Comparison of Methods and Software for Modeling Nonlinear Dependencies: A Fraud Application

By

Richard Derrig, PhD,	**Louise Francis, FCAS, MAAA**
OPAL Consulting LLC	**Francis Analytics and Data Mining**
41 Fosdyke Street	**706 Lombard Street**
Providence	**Philadelphia**
Rhode Island, 02906, U.S.A.	**Pennsylvania, 19147, U.S.A.**
Phone: 001-401-861-2855	**Phone: 001-215-923-1567**
email: richard@derrig.com	**email: louise_francis@msn.com**

Abstract:

In recent years a number of approaches for modeling data containing nonlinear and other complex dependencies have appeared in the literature. These procedures include classification and regression trees, neural networks, regression splines and naïve Bayes. Viaene et al (2002) compared several of these procedures, as well as a classical linear model, 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 common data mining 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 in a large claim database. Among the software programs we will investigate are MARS, CART, S-PLUS, TREENET and Insightful Miner. 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 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 two 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

International Congress of Actuaries – Paris – May 28–June 2, 2006

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 four procedures of interest for the purposes of this paper are classification and regression trees, neural networks, regression splines and naïve Bayes. Viaene et al (2002) applied a similar set of procedures, as well as a classical general linear model, logistic regression, on a small single data set of 1400 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. [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 some publicly available common data mining approaches 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 MARS, CART, S-PLUS, TREENET and Insightful Miner. 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[3]. 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 is the modeling target.

Nine 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 nine methods and two targets. All nine modeling outcomes are 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[4] and note the type of data that are important for this analysis. This entire exercise should provide practicing actuaries with guidance on 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 and nonadditivity 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. The specific software procedures are covered in Section 3. Comparative outcomes for the variables ("importance") and software ("AUROC") are reported in Sections 4 and 5. Conclusions are shown in Section 6.

SECTION 1. NONLINEARITY AND NONADDITIVITY IN INSURANCE DATA

Nonlinearity

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[5].

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.

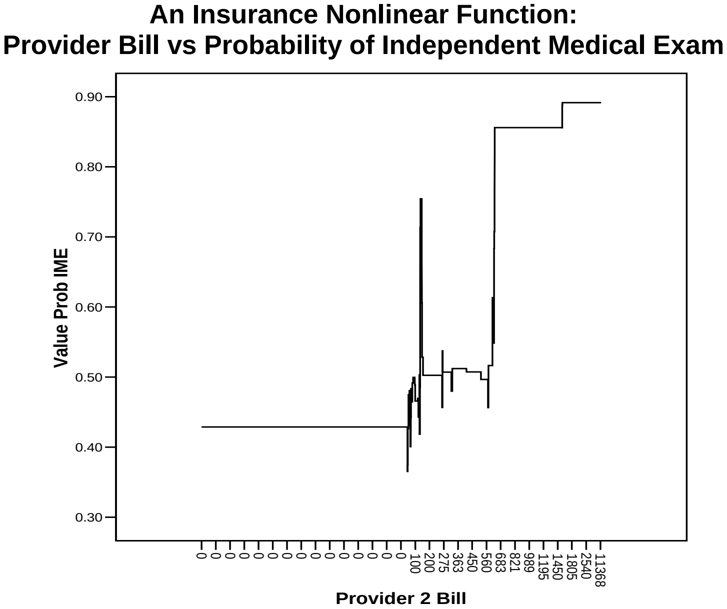

Figure 1 -1

Nearly all regression and econometrics 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 scatterplots 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[6], 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

The log link: h(Y) = *ln*(Y)

The inverse link: h(Y) = $\frac{1}{Y}$

The logit link: h(Y) = $\ln(\frac{Y}{1-Y})$

The probit link: h(Y) = $\Phi(Y), \Phi$ 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.

Five data mining methods were identified that meet the requirements: decision trees, ensemble trees neural networks, naieve bayes and regression splines. A detailed exposition and illustration of how each of these methods models nonlinearity is supplied in Derrig and Francis (2005). Technical definitions of the major methods can be found in Viaene et al.(2002).

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:

$$Y = a + b_1X_1 + b_2X_2 + b_3{*}X_1{*}X_2 \qquad (1)$$

In the case of a two-way interaction, the interaction terms appear as products of the two variables.

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 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 data mining techniques used in this paper have efficient methods for handling interactions. See Brieman et al. (1993) and Francis (2001, 2003c) for a more detailed description of how the methods model interactions.

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[7] arising from injury coverages: Personal Injury Protection (PIP)/ Medical payments excess of PIP[8], 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[9]. 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 their presence, outcome, and formulaic savings amounts.

The claims handling technique 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[10]. 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[11]. 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.

For purposes of this analysis and demonstration of non-linear models and software, we employ twenty-one potentially predicting variables and two 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 (DM) data 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

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

Table 2-1

Eight predicting variables, and both target variables (IME and SIU), 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 (1), no attorney
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%)
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 10% 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)[12]. Nationally, about 4% and 2% of BIL 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[13]. 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. We now turn to descriptions of the types of models, and the software that implements them, in the next section before we describe how they are applied to model the IME and SIU target variables.

SECTION 3. SOFTWARE FOR MODELING NON-LINEAR DEPENDENCIES

Software Products

Five software products were included in our fraud comparison: They are CART, MARS, Treenet, S-PLUS (R) and Insightful Miner.

CART, MARS and TREENET are Salford Systems stand-alone software products that each performs one technique. CART (Classification and Regression Trees) does tree analysis, MARS implements multivariate adaptive regression splines, and TREENET applies stochastic gradient boosting using the method described by Freidman (2001). All the products contain a procedure for handling missing values using surrogate variables which are described in Derrig and Francis (2005). Different versions of CART, MARS and TREENET handle different size databases. The number of levels on 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 (especially with MARS) 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[14]. Different statistics are used for CART and Treenet (based on Brieman et, al.(1993)) versus MARS[15].

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 by 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 a free statistical software language 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. These were: the tree function for decision trees, the nnet function (supplied by Venables and Ripley, 1999) for neural networks 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 creating a missing values category is not feasible[16]) to make a provision for the 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, although the neural network function was somewhat slow. 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[17].

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.

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 test or 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 data. Some (CART, Insightful Miner Tree) only allow cross-validation. Others (TREENET) use a validation sample. S-PLUS (R) and MARS allow either approach[18] 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 or cross-validation were 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 for the final testing and comparison of the models. Two key statistics often used to compare models accuracy are sensitivity and specificity. The sensitivity is the percentage of events (i.e., claims 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 3-1, 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 3-1

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 4. 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 ten model/software combinations applied to our Example 2, the decision to investigate via IMEs or referral to SIU. We model the presence of each investigative technique for the DCD subset of automobile bodily injury liability (third party) claims from 1995-1997 accident years.[19] 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 ten model/software combinations we will use here are also of three categories: five tree models, three alternative models (two neural and one regression spline), and two benchmark models (Naïve Bayes and logistic). They are:

1) TREENET
2) Iminer Tree
3) SPLUS Tree
4) CART
5) Iminer Ensemble
6) Naïve Bayes
7) MARS
8) Iminer Neural
9) SPlus Neural
10) Logistic

In this analysis, approximately a third of the data, about 50,000 records, were used for the final testing and comparison of the models. Two key statistics often used to compare models accuracy are sensitivity and specificity. The sensitivity is the percentage of events (i.e., claims 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. Model performance is covered in the next section, section 5, 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).

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 derived one. Different procedures are used for different methods and different products.

Three software products, CART, MARS 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. (Hastie, et al., p271-272.)

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.

MARS uses the generalized cross-validation statistic to rank variables (see Francis 2003b for a description of generalized cross-validation)

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 and neural network functions. For the logistic regression, information about the variables contribution to sum of squared variation explained by the model was used to rank it.

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. Rather, the predicted values, along with the independent variables were output to a file. These were used by another procedure, MARS[20] to determine an importance ranking. To impute rankings, the Iminer's predicted value was used as a dependent variable in MARS and the other variables were used to model it. The resulting importance rankings were used in Tables 4-1 and 4-2.

Each model/software combination output 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 4-1A displays the importance results for predicting an IME using the five tree models while Table 4-1B 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 ten variables (Iminer 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**	**Iminer Tree**	**S Plus Tree**	**CART**	**Iminer Ensemble**
Provider 2 Bill	1 (100)	*	2 (91)	1 (100)	1 (100)
Attorneys Per Zip	2 (80)	*	5 (26)	13 (9)	*
Territory	3 (71)	6 (29)	4 (32)	11 (11)	4 (61)
Health Insurance	4 (61)	1 (100)	1 (100)	3 (68)	3 (79)
Injury Type	5 (50)	2 (85)	6 (24)	5 (47)	2 (81)
Provider 1 Bill	6 (47)	*	3 (51)	4 (58)	*
Provider 1 Type	7 (31)	7 (26)	9 (7)	*	10 (16)
Report Lag	8 (31)	*	7 (16)	8 (18)	11 (12)
Attorney	9 (25)	4 (46)	12 (3)	*	*
Age	10 (23)	*	*	17 (2)	*
Provider 2 Type	11 (19)	3 (83)	8 (9)	*	8 (24)
Income Household/Zip	12 (18)	*	*	10 (13)	*
Avg. Household Price/Zip	13 (17)	11 (13)	*	15 (5)	*
Providers per City	14 (17)	*	*	9 (15)	6 (30)
Claimants per City	15 (16)	5 (30)	*	*	9 (22)
Providers/Zip	16 (16)	*	*	*	*
Households/Zip	17 (16)	*	11 (3)	*	*
Treatment Lag	18 (14)	8 (21)	10 (4)	18 (2)	7 (26)
Distance MP1 Zip to Clt Zip	19 (13)	9 (10)	*	20 (0.1)	*
Emergency Treatment	20 (4)	10 (16)	*	7 (20)	*
Policy Type	21 (3)	*	*	19 (2)	*

Note: * represents insignificance of variable in the model.

Table 4-1A

The same set of model/software combinations was used with the same set of twenty-one predicting variables to predict the use of special investigation or SIU. Tables 4-2A and 4-2B show the corresponding ranking of variables by importance for each of the ten model combinations. The number of significant variables ranges from eight for Iminer Tree to all twenty one for TREENET.

Software Ranking of Variables for IME Decision By Importance Rank and Value					
	(6)	**(7)**	**(8)**	**(9)**	**(10)**
Variable	**Naïve Bayes**	**Mars**	**Iminer Neural**	**S Plus Neural**	**Logistic**
Provider 2 Bill	1 (100)	2 (72)	*	2 (74)	10 (1)
Attorneys Per Zip	*	14 (15)	*	5 (38)	11 (1)
Territory	11 (10)	9 (34)	6 (29)	4 (49)	*
Health Insurance	3 (47)	1 (100)	1 (100)	1 (100)	1 (100)
Injury Type	2 (51)	3 (69)	2 (85)	6 (30)	2 (51)
Provider 1 Bill	8 (32)	5 (48)	*	3 (51)	*
Provider 1 Type	*	15 (12)	7 (26)	9 (15)	6 (8)
Report Lag	* 9 (28)	4 (59)	*	7 (23)	13 (1)
Attorney	*	10 (19)	*	12 (4)	5 (18)
Age	*	*	4 (46)	16 (1)	*
Provider 2 Type	10 (18)	*	3 (83)	8 (20)	3 (47)
Income Household/Zip	*	*	*	*	9 (2)
Avg. Household Price/Zip	*	8 (37)	11 (13)	*	*
Providers per City	4 (44)	15 (12)	*	17 (1)	*
Claimant per City	6 (32)	7 (40)	5 (30)	14 (2)	12 (1)
Providers/Zip	*	11 (17)	*	15 (2)	8 (2)
Household/Zip	*	*	*	11 (12)	7 (2)
Treatment Lag	7 (32)	6 (41)	8 (21)	10 (15)	4 (24)
Distance MP1 Zip to Clt Zip	5 (40)	12 (17)	9 (16)	18 (1)	*
Emergency Treatment	*	*	10 (16)	13 (3)	*
Policy Type	*	*			*

Note: * represents insignificance of variable in the model.

Table 4-1B

Software Ranking of Variables for SIU Decision By Importance Rank and Value					
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
Variable	**TREENET**	**Iminer Tree**	**S Plus Tree**	**CART**	**Iminer Ensemble**
Providers/Zip	1 (100)	*	1 (100)	8 (37)	3 (67)
Provider 2 Type	2 (98)	*	10 (3)	15 (34)	11 (35)
Territory	3 (92)	4 (42)	5 (18)	3 (84)	6 (47)
Health Insurance	4 (64)	*	3 (33)	7 (52)	13 (29)
Provider 1 Bill	5 (59)	2 (50)	2 (51)	2 (85)	4 (56)
Injury Type	6 (52)	8 (13)	7 (6)	5 (59)	7 (46)
Attorney	7 (47)	*	8 (4.5)	17 (13)	18 (23)
Provider 1 Type	8 (38)	*	4 (29)	4 (69)	9 (42)
Age	9 (31)	*	*	*	20 (19)
Provider 2 Bill	10 (30)	3 (42)	*	1 (100)	5 (49)
Report lag	11 (28)	*	*	6 (54)	19 (20)
Average House Price	12 (28)	*	*	15 (18)	10 (35)
Attorneys/zip	13 (22)	7 (18)	6 (8)	14 (20)	8 (44)
Distance to Provider	14 (20)	6 (19)	*	19 (4)	14 (27)
Emergency Treatment	15 (19)	*	*	13 (27)	16 (26)
Income/Cap Household	16 (18)	*	11 (3)	9 (4.5)	17 (25)
Claimants per City	17 (17)	5 (27)	*	12 (30)	1 (100)
Treatment Lag	18 (16)	*	9 (34)	18 (12)	12 (30)
Households/Zip	19 (16)	*	*	16 (16)	15 (27)
Policy Type	20 (8)	*	*	*	21 (14)
Providers per City	21 (6)	1 (100)	12 (1)	11 (30)	2 (71)

Note: * represents insignificance of variable in the model.

Table 4-2A

Software Ranking of Variables for SIU Decision
By Importance Rank and Value

	(6)	**(7)**	**(8)**	**(9)**	**(10)**
Variable	**Naïve Bayes**	**Mars**	**Iminer Neural**	**S Plus Neural**	**Logistic**
Providers/Zip	*	*	*	1(100)	*
Provider 2 Type	*	2 (60)	6 (37)	10 (5)	6 (39)
Territory	11 (22)	*	*	3 (21)	*
Health Insurance	10 (23)	3 (52)	9 (15)	9 (5)	7(28)
Provider 1 Bill	7 (33)	4 (35)	*	2 (79)	14 (2)
Injury Type	4 (47)	7 (24)	2 (43)	11 (3)	2 (71)
Attorney	5 (40)	10 (22)	4 (39)	12 (3)	3 (63)
Provider 1 Type	*	5 (30)	5 (38)	5 (18)	1 (100)
Age	*	9 (22)	*	*	*
Provider 2 Bill	1 (100)	*	*	4 (19)	13 (5)
Report lag	6 (40)	*	*	*	11 (17)
Average House Price	*	*	*	*	*
Attorneys/zip	*	8 (23)	*	*	12 (7)
Distance to Provider	*	11 (20)	10 (8)	*	4 (58)
Emergency Treatment	9 (24)	*	8 (19)	6(8)	5 (49)
Income/Cap Household	*	*	*	*	9 (27)
Claimants per City	3 (83)	*	3 (40)	*	*
Treatment Lag	8 (27)	6 (27)	*	8 (6)	15 (2)
Households/Zip	*	*	*	13 (1)	8 (28)
Policy Type	*	*	*	*	*
Providers per City	2 (93)	1 (100)	1 (100)	7 (6)	10 (22)

Note: * represents insignificance of variable in the model.

Table 4-2B

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, provider 2 bill ranks as the most important variable in the TREENET application for the IME target but it is insignificant in both the Iminer Tree and Neural models. Note, however, it is deemed highly important in all other 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 of 2,113, across all other models of 2,798, for a total importance score of 4,911. 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 4-3 and 4-4 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 three summary scores are highly (Pearson) correlated as for example, the tree model summary ranks and the other model summary ranks have correlation coefficients of 0.74 for IME and 0.65 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, Other Models and Total

			Total Score	**Tree Score**	**Other Score**
Variable	**Variable type**	**Total Score**	**Rank**	**Rank**	**Rank**
Health Insurance	F	16529	1	2	1
Provider 2 Bill	F	12514	2	1	3
Injury Type	F	10311	3	3	2
Territory	F	5180	4	4	7
Provider 2 Type	F	4911	5	6	4
Provider 1 Bill	F	4711	6	5	5
Attorneys Per Zip	DV	2731	7	7	14
Report Lag	DV	2650	8	10	8
Treatment Lag	DV	2638	9	13	6
Claimant per City	DV	2383	10	12	9
Provider 1 Type	F	1794	11	9	13
Providers per City	DV	1708	12	11	11
Attorney	F	1642	13	8	16
Distance MP1 Zip to Clt Zip	DV	1134	14	18	10
AGE	F	1048	15	17	12
Avg. Household Price/Zip	DM	907	16	16	15
Emergency Treatment	F	660	17	14	18
Income Household/Zip	DM	329	18	15	20
Providers/Zip	DV	288	19	20	17
Household/Zip	DM	242	20	19	19
Policy Type	F	4	21	21	21

Table 4-3

Important Variable Summarizations for SIU
Tree Models, Other Models and Total

			Total Score	**Tree Score**	**Other Score**
Variable	**Variable Type**	**Total Score**	**Rank**	**Rank**	**Rank**
Providers per City	DV	9751	1	5	1
Provider 1 Bill	F	8002	2	2	5
Providers/Zip	DV	7687	3	1	9
Provider 2 Bill	F	6233	4	4	6
Provider 1 Type	F	6040	5	7	2
Injury Type	F	5952	6	9	3
Territory	F	5473	7	3	15
Claimants per City	DV	4984	8	6	8
Provider 2 Type	F	4784	9	10	7
Health Insurance	F	4463	10	8	10
Attorney	F	3543	11	13	4
Report lag	DV	1900	12	12	14
Emergency Treatment	F	1899	13	17	11
Distance to Provider	DV	1896	14	16	12
Attorneys/zip	DV	1622	15	11	17
Treatment Lag	DV	1608	16	14	13
Average House Price	DM	745	17	15	21
Age	F	655	18	18	19
Households/Zip	DM	646	19	19	16
Income/Cap Household	DM	598	20	20	18
Policy Type	F	8	21	21	21

Table 4-4

We next turn to consideration of model performance as a whole in Section 5.

SECTION 5. ROC CURVES AND LIFT FOR SOFTWARE: TREES, NEURAL NETWORKS, REGRESSION SPLINES, NAIVE BAYES AND LOGISTIC MODELS

The sensitivity and specificity measures discussed in Section 3 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[21]. 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[22]. However, other cutoff values can be used.

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[23]. 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 raised, 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). 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[24]. 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)

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[25].

Table 5-1 shows the values of AUROC for each of ten model/software combinations in predicting an IME 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 ± one standard deviation. TREENET, MARS, and SPLUS Neural all do well with AUROC values about 0.7, significantly better than the logistic model. All Iminer models (Tree, Ensemble, Neural and Naïve Bayes) have AUROC values significantly below the top three performers, with two (Tree and Ensemble) significantly below the Logistic and Naïve Bayes benchmarks.

Area Under the ROC Curve - IME					
	CART Tree	**S-PLUS Tree**	**Iminer Tree**	**Treenet**	**Iminer Ensemble**
AUROC	0.669	0.688	0.629	0.701	0.649
Lower Bound	0.661	0.680	0.620	0.693	0.641
Upper Bound	0.678	0.696	0.637	0.708	0.657
	SPLUS Neural	**Iminer Neural**	**MARS**	**Iminer Naïve Bayes**	**Logistic**
AUROC	0.696	0.668	0.697	0.676	0.677
Lower Bound	0.688	0.660	0.690	0.669	0.669
Upper Bound	0.704	0.676	0.705	0.684	0.685

Table 5-1

Table 5-2 shows 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 ten 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					
	CART Tree	**S-PLUS Tree**	**Iminer Tree**	**Treenet**	**Iminer Ensemble**
AUROC	0.607	0.616	0.565	0.643	0.539
Lower Bound	0.598	0.607	0.555	0.634	0.530
Upper Bound	0.617	0.626	0.575	0.652	0.548
	SPLUS Neural	**Iminer Neural**	**MARS**	**Iminer Naïve Bayes**	**Logistic**
AUROC	0.611	0.623	0.628	0.615	0.612
Lower Bound	0.601	0.614	0.618	0.605	0.603
Upper Bound	0.621	0.632	0.637	0.625	0.621

Table 5-2

TREENET performs significantly better than all other model/software combinations with MARS and Iminer Neural also performing well. All three perform significantly better than the Logistic. Iminer Tree and Ensemble again do poorly on the SIU holdout sample. Figures 5-1 to 5-4 show the ROC curves for TREENET compared to the Logistic for both IME and SIU[26]. 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

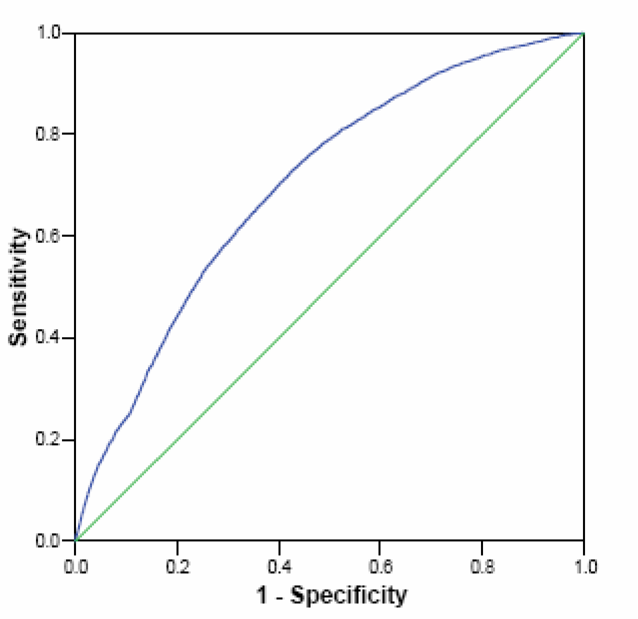

Figure 5-1

TREENET ROC Curve – SIU
AUROC = 0.677

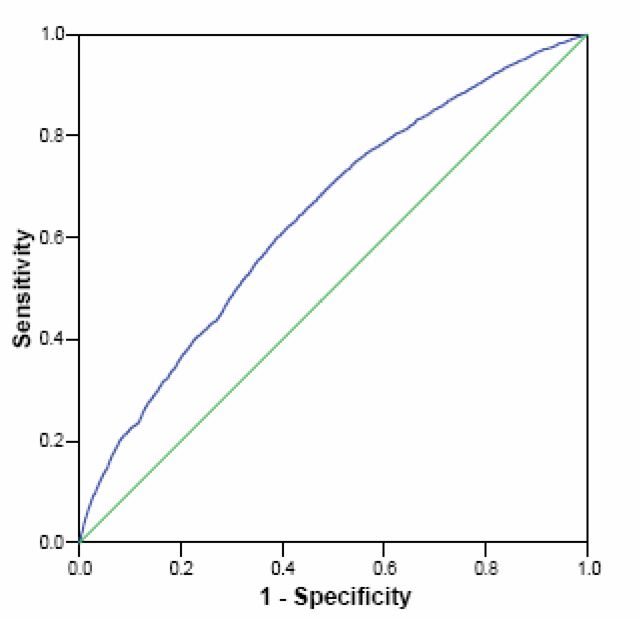

Figure 5-2

Logistic ROC Curve – IME
AUROC = 0.643

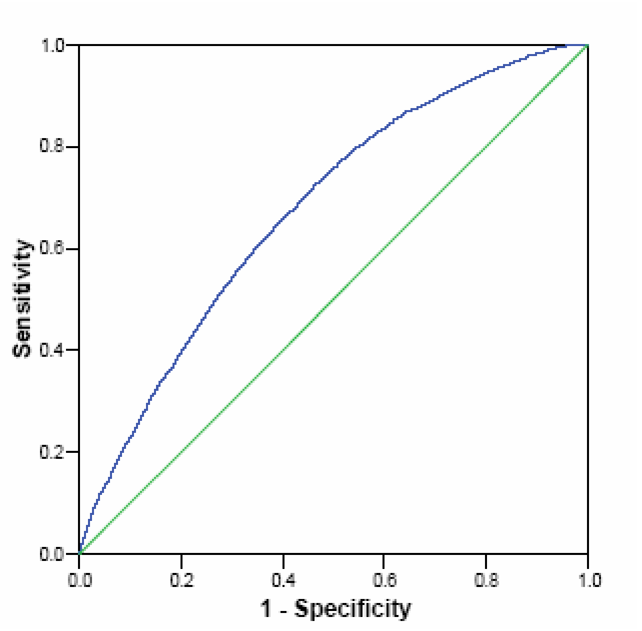

Figure 5-3

Logistic ROC Curve – SIU
AUROC = 0.612

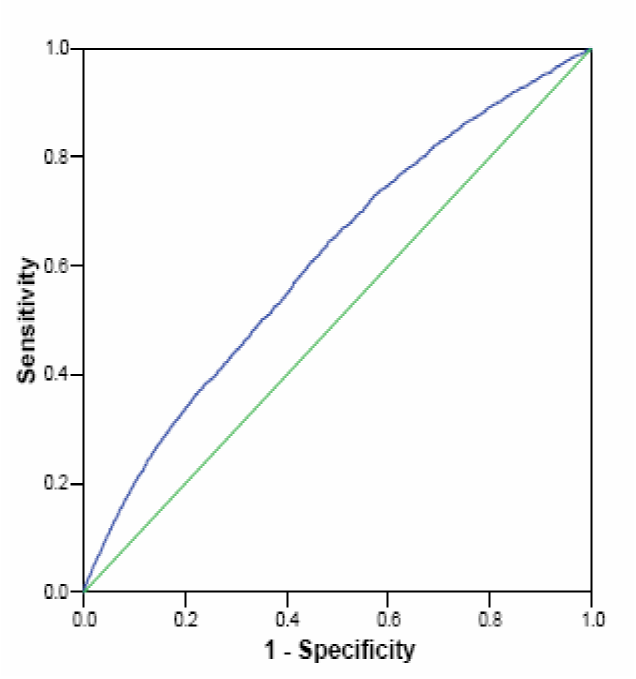

Figure 5-4

Finally, Table 5-3 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, MARS and SPLUS Tree do better than the benchmarks while CART Tree, Iminer Tree, and Iminer Ensemble do worse. Both SPLUS and Iminer Neural have mixed results.

Ranking of Methods By AUROC				
Method	**SIU AUROC**	**SIU Rank**	**IME Rank**	**IME AUROC**
Treenet	0.643	1	1	0.701
MARS	0.628	2	2	0.697
Iminer Neural	0.623	3	8	0.668
S-PLUS Tree	0.616	4	4	0.688
Iminer Naïve Bayes	0.615	5	6	0.676
Logistic	0.612	6	5	0.677
SPLUS Neural	0.611	7	3	0.696
CART Tree	0.607	8	7	0.669
Iminer Tree	0.565	9	10	0.629
Iminer Ensemble	0.539	10	9	0.649

Table 5-3

SECTION 6. 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, MARS, Neural Networks together with benchmark procedures of Naïve Bayes and Logistic regression. Those ten 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. Two target categorical variables were selected to model: The decision to request an independent medical examination (IME) or a special investigation (SIU). The two 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.

Nine 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 nine methods and two targets. All nine 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, MARS, and SPLUS Tree do better than the benchmark Naïve Bayes and Logistic methods, while CART tree, Iminer tree, and Iminer Ensemble do worse. Both SPLUS Neural and Iminer Neural have mixed results depending on the target. 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 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.

Derrig, R.A. and L. Francis, The Horse Race of Common Data Mining Applied to Auto Injury Claim Handling Decisions, Models and Software Implementations, Working Paper, 2005.

Derrig, R.A. and K.M. Ostaszewski, Fuzzy Sets Methodologies in Actuarial Science, Chapter 16, Practical Applications of Fuzzy Technologies, Hans-Jurgen Zimmermann, Editor (The Handbooks of Fuzzy Sets Series, D. Dubois and M. Prade, Editors), pp. 531-553, 1999.

Derrig, R.A. and H.I. Weisberg, AIB PIP Claim Screening Experiment Final Report: Understanding and Investigating the Claim Investigation Process, *AIB Filing on Cost Containment and Fraudulent Claims Payments*, DOI Docket R98-41, Boston, 1998.

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.

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

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.

Marzban, C., A Comment on the ROC curve and the Area Under it as Performance Measures, University of Oklahoma, 2004.

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.

Ostaszewski, K.M., An Investigation into Possible Applications of Fuzzy Sets Methods in Actuarial Science, *Society of Actuaries*, Schaumburg, Illinois, 1993.

Stephenson, D. B., Use of the “Odds Ratio” for Diagnosing Forecast Skill, Weather Forecasting, 2000, 15, 221-232.

Venables, W.N. and Ripley, B.D., Modern Applied Statistics with S-PLUS, third edition, Springer, 1999

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] See section 2 for an overview of the database and descriptions of the variables used for this paper.

[4] 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.

[5] See, for example, 2004 Discussion Paper Program, Applying and Evaluating Generalized Linear Models, May 16-19, 2004, Casualty Actuarial Society.

[6] This was the text used by the Casualty Actuarial Society for the exam on applied statistics during the 1980s

[7] 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.

[8] Combined payments under PIP and Medical Payments are reported to DCD.

[9] 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.

[10] 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.

[11] 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.

[12] The IRC also includes an index bureau check as one of the claims handling activities.

[13] 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).

[14] See Section 5 for the importance of variables in our study.

[15] MARS uses generalized cross validation. See Francis (2003b) for a description of the procedure.

[16] This would convert the numeric variable into a categorical variable with a level for every numeric value that is in the training data

[17] Generally by collapsing sparsely populated categories into an “all other” category

[18] In general, some programming is required to apply either approach in S-PLUS (R).

[19] The data set is described in more detail in Section 2 above.

[20] CART or TREENET could also have been used. However, we have a licensed copy only of MARS. Free access to the other products was generously supplied by Salford Systems.

[21] See Ostaszewski (1993) or Derrig and Ostaszewski (1999).

[22] 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

[23] A ROC curve is one example of a so-called “gains” chart.

[24] 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).

[25] The details of the formula were supplied by SPSS.

[26] All twenty ROC curves are available from the authors.