable. We first note that, in general, the model predictions as measured by AUROC are significantly lower than for IME across all eight model/software combinations. This reduction in AUROC values may be a reflection of the explanatory variables used in the analysis; i.e., they may be more informative about claim build-up, for which IME is the principal investigative tool, than about claim fraud, for which SIU is the principal investigative tool.

Area Under the ROC Curve – SIU Decision				
	CART Tree	**S-PLUS Tree**	**Iminer Tree**	**TREENET**
AUROC	0.607	0.616	0.565	0.643
Lower Bound	0.598	0.607	0.555	0.634
Upper Bound	0.617	0.626	0.575	0.652
	Iminer Ensemble	**Random Forest**	**Iminer Naïve Bayes**	**Logistic**
AUROC	0.539	0.677	0.615	0.612
Lower Bound	0.530	0.668	0.605	0.603
Upper Bound	0.548	0.686	0.625	0.621

Table 6-2A

Area Under the ROC Curve – SIU Favorable				
	CART Tree	**S-PLUS Tree**	**Iminer Tree**	**TREENET**
AUROC	0.598	0.616	0.547	0.678
Lower Bound	0.584	0.607	0.555	0.667
Upper Bound	0.612	0.626	0.575	0.689
	Iminer Ensemble	**Random Forest**	**Iminer Naïve Bayes**	**Logistic**
AUROC	0.575	0.645	0.607	0.610
Lower Bound	0.530	0.631	0.593	0.596
Upper Bound	0.548	0.658	0.625	0.623

Table 6-2B

TREENET and Random Forest perform significantly better than all other model/software combinations on the favorable target variables. Both perform significantly better than the Logistic. Iminer Tree and Ensemble again do poorly on the IME and SIU Favorable holdout samples.

Figures 6-1 to 6-4 show the ROC curves for TREENET compared to the Logistic for both IME and SIU[38]. As we can see, a simple display of the ROC curves may not be sufficient to distinguish performance of the models as well as the AUROC values.

TREENET ROC Curve – IME
AUROC = 0.701

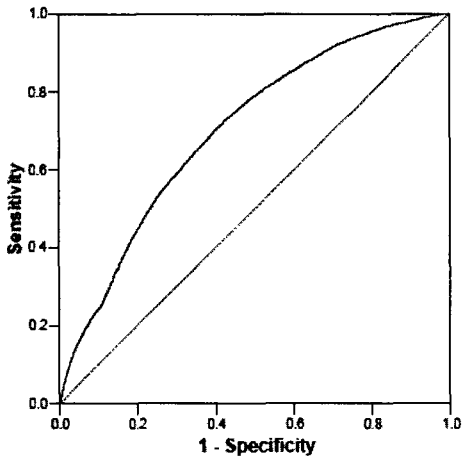

Figure 6-1

TREENET ROC Curve – SIU
AUROC = 0.677

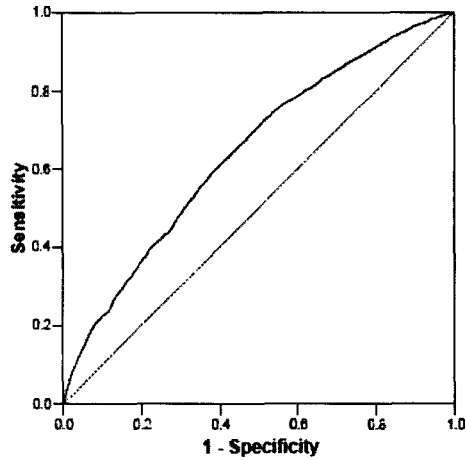

Figure 6-2

Logistic ROC Curve – IME
AUROC = 0.643

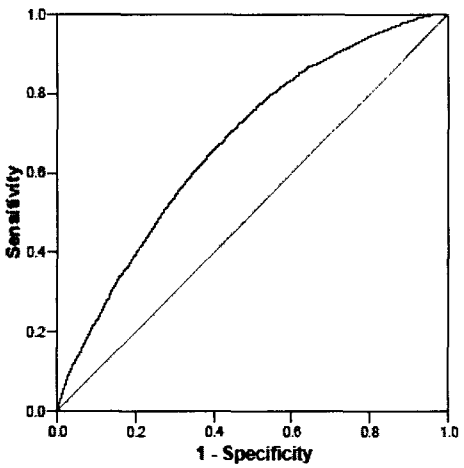

Figure 6-3

Logistic ROC Curve – SIU
AUROC = 0.612

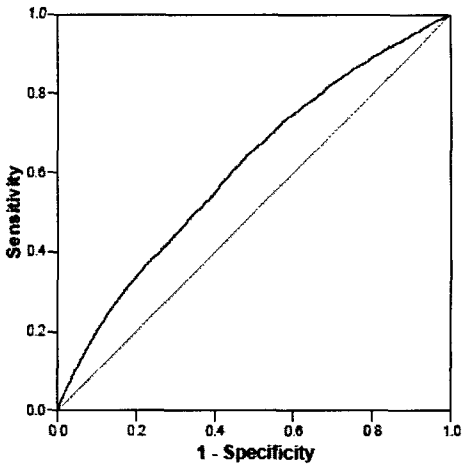

Figure 6-4

Finally, Table 6-3 displays the relative performance of the model/software combinations according to AUROC values and their ranks. With Naïve Bayes and Logistic as the benchmarks, TREENET, Random Forest and SPLUS Tree do better than the benchmarks while CART Tree, Iminer Tree, and Iminer Ensemble do worse.

Ranking of Methods By AUROC - Decision				
Method	**SIU AUROC**	**SIU Rank**	**IME Rank**	**IME AUROC**
Random Forest	0.645	1	1	0.703
TREENET	0.643	2	2	0.701
S-PLUS Tree	0.616	3	3	0.688
Iminer Naïve Bayes	0.615	4	5	0.676
Logistic	0.612	5	4	0.677
CART Tree	0.607	6	6	0.669
Iminer Tree	0.565	7	8	0.629
Iminer Ensemble	0.539	8	7	0.649

Table 6-3A

Ranking of Methods By AUROC - Favorable				
Method	**SIU AUROC**	**SIU Rank**	**IME Rank**	**IME AUROC**
TREENET	0.678	1	2	0.683
Random Forest	0.645	2	1	0.692
S-PLUS Tree	0.616	3	5	0.664
Logistic	0.610	4	3	0.677
Iminer Naïve Bayes	0.607	5	4	0.670
CART Tree	0.598	6	7	0.651
Iminer Ensemble	0.575	7	6	0.654
Iminer Tree	0.547	8	8	0.591

Table 6-3B

Finally, Figures 6-5A&B show the relative performance in a graphic. Procedures would work equally on both IME and SIU if they lie on the 45 degree line. To the extent that performance is better on the IME targets, procedures would be above the diagonal. Better performance is shown by positions farther to the right and closer to the top of the square. This graphic clearly shows that TREENET and Random Forest procedures do better than the other tree procedures and the benchmarks.

Plot of AUROC for SIU vs. IME Decision

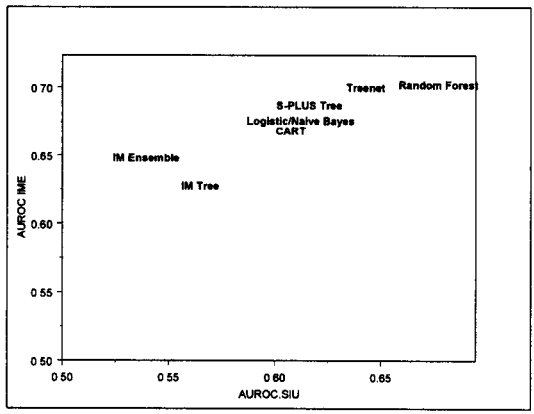

Figure 6-5A

Plot of AUROC for SIU vs. IME Favorable

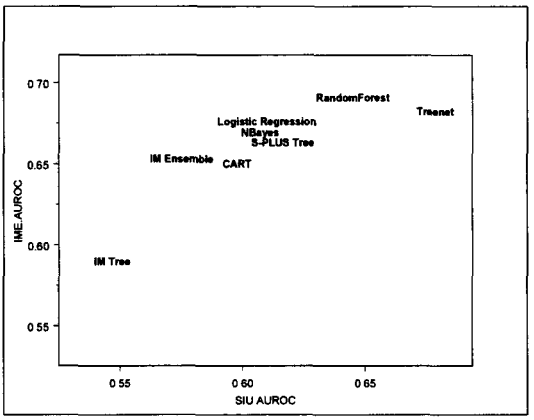

Figure 6-5B

SECTION 7. CONCLUSION

Insurance data often involves both large volumes of information and nonlinearity of variable relationships. A range of data manipulation techniques have been developed by computer scientists and statisticians that are now categorized as data mining, techniques with principal advantages being precisely the efficient handling of large data sets and the fitting of non-linear functions to that data. In this paper we illustrate the use of software implementations of CART and other tree-based methods, together with benchmark procedures of Naïve Bayes and Logistic regression. Those eight model/software combinations are applied to data arising in the Detail Claim Database (DCD) of auto injury liability claims in Massachusetts. Twenty-one variables were selected to use in prediction models using the DCD and external demographic variables. Four target categorical variables were selected to model: The decision to request an independent medical examination (IME) or a special investigation (SIU) and the favorable outcome of each investigation. The two decision targets are the prime claim handling techniques that insurers can use to reduce the asymmetry of information between the claimant and the insurer in order to distinguish valid claims from those involving buildup, exaggerated injuries or treatment, or outright fraud.

Eight modeling software results were compared for effectiveness of modeling the targets based on a standard procedure, the area under the receiver operating characteristic curve (AUROC). We find that the methods all provide some predictive value or lift from the predicting variables we make available, with significant differences among the eight methods and four targets. Seven modeling outcomes are compared to logistic regression as in Viaene et al. (2002) but the results here are different. They show some software/methods can improve on the predictive ability of the logistic model. TREENET, Random Forest and SPLUS Tree do better than the benchmark Naïve Bayes and Logistic methods, while CART

tree, Iminer tree, and Iminer Ensemble do worse. That some model/software combinations do better than the logistic model may be due to the relative size and richness of this data set and/or the types of independent variables at hand compared to the Viaene et al. data.

We show how "important" each variable is within each software/model tested and note the type of data that are important for this analysis. In general, variables taken directly from DCD fields and variables derived as demographic type variables based on DCD fields do better than variables derived from external demographic data. Variables relating to the injury and medical treatment dominate the highly important variables while the presence of an attorney, age of the claimant, and policy type, personal or commercial, are less important in making the decision to invoke these two investigative techniques.

No general conclusions about auto injury claims can be drawn from the exercise presented here except that these modeling techniques should have a place in the actuary's repertory of data manipulation techniques. Technological advancements in database assembly and management, especially the availability of text mining for the production of variables, together with the easy access to computer power, will make the use of these techniques mandatory for analyzing the nonlinearity of insurance data. As for our part in advancing the use of data mining in actuarial work, we will continue to test various software products that implement these and other data mining techniques (e.g. support vector machines).

REFERENCES

Allison, P., Missing Data, Sage Publications, 2002.

Automobile Insurers Bureau of Massachusetts, Detail Claim Database Claim Distribution Characteristics, Accident Years 1995-1997, Boston MA, 2004.

Brieman, L., J. Freidman, R. Olshen, and C. Stone, Classification and Regression Trees, Chapman Hall, 1993.

Conger, R.F., The Construction of Automobile Rating Territories in Massachusetts, *Casualty Actuarial Society Forum*, pp. 230-335, Fall 1987.

Derrig, R.A., and L. Francis, Comparison of Methods and Software for Modeling Nonlinear Dependencies: A Fraud Application, Working Paper, 2005.

Flach, P., H. Blockeel, C. Ferri, J. Hernandez-Orallo, and J. Struyf, Decision Support for Data Mining: An Introduction to ROC Analysis and Its Applications, Chapter 7 in Data

Mining and Decision Support, D. Mlandenic, Nada Lavrac, Marko Bohanec and Steve Moyle Eds., Kluwer Academic, N.Y., 2003.

Fox, J, An R and S-PLUS Companion to Applied Regression, SAGE Publications, 2002.

Francis, L.A., An introduction to Neural Networks in Insurance, Intelligent and Other Computational Techniques in Insurance, Chapter 2, A.F. Shapiro and L.C. Jain, Eds, World Scientific, pp. 51-1, 2003a.

Francis, L.A., Martian Chronicles: Is MARS better than Neural Networks? *Casualty Actuarial Society Forum,* Winter, pp. 253-320, 2003b.

Francis, L.A., Practical Applications of Neural Networks in Property and Casualty Insurance, Intelligent and Other Computational Techniques in Insurance, Chapter 3, A.F. Shapiro and L.C. Jain, Editors, World Scientific, pp. 104-136, 2003c.

Francis, L.A., Neural Networks Demystified, *Casualty Actuarial Society Forum,* Winter, pp. 254-319, 2001.

Francis, L.A., A Comparison of TREENET and Neural Networks in Insurance Fraud Prediction, presentation at CART Data Mining Conference, 2005.

Friedman, J., Greedy Function Approximation: The Gradient Boosting Machine, *Annals of Statistics,* 2001.

Hassoun, M.H., Fundamentals of Artificial Neural Networks, MIT Press, Cambridge MA, 1995.

Hastie, T., R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer, New York., 2001.

Hosmer, David W. and Stanley Lemshow, Applied Logistic Regression, John Wiley and Sons, 1989.

Insurance Research Council, Auto Injury Insurance Claims: Countrywide Patterns in Treatment Cost, and Compensation, Malvern, PA, 2004a.

Insurance Research Council, Fraud and Buildup in Auto Injury Insurance Claims, Malvern, PA, 2004b.

Jacard, J and Turrisi, Interaction Effects in Multiple Regression, SAGE Publications, 2003.

Kantardzic, M., Data Mining, John Wiley and Sons, 2003.

Lawrence, Jeannette, Introduction to Neural Networks: Design, Theory and Applications, *California Scientific Software,* 1994.

Marzban, C., A Comment on the ROC curve and the Area Under it as Performance Measures, 2004. *Weather and Forecasting,* Vol. 19, No. 6, 1106-1114.

McCullagh, P. and J. Nelder, General and Linear Models, Chapman and Hall, London, 1989.

Miller, R.B. and D.B. Wichern, Intermediate Business Statistics, Holden-Day, San Francisco, 1997.

Neter, John, Mihael H. Kutner, William Waserman, and Christopher J. Nachtsheim, Applied Linear Regression Models, 4th Edition, 1985.

Salford Systems, Data mining with Decision Trees: Advanced CART Techniques, Notes from Course, 1999.

Smith, Murry, (1996), Neural Networks for Statistical Modeling, International Thompson Computer Press, 1996.

Viaene, S., R.A. Derrig, and G. Dedene, A Case Study of Applying Boosting Naïve Bayes for Claim Fraud Diagnosis, IEEE Transactions on Knowledge and Data Engineering, 2004, Volume 16, (5), pp. 612-620, May.

Viaene, S., R.A. Derrig, and G. Dedene, Illustrating the Explicative Capabilities of Bayesian Learning Neural Networks for Auto Claim Fraud Detection, Intelligent and Other Computational Techniques in the Insurance, Chapter 10, A.F. Shapiro and L.C. Jain, Editors, World Scientific, pp. 365-399, 2003.

Viaene, S., B. Baesens, G. Dedene, and R. A. Derrig, A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Fraud Detection, *Journal of Risk and Insurance*, 2002, Volume 69, (3), pp. 373-421.

Weisberg, H.I. and R.A. Derrig, Methodes Quantitatives Pour La Detection Des Demandes D'Indemnisation Frauduleuses, *RISQUES*, No. 35, Juillet-Septembre, 1998, Paris.

Zhou, X-H, D.K. McClish, and N.A. Obuchowski, Statistical Methods in Diagnostic Medicine, John Wiley and Sons, New York, 2002.

[1] A good up-to-date and comprehensive source for a variety of data manipulation procedures is Hastie, Tibshirani, and Friedman (2001), Elements of Statistical Learning, Springer.

[2] They also found that augmenting the categorized red flag variables with some other claim data (e.g. age, report lag) improved the lift as measured by AUROC across all methods but the logistic model still did as well as the other methods (Viaene et al., 2002, Table 6, p.400-401).
[3] A wider set of data mining techniques is considered in Derrig, R.A. and L A. Francis, Comparison of Methods and Software Modeling Nonlinear Dependencies: A Fraud Application, Congress of Actuaries, Paris, June 2006
[4] See section 2 for an overview of the database and descriptions of the variables used for this paper.
[5] The relative importance of the independent variables in modeling the dependent variable within these methods are analogous to statistical significance or p-values in ordinary regression models.
[6] See, for example, 2004 Discussion Paper Program, Applying and Evaluating Generalized Linear Models, May 16-19, 2004, Casualty Actuarial Society.
[7] This was the text used by the Casualty Actuarial Society for the exam on applied statistics during the 1980s
[8] Claims that involve only third party subrogation of personal injury protection (no fault) claims but no separate indemnity payment or no separate claims handling on claims without payment are not reported to DCD.
[9] Combined payments under PIP and Medical Payments are reported to DCD.
[10] With a large holdout sample, we are able to estimate tight confidence intervals for testing model results in section 6 using the area under the ROC curve measure.
[11] This fact is a matter of Massachusetts law which does not permit IMEs by one type of physician, say an orthopedist, when another physician type is treating, say a chiropractor. This situation may differ in other jurisdictions.
[12] Because expert bill review systems became pervasive by 2003, reaching 100% in some cases, DCD redefined the reported MA to encompass only peer reviews by physicians or nurses for claims reported after July 1, 2003..
[13] The standard Massachusetts auto policy has a cooperation clause for IME both in the first party PIP coverage and in the third party BI liability coverage.
[14] The IRC also includes an index bureau check as one of the claims handling activities
[15] Prior studies of Massachusetts Auto Injury claim data for fraud content included Weisberg and Derrig (1998, Suspicion Regression Models) and Derrig and Weisberg (1998, Claim Screening with Scoring Models).
[16] See Section 5 for the importance of the provider 2 bill variable in the decision to investigate claims for fraud (SIU) and/or buildup (IME).
[17] There are Tree Software models that may split nodes into three or more branches. SPSS classification trees is an example of such software.
[18] For binary categorical data assumed to be generated from a binomial distribution, entropy and deviance are essentially the same measure. Deviance is a generalized linear model concept and is closely related to the log of the likelihood function.
[19] Hastie et al., p. 301 Note that Hastie et al. describe other error and weight functions. [endnote]
[20] Note that the ensemble tree methods employ all 21 variables in the models. See tables 5-1 and 5-2.
[21] The ROC curve results in Section 6 show that TREENET generally provides the best prediction models for the Massachusetts data.
[22] The numeric variables were grouped into five bins or into quintiles in this instance.
[23] The software product MARS also was used to rank variables in importance. MARS implements multivariate adaptive regression splines and is described in Francis (2003).
[24] The SAS code is generally relatively easy to edit if some other language is used to implement the model
[25] See Section 5 for the importance of variables in our study.
[26] S-PLUS would convert the numeric variable into a categorical variable with a level for every numeric value that is in the training data, including missing data, but the result would have far too many categories to be feasible.
[27] Generally by collapsing sparsely populated categories into an "all other" category

[28] It also contains some dimension reduction methods such as clustering and Principal Components which are also contained in S-PLUS.
[29] In general, some programming is required to apply either approach in S-PLUS (R)
[30] The data set is described in more detail in Section 2 above.
[31] Pruning is not feasible or necessary for the example tree methods such as TREENET or Random TREENET.
[32] The S-PLUS tree graph does not print out the values of categorical variables, although it displays the values of the numeric variables. For categorical variables letters are assigned and displayed instead of the category values.
[33] See Ostaszewski (1993) or Derrig and Ostaszewski (1999).
[34] One way of dealing with values equal to the cutoff point is to consider such observations as one-half in the event group and one-half in the non-event group
[35] A ROC curve is one example of a so-called "gains" chart.
[36] ROC curves were developed extensively for use in medical diagnosis testing in the 1970s and 1980s (Zhou et al. 2004 and more recently in weather forecasting (Marzban, 2004) and (Stephenson, 2000).
[37] The details of the formula were supplied by SPSS.
[38] All twenty ROC curves are available from the authors.

Acknowledgement: The authors gratefully acknowledge the production assistance of Eilish Browne and helpful comments from Rudy Palenik on a prior version of the paper.