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

by Richard A. Derrig and Louise Francis

ABSTRACT

One of the most commonly used data mining techniques is decision trees, also referred to as classification and regression trees or C&RT. Several new decision tree methods are based on ensembles or networks of trees and carry names like TreeNet and Random Forest. Viaene et al. compared several data mining procedures, including tree methods and logistic regression, for modeling expert opinion of fraud/no fraud using a small fixed data set of fraud indicators or "red flags." They found that simple logistic regression did as well at matching expert opinion on fraud/no fraud as the more sophisticated procedures. In this paper we will introduce some publicly available regression tree approaches and explain how they are used to model four proxies for fraud in insurance claim data. We find that the methods all provide some explanatory value or lift from the available variables with significant differences in fit 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, claim investigation, decision trees

1. Introduction

In the past decade, computationally intensive techniques collectively known as *data mining* have gained popularity for explanatory and predictive applications in business. Many of the techniques, such as neural network analysis, have their roots in the artificial intelligence discipline. Data mining procedures include several that should be of interest to actuaries dealing with large and complex data sets. One of the most popular of the data mining tools, decision trees, originated in the statistics discipline, although an implementation of trees or classification and regression trees (C&RT) known as C4.5 was independently developed by artificial intelligence researchers. The seminal book by Brieman et al. (1993) provided an introduction to decision trees that is still considered the standard resource on the topic. Two reasons for the popularity of decision-tree techniques are (1) the procedures are relatively straightforward to understand and explain, and (2) the procedures address a number of data complexities, such as nonlinearities and interactions, that commonly occur in real data. In addition, software for implementing the technique, including both free open source as well as commercial implementations, has been available for many years.

While recursive partitioning, a common approach to estimation, underlies all the implementations of trees, there are many variations in the particulars of fitting methods across software products. For instance, different kinds of trees can be fit, including the classic single trees and the newer ensemble trees. Also, different goodness-of-fit measures can be used to optimize partitioning in creating the final tree, including deviance and the Gini index.

The four objectives of this paper are to

- describe the principal variations in tree methods;
- illustrate the application of tree methods to the identification of key parameters for the successful claim investigation on suspicion of fraud;[1]
- compare the accuracy of a number of tree-based data mining methods; and
- assess the impact of a few modeling tools and methods that different software implementations of tree-based methods incorporate.

A number of different tree methods, as well as a number of different software implementations of tree-based data mining methods, will be compared for their explanatory accuracy in the fraud application. Including the two baseline methods, eight combinations of methods and software are compared in this study. Our comparisons include several software implementations in order to show that specific implementations of the decision tree algorithms matter.

It should be noted that the tree-based software compared in this paper incorporate both algorithms and modeling techniques. The software products differ not only with respect to algorithms but also with respect to their modeling capabilities. Thus, graphical and statistical diagnostics, procedures for validation, and methods for controlling for over-parameterization vary across the software implementations, and this variability contributes to differences in accuracy (as well as practical usefulness) of the products.

The fraud analyses in this paper use data from a personal automobile bodily injury closed-claim database to explain the outcomes of four different fraud surrogates. This application is a classification application, where the modeler's objective is the identification of two or more distinct groups. Obviously, these methods can be used in other classification problems, such as the decision to underwrite specific types of risks.

Our selection of tree methods will be compared to two "baseline" prediction methods. The baseline prediction methods are (1) logistic re-

[1] See Derrig, 2002, for a general discussion of fraud in insurance claims.

gression and (2) naïve Bayes. The baseline methods were selected as computationally efficient procedures that make simplifying assumptions about the relationship between explanatory and target variables. We use straightforward implementations of the two methods without an attempt to optimize the hyperparameters.[2] Viaene et al. (2002) applied a wider set of procedures, including neural networks, support vector machines, and a classic general linear model, logistic regression, on a small single data set of insurance claim fraud indicators or "red flags" as predictors of expert opinion on the suspicion of fraud. They found that simple logistic regression did as well as the more sophisticated procedures at predicting expert opinion on the presence of fraud.[3] 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. There will be a number of distinct differences between the data and modeling targets used in our analysis and that of Viaene et al. They applied their methods to a database with only 1,400 records, while our database contained approximately 500,000 records, more typical of a database size for current data mining applications. In addition, most of the predictors used by Viaene et al. were binary, that is, they could take on only two values, whereas the data for this study contain a more common mixture of numeric variables and categorical variables with many potential values, such as treatment lag in days and zip code.

A wide variety of statistical software is now available for implementing fraud and other explanatory and predictive models through clustering and data mining. In this paper we will introduce a variety of C&RT (pronounced "cart," but in this paper CART refers to a specific software product) approaches[4] and explain in general how they are used to model complex dependencies in insurance claim data. We also investigate the relative performance of a few software products that implement these models. As an illustrative example of relative performance, we test for the key claim variables in the decision to investigate for excessive or fraudulent practices in a large claim database. The software programs we will investigate are CART, S-PLUS/R-TREE, TreeNet, Random Forests, and Insightful Tree and Ensemble from the Insightful Miner package. The naïve Bayes benchmark method is from Insightful Miner, while logistic regression is from R/S-PLUS. The data used for this analysis are the auto bodily injury liability closed claims reported to the Detailed Claim Database (DCD) of the Automobile Insurers Bureau of Massachusetts from accident years 1995 through 1997.[5] Three types of variables are employed. Several variables thought to be related to the decision to investigate are included in the DCD, such as outpatient provider medical bill amounts. A few other variables 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 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 in

[2] In other words, the baseline methods were applied in an automated way. No attempt was made to optimize variable selection or variable transformation, or search for significant interaction terms. The baseline methods are intended to be simple and easily implemented examples of their class of modeling procedures and thus serve as an easy-to-hit modeling target that may be able to be improved.

[3] 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, pp. 400–401).

[4] A wider set of data mining techniques is considered in Derrig and Francis (2006).

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

terms of a reduction or denial of the otherwise indicated claim payment, are the four modeling targets.

Eight modeling software results for each modeling target are compared for effectiveness based on a standard evaluation technique, the area under the receiver operating characteristic curve (AUROC) as described in Section 4. We find that the methods all provide some explanatory value or lift from the DCD variables, used as independent variables, with significant differences in accuracy among the eight methods and four targets. 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 significantly on the explanatory ability of the logistic model, while some software/methods are less accurate. The different 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.[6] This exercise should provide practicing actuaries with guidance on regression tree software and market methods to analyze complex and nonlinear relationship commonly found in all types of insurance data.

The paper is organized as follows. Section 1 covers the general setting for the paper. 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 appear in Section 3. In Section 4 we describe software for modeling nonlinearities. Comparative outcomes for each software implementation are described in Section 5 with numerical results shown in Section 6. Implications for the use of the software models for explanatory and predictive applications are discussed in Section 7.

[6]In a companion paper we show how "important" each variable is within and across software implementations (Derrig and Francis 2006).

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 been 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] are reported to DCD. 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.[9] This will allow us to divide the sample into large training, test, and holdout subsamples, 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), and attorney presence. Note that many insurance databases, including the DCD, do not contain data or variables indicating whether a particular claim is suspected of fraud or abuse. For such databases, other approaches, such as unsupervised learning methods, might be applied.[11] In the DCD data, there are three claims handling techniques for mitigat-

[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. Our sample removed all personal and company identifying information to form an analytic subset of the actual data.

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

[9]BIL claim payments are the sum of the liability claim payment plus a no-fault subrogation paid by the third-party carrier. Thus they are representative of the full third-party liability on each claimant's injuries.

[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]See Brockett et al. (2002) for the use of the unsupervised PRIDIT method to assign suspicion of fraud scores.

ing claims cost of fraud or abuse that are reported when present, as well as outcome, and formulaic savings amounts for each of the techniques. These variables can serve as surrogates of suspicion of fraud and abuse but they stand on their own as applied investigative techniques.

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.[12] 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, and sometimes also by expert systems that review the billing and treatment patterns.[13] Favorable outcomes are reported by insurers when the damages are mitigated, when the billing and treatment are curtailed, and when the claimant refuses to undergo the IME or does not show.[14] 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 (Derrig and Weisberg 2004).

Special Investigation (SIU) is reported when claims are handled through nonroutine investigative techniques (accident reconstruction, examinations under oath, and surveillance are the expensive 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 within the special investigation category 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 special investigation.

For purposes of this analysis and demonstration of models and software, we employ 21 potential explanatory variables and four target variables. The target variables are prescribed field variables of DCD. Thirteen predicting variables are numeric, two from DCD fields (F), eight from internal demographic type derived data (DV), and three from external demographic data (DM), as shown in Table 1. A frequent data-mining practice is to "derive" explanatory or predictive variables from the primary dataset to be "mined" by creating summary statistics of informative subsets such as RANK ATT/ZIP, the rank of a simple count of the number of attorneys in the Massachusetts zip code with BIL claims. While many such variables are possible, we use only a representative few such derived variables, denoted by DV.

The choice of predictor variables was guided by prior published research on insurance fraud and data mining. Thus, certain provider-related variables, such as attorney involvement, the amount of the provider 1 and provider 2 bills, and the type of medical provider are included. In addition, certain variables related to claimant behavior, such as amount of time between occur-

[12]This fact is a matter of Massachusetts law which at the time appeared to permit only IMEs by the same type of physician, say a chiropractor, as the type is treating physician. Recently, the Massachusetts Supreme Court in *Boone v. Commerce Insurance, 451 Mass. 198 (2008)* clarified that IMEs must be conducted by physicians of a similar, but not necessarily exactly the same, specialty. This situation may differ in other jurisdictions.

[13]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.

[14]The standard Massachusetts auto policy has a claimant cooperation clause for IME both in the first party PIP coverage and in the third party BI liability coverage.

Table 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
Claimant Age	155,438	DV	0	104	34.15	15.55
Claim Report Lag (Days)	162,709	DV	0	2,793	47.94	144.44
Treatment Lag (Days)	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	1	1,305	30.84	91.65
Valid N (listwise)	70,397					

N=Number of nonmissing 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. 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%
1993 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%)
1 SIU Done	162,761	F	Special Investigation Done (7%), No SIU (93%)
2 IME Done	162,761	F	Independent Medical Examination Done (8%), No IME (92%)
3 SIU Favorable	162,761	F	Special Investigation Favorable (3.4%), Not Favorable/Not Done (95.6%)
4 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 nonmissing 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.

rences of the accident and reporting of the claim and amount of time between occurrences of the accident and the first outpatient medical treatment are included. Geographic and claim risk indicators, i.e., rating territory and distance from claimant to provider are also used. The four rank variables[15] are calculated to represent the relative claim, attorney, and provider activity at zip code and city levels. One important caveat of this analysis is that it is based on closed claims, so some of the variables, such as the amount billed by outpatient medical providers, may not be fully known until the claim is nearly closed. When building a model to detect fraud and abuse *prospectively*, the modeler will be restricted to information available relatively early in the life of a claim or to probabilistic estimates of final values dependent on that early information. Tables 1 and 2 list the explanatory variables we use that are numeric and categorical, respectively.

[15]Each of the rank variables uses the ranking, not the actual data values.

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

Similar claim investigation variables are now being collected by the Insurance Research Council (IRC) in their periodic sampling of countrywide injury claims (IRC 2004a, pp. 89–104).[16] 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 liability claims and (2) Massachusetts is a fairly urban state with high claim frequencies and more dubious claims.[17] In fact, a recent IRC study shows Massachusetts has the highest percentage of BI claims in no-fault states that are suspected of fraud (23%) and/or buildup (41%) (IRC 2004b, p. 25). It is, therefore, entirely consistent for the Massachusetts claims to exhibit more nonroutine claim handling techniques. Favorable outcomes average about 50% 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 programs that implement them, in the next two sections before we describe how they are applied to model the IME and SIU target variables.

3. Data mining and software models

3.1. How tree models handle data complexity

Traditional actuarial and statistical techniques often assume that the functional relationship between the independent variables and the dependent or target variable is linear or that some transformation of the variables exists so that techniques can treat the relationship as linear. Insurance data, however, often contain variables where the relationship among variables is complex and not susceptible to transformation to the linear world. 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 complex relations without any assumption being made about the nature of the nonlinearity. We illustrate how each of our tree methods reviewed in this paper models nonlinearities in the following two relatively simple examples. The variables in this example were selected because of a known nonlinear relationship between independent and dependent variables.

Table 3. Example 1 and 2 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%

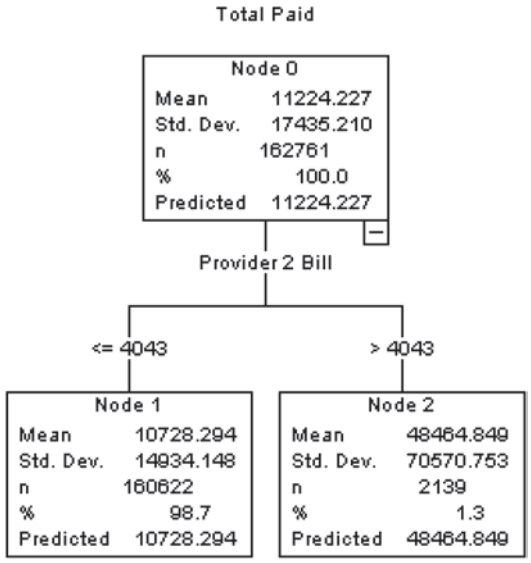
EXAMPLE 1 The dependent variable, a numeric variable, is total paid losses and the independent variable is provider 2 bill. Table 3 displays average paid losses at various bands of provider 2 bill.

EXAMPLE 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.

[16]The IRC also includes an index bureau check as one of the claims handling activities but this practice is universal in Massachusetts.

[17]Prior studies of Massachusetts Auto Injury claim data for fraud content included Weisberg and Derrig 1998, (suspicion regression models), Derrig and Weisberg 1998, (claim screening with scoring models) and Derrig and Weisberg 2004, (effect of investigative technique on claim settlements).

Figure 1. CART example of parent and children nodes: Total paid as a function of provider 2 bill



3.2. Trees

In this section we give a brief introduction to decision tree methodology. Hadidi (2003) provides a more thorough introduction to trees and their use in insurance. Trees, also known as classification and regression trees (C&RT), 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 to the dependent variable.[18] A goodness-of-fit statistic is used to select the split which maximizes the difference between the two children 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 1, with the value \$4,043 for provider 2 bill as the cutpoint in this example.

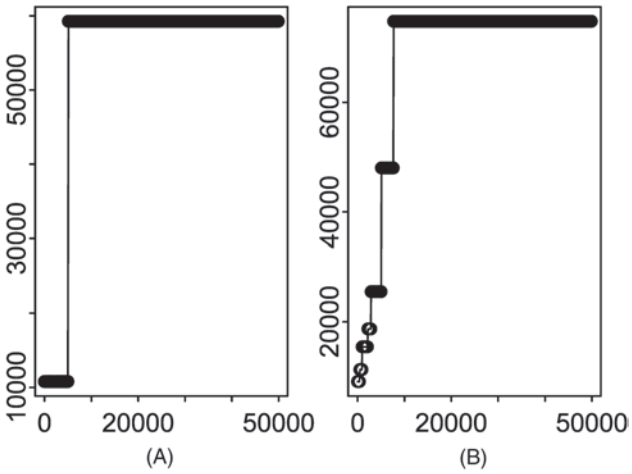
[18]There are Tree Software models that may split nodes into three or more branches. The CHAID and Exhaustive CHAID techniques in SPSS classification trees are an example of such software.

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 a residual sum of squares or some function of the residual errors such as the R^2 or F statistic. For this example, all values of provider 2 bill are searched and a split is made at the value \$4,043. All claims with provider 2 bills less than or equal to \$4,043 go to the left node and "explain" a total paid of \$10,728 and all claims with provider 2 bill greater than \$4,043 go to the right node, and "explain" a total paid of \$48,465. This is depicted in Figure 1. The tree graph shows that the total paid mean is significantly lower for the claims with provider 2 bills less than \$4,043.

Alternatively, when a predictor variable is categorical, all possible two-way groupings of the categories of the variable are tested. The grouping that optimizes the goodness-of-fit measure for the variable is the one that is used to partition the data into two groups that are significantly different with respect to the value of the dependent variable. For instance, if injury type is used to model paid losses, the data can be split into a group including back/neck injuries, neck sprains, other sprains, minor lacerations, and a group that includes all other injuries. The mean claim payments of these two groups are approximately \$9,000 and \$25,000, respectively.

One statistic often used as a goodness-of-fit measure to optimize tree splits is the 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 7.00×10^{13}. The SSE declines to 4.65×10^{13} after the data are par-

Figure 2A and B. CART example with two and seven nodes: Total paid as a function of provider 2 bill



titioned using $4,043 as the cutpoint. Any other partition of the provider bill produces a larger SSE than 4.65×10^{13}. For instance, if a cutpoint of $10,000 is selected, the SSE is 4.76×10^{13}.

The two nodes in Figure 1 can each be split into two more children nodes and these can then be further split. The sequential splitting continues until no (significant) 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, i.e., the final nodes, are 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 2A, with only two terminal nodes, the fitted function is flat until $4,043, steps up to a higher value, and then remains flat. Figure 2B displays the predicted values of a tree with seven terminal nodes. The steps or increases are more gradual for this function.

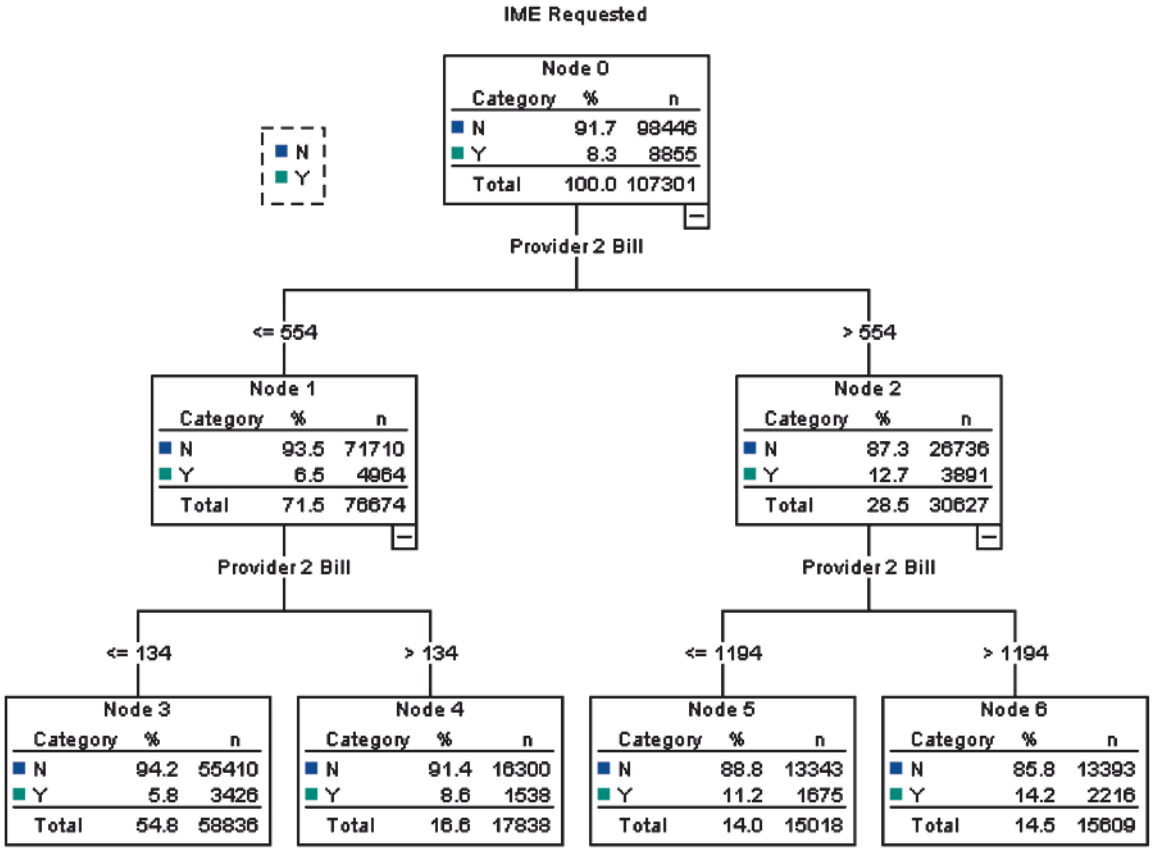
The procedure for modeling data where the dependent variable is categorical (binary in our example) is similar to that of a numeric variable. For instance, one of the binary fraud surrogates is independent medical exam (IME) requested, yes or no. The target classifications are claimants for whom an IME was requested and the group of (presumably legitimate) claimants where an IME was not requested. At each step, the tree procedure selects the split that best improves or lowers node heterogeneity. 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 are used in different products to select the optimal split. These include entropy/deviance and Gini index. Kantardzic (2003), Breiman et al. (1993), and Venibles and Ripley (1999) describe the computation and application of the Gini index and entropy/deviance measures.[19] 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 terminal node 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 of interest; otherwise it is assigned to the remaining class.

Figure 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 3B.

Tree models use categorical as well as numeric independent variables in modeling complex data.

[19]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.

Figure 3A. CART example with seven nodes IME proportion as a function of provider 2 bill



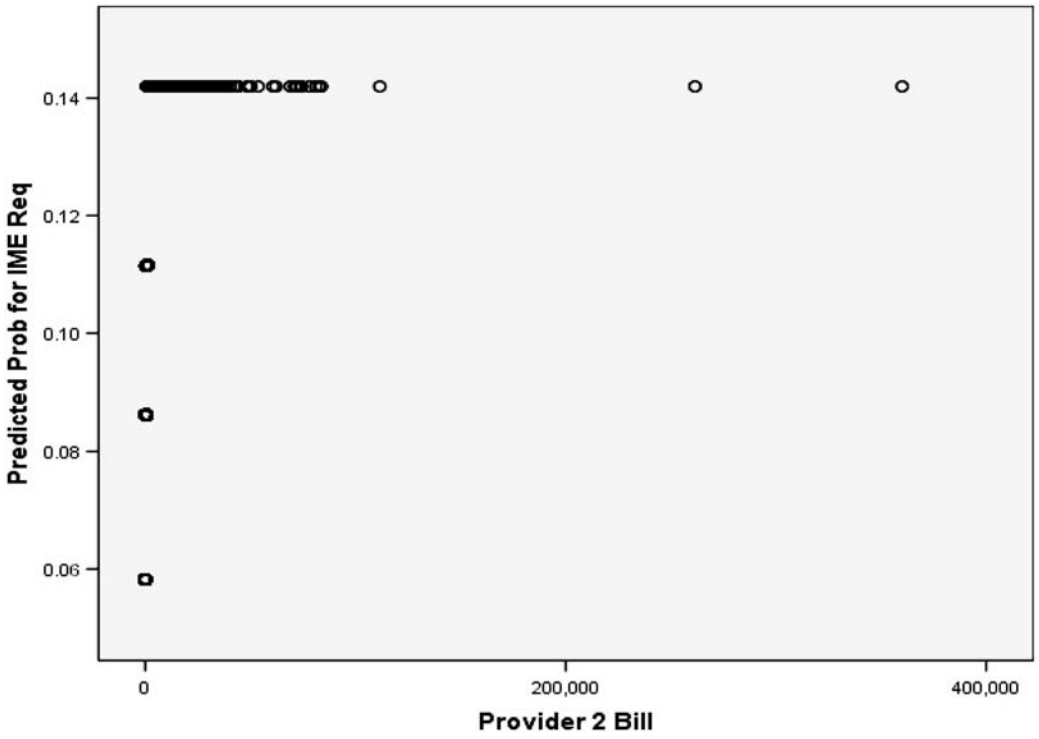
However, because the levels on categorical data may not be ordered, all possible two-way splits of categorical variables must be considered before the data are partitioned.

The final tree model in a data mining application will typically use many predictor variables, with different predictor variables incorporated into the model at different nodes. To prevent overfitting, the technique of pruning is used. This technique uses a goodness-of-fit measure to "prune" or eliminate branches and nodes based on eliminating those that do not contribute a significant improvement in the model's fit. Pruning may be thought of as analogous to backwards stepwise regression, where the modeler fits a full regression incorporating all possible predictor variables, and then one by one removes those variables that are not deemed "significant" by a goodness-of-fit measure, such as the F statistic.

3.3. 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. The ensemble model's prediction is a weighted average of the single tree predictions. In the data mining literature the technique is often referred to as "boosting" (Hastie, Tibshirani, and Friedman 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 on a test sample, is computed, and, in applications involving

Figure 3B. CART example with four step functions: IME proportion as a function of provider 2 bill



numeric dependent variables, is used in another round of fitting as a dependent variable.[20] For models with categorical dependent variables, the error is often computed as a rate or proportion, i.e., the number of total records for a given node of a tree that were misclassified, compared to the sample size of the node. For both numeric and categorical dependent variables, the error is also typically used in the computation of weights in subsequent rounds of fitting, with records containing larger errors receiving higher weighting in the next round of estimation.

One algorithm for computing weights is described by Hastie, Tibshirani, and Friedman.[21] Consider an ensemble of trees $1,2,\ldots,M$. The error for the mth tree measures the departure of the actual from the fitted value on the test data after the mth model has been fit. When the dependent variable is categorical, as it is in the illustrative fraud application in this paper, a common error measure used in boosting is:

$$\text{err}_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq F_m(\mathbf{x}_i))}{\sum_{i=1}^{N} w_i} \qquad (3.1)$$

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 categorical indicator dependent variable, $\mathbf{x}$ is a matrix of predictors, and $F_m(\mathbf{x}_i)$ is the prediction for the ith record of the mth tree.

[20] If only two models were fit, one to the dependent variable and one to the residual from that model, the predicted value would be $\hat{Y}_1 + wt_1\hat{e}_1$, where $\hat{Y}_1$ is the first round fitted value for the dependant variable and $\hat{e}_1$ is the fitted value of $Y - \hat{Y}_1$ and wt_1 is a weight. Therefore, $E(Y) = \hat{Y}_1 + wt_1 E(Y - \hat{Y}_1)$. That is, each tree in the sequence further refines the fit by estimating the error for the observation and adding it back to the prediction. When the dependent variable is categorical, one can think of the actual value for an observation as either 0.0 or 1.0, and the error as a positive or negative value between 0.0 and 1.0 representing the deviation between the probability of the target variable from the model and the actual value of the dependent variable. The actual implementation in boosting is typically more complex than this intuitive description.

[21] See p. 301. Note that the literature also describes other error and weight functions.

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

$$\alpha_m = \log\left(\frac{1-\text{err}_m}{\text{err}_n}\right) \qquad (3.2)$$

and the new weight is $w_{i,m+1} = w_m \exp(\alpha_m I(y_i \neq F_m(\mathbf{x}_i)))$.

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

The formulas presented above describe a general approach to boosting.[22] The specific boosting procedures implemented differ among software products. For instance, TreeNet (Freidman 2001) uses stochastic gradient boosting.[23] One of the modeling issues one encounters when applying ensemble models is that of overfitting; that is, the model being estimated contains many parameters and its fit is generally assessed by applying a goodness-of-fit measure only to the training sample. The model may provide an excellent fit to the training data but a poor fit to validation or holdout data. 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 validation[24] data. Another procedure used by TreeNet to control overfitting is *shrinkage* or *regularization*. Regularization refers to any procedure that prevents overfitting. For instance, variable selection techniques such as stepwise regression mitigate overfitting by selecting only significant variables for inclusion in the model. One kind of regularization is shrinkage, where credibility is perhaps the most familiar example of shrinkage to actuaries. The approach shrinks or reduces the parameter estimate associated with a given predictor variable[25] in a given model. For instance, a credibility procedure might "shrink" the weight assigned to a given class, moving the estimate for the class towards the overall mean. The objective of shrinkage is to reduce the tendency of a model with many parameters to overfit, or to fit noise, rather than pattern in the data.[26]

Alternatively, the Insightful Miner (Iminer) Ensemble model employs a simpler implementation of boosting which does not use shrinkage and applies nonstochastic boosting using 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. Note that:

- Using a large collection of trees allows many different variables to be used. Some of these would not be used in smaller simpler models.[27]
- Many different models are used. The predictive modeling literature (Hastie et al. 2001; Francis 2003) indicates that composites of multiple models perform better than the prediction of a single model.[28]
- Different training and validation records are used in parameterizing the model (with stochastic gradient boosting as implemented in TreeNet and bagging as implemented in Random Forest, but not the Iminer Ensemble tree). This makes the procedure more robust to the influence of a few extreme observations.

[22] Boosting has been applied to fraud detection modeling in Viaene, Derrig, and Dedene (2004).

[23] TreeNet is compared to the non-tree method of neural networks for fraud detection in Francis (2005).

[24] A validation sample is used by many modeling procedures to tune the parameters of a model. That is, it is a separate sample from the training sample that is incorporated as part of the goodness of fitting used in estimating the model. However, as it is not the sample the model was fit to, it is resistant to overfitting. An additional sample, referred to as the test sample, is used for a final test of the models goodness of fit, after a final model is developed.

[25] Or it may shrink or reduce the parameter estimate associated with a function, as in the case of stochastic gradient boosting

[26] See Harrell (2001), pp. 61–64, for a more detailed description of shrinkage.

[27] Note that the ensemble tree methods employ all 21 variables in the models.

[28] The ROC curve results in Section 6 show that TreeNet and Random Forest generally provide the best explanatory models for the Massachusetts data.

The method of fitting many (often 100 or more) small trees results in fitted curves which are almost smooth. Figures 4A and 4B display two nonlinear functions fit to total paid and IME requested variables by the TreeNet ensemble model and show the increased flexibility of the output functions compared to the simple tree step functions.

3.4. Ensemble models—bagging

Bagging is an ensemble approach based on resampling or bootstrapping. Bagging is an acronym for "bootstrap aggregation" (Hastie et al. 2001). 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 many random samples of records in the data to fit many trees. For instance, an analyst may decide to take 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 50% sample. The trees fit are unpruned and are not necessarily small trees with 5 to 10 terminal nodes as with boosting. Each tree may have a different number of terminal nodes. By averaging the predictions of a number of bootstrap samples, typically using a simple average of all the models fit, bagging reduces the prediction variance.[29] The implementation of bagging used in this paper is known as Random Forest. Brieman (2001) points out that using different variables, as well as different records in the different trees in the Random Forest ensemble, seem to reduce the correlation between the different models fit and improve the accuracy of the overall prediction. For the analysis in this paper, one-third of the dataset was sampled as a training set for each tree fit, while one-third was used as a validation or "out of bag" sample for assessing the goodness-of-fit of the tree at that iteration. The remaining third was the test sample. (See "Validation and Testing" in Section 4 for a more detailed discussion of training, validation, and test samples).

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

3.5. Naïve Bayes

The naïve Bayes method is relatively simple and easy to implement. In our comparison, we treat it as one of two benchmark data mining methods. 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 a logistic regression models as the second benchmark.

The naïve Bayes method was developed for categorical data. Specifically, both dependent and independent variables are categorical.[30] 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.[31]

The key simplifying assumption of the naïve Bayes method is the assumption of independence. Given each category of the dependent variable, 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

[29] Hastie, Tibshirani, and Friedman describe how the estimate resulting from bagging is similar to a posterior Bayesian estimate.

[30] Although our implementation of naïve Bayes uses only categorical predictor variables, some implementations allow numerical predictors.

[31] The numeric variables were grouped into five bins or into quintiles in this instance.

Figure 4A. Ensemble prediction of total paid

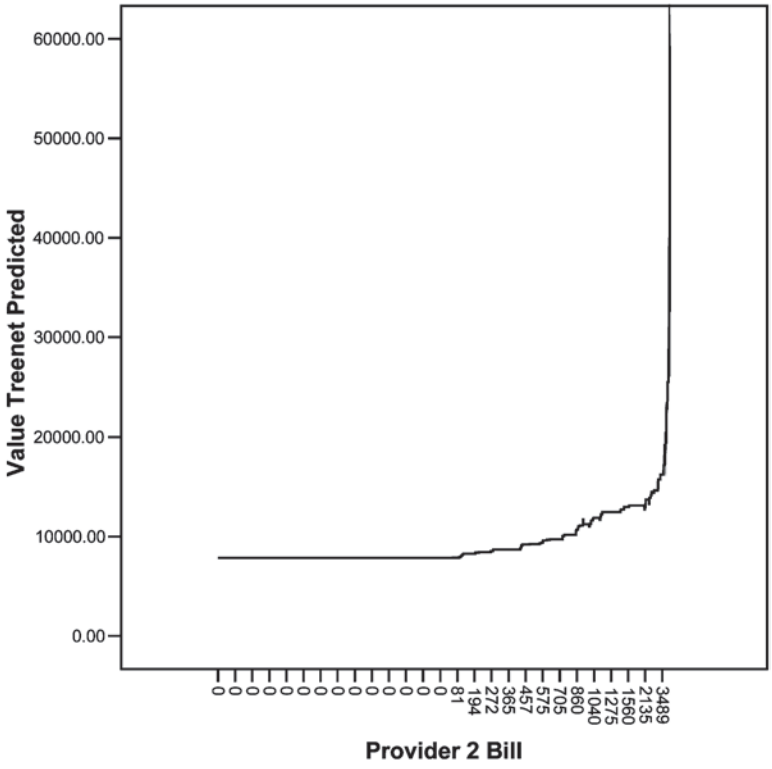


Figure 4B. Ensemble prediction of IME requested

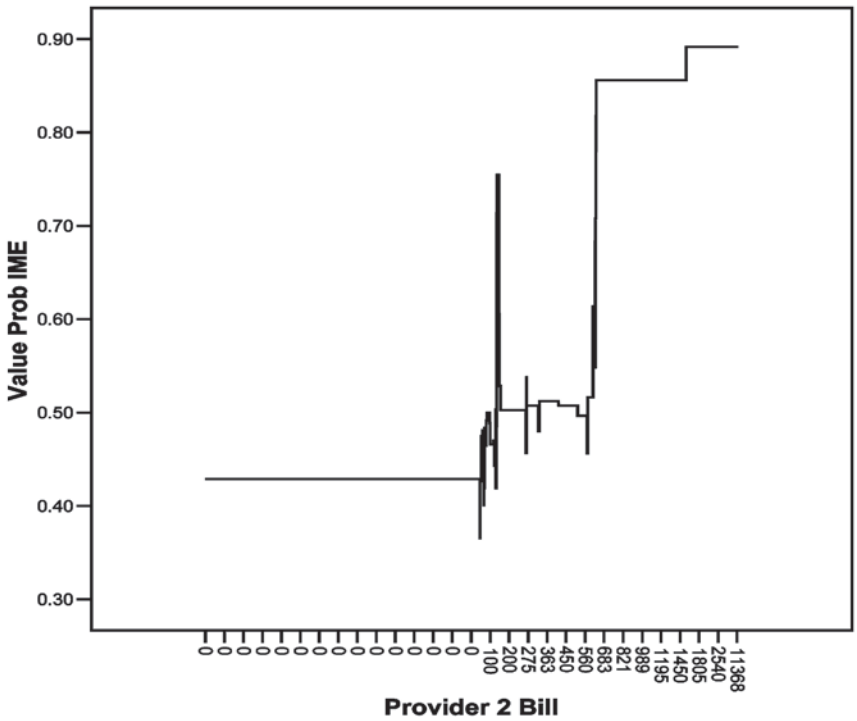


Figure 5A. Random Forest prediction of Total Paid

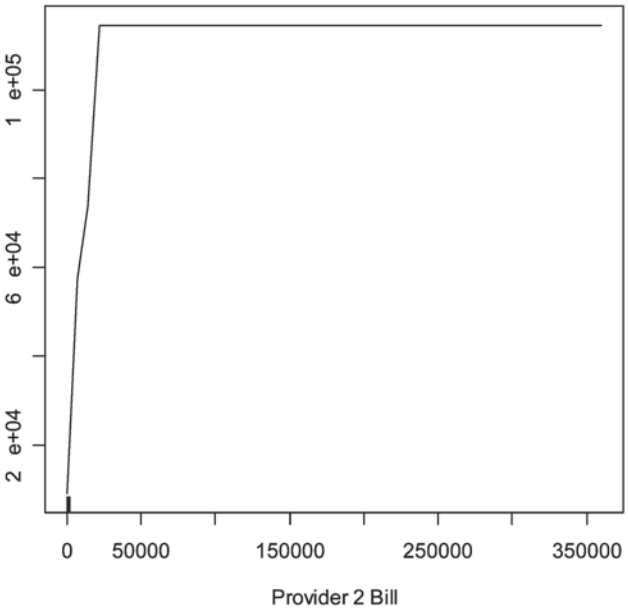
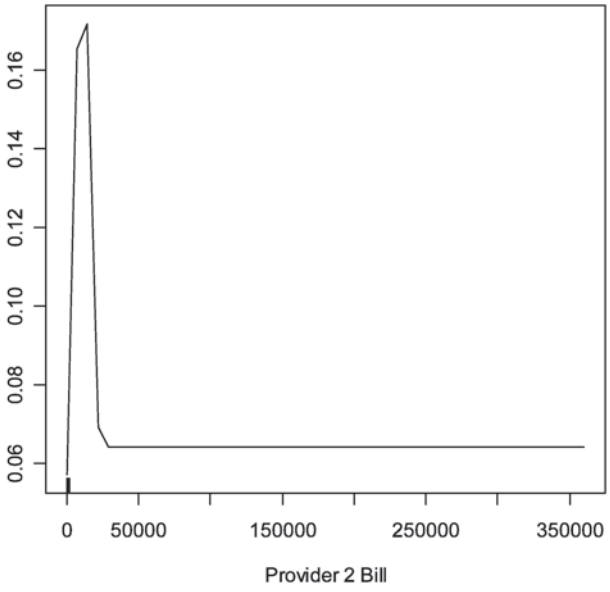


Figure 5B. Random Forest prediction of IME



vector $X = \{x_i\}$ is in category $C = \{c_j\}$ of the dependent variable.

$$P(c_j \mid \mathbf{X}) = P(\mathbf{X} \mid c_j)P(c_j)/P(\mathbf{X}). \qquad (3.3)$$

Because of the naïve 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

$$P(\mathbf{X} \mid c_j) = \prod_i P(x_i \mid c_j). \qquad (3.4)$$

The method is described in more detail in Kantardzic (2003). To illustrate the use of naïve 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 bills, only four categories are created by the binning procedure. The new variable was used to estimate the IME targets. Figure 6 displays a bar plot of the predicted probability of an IME for each of the groups. Figure 7 displays the fitted function. This function is a step function which changes value at each boundary of a provider 2 bill bin.

3.6. 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 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 2003).

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

$$Y = a + b_1X_1 + b_2X_2 + b_3 * X_1 * X_2. \qquad (3.5)$$

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

Figure 6. Bayes predicted probability IME requested vs. quintile of provider 2 bill

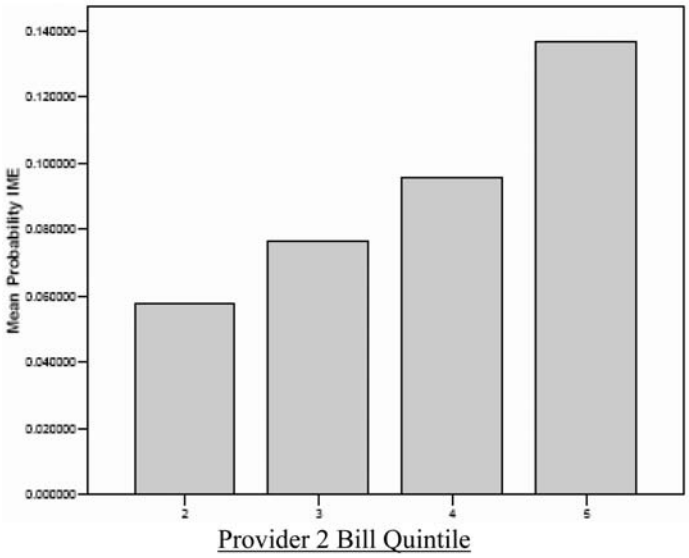
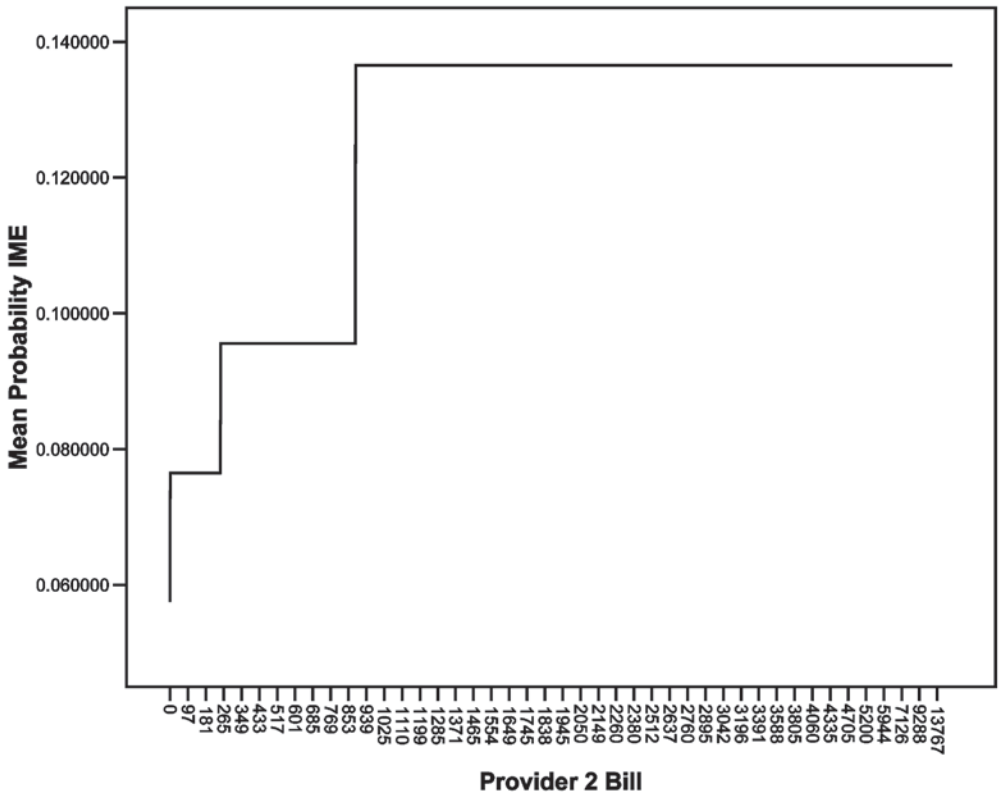


Figure 7. Naïve Bayes predicted IME vs. provider 2 bill



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 (Jaccard 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. Each of the tree-based data mining techniques used in this paper has efficient methods for searching for significant interactions.

- 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 data mining 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.
- 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 analytic comparison study, no attempt was made to incorporate interaction terms as this procedure lacks an efficient way to search for the significant interactions.

3.7. 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 (at each node) of partitioning the data until no further improvement in fit can be obtained. Different variables can be used to partition the data at different nodes.

4. Software for modeling nonlinear dependencies and testing the models

4.1. Software for modeling nonlinear dependencies

Four software products were included in our fraud comparison: They are CART, TreeNet, S-PLUS (R) and Insightful Miner.[32] One of the products, R, is open source software that can be downloaded and used for free,[33] while the other products are commercial software. A description of some of the features of each product, including features that are useful to the practitioner but have no impact on the model's accuracy can be found in a related pdf document posted on the Casualty Actuarial Society's *Variance* Web Site.

4.2. Validating and testing

It is common for data mining practitioners to partition the data into three groups (Hastie, Tibshirani, and Friedman 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"

[32] Software products used in the comparison were based on (1) software licensed to the authors, (2) free software, and (3) software that the authors were granted temporary use of by the company licensing the software.

[33] See Fox 2002 for a discussion of R.

sample, is used to obtain an unbiased test of the model's accuracy. Cross validation is an alternative approach to a validation sample that is especially appropriate when the sample size used in the analysis is relatively modest. 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 five-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.[34] S-PLUS (R) allows either approach[35] 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 or logistic regression[36] to tune the parameters of the model during fitting.[37]

In this analysis, approximately a third of the data, about 50,000 records, was used as the hold-out sample for the final testing and comparison of the models. Two key statistics often used to compare model 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 those events. The *specificity* is the percentage of non-events (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, often referred to as a confusion matrix (Hastie, Tibshirani, and Friedman 2001), presents an example of the calculation. In the example confusion matrix, 800 of 1,000 nonevents are predicted to be nonevents so the sensitivity is 80%. The specificity is 67% since 400 of 600 true positives are accurately predicted.

Table 4. 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%

The sensitivity and specificity measures discussed above 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.[38] A common cutoff point is 50% and records with scores greater than 50% are classified as events and records with scores below that value are classified as nonevents.[39] 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

[34]The TreeNet software also allows cross-validation, but it is easier and probably more typical to use a validation sample.

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

[36]For a more extensive discussion of logistic regression see Hosmer and Lemshow (1989).

[37]With parametric models such as logistic regression, validation samples are not usually part of the estimation procedure. The implementation of naïve Bayes used here did not provide a validation procedure for tuning the model.

[38]See Ostaszewski (1993) or Derrig and Ostaszewski (1995).

[39]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.

characteristic (ROC) curve.[40] 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 is.[41] ROC curves have been used in prior studies of insurance claims and fraud detection regression models (Derrig and Weisberg 1998; 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 0.50 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 nonparametric method was used to compute the standard error of the AUROC.[42]

We show the AUROC results for our analyses in Section 6.

5. Modeling the decision to investigate and favorable outcome

The remainder of this paper is devoted to illustrating the usefulness and effectiveness of eight model/software combinations applied to our four fraud applications, the decision to investigate via IMEs or referral to SIU, and favorable outcomes from IME or SIU referrals. 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.[43] We employ 21 potentially predicting variables of three types: (1) 11 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.

[40] A ROC curve is one example of a so-called "gains" chart.

[41] ROC curves were developed extensively for use in medical diagnosis testing in the 1970s and 1980s (Zhou, McClish, and Obuchowski 2002 and more recently in weather forecasting (Marzban 2004).

[42] The details of the formula are supplied in SPSS documentation that can be found at the SPSS Web Site, www.spss.com.

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

The eight model/software combinations we will use here are

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

Numbers 1–6 are six tree models, and 7–8 are benchmark models (naïve Bayes and logistic).

CART and TreeNet are Salford Systems standalone software products that perform 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 procedure used here in the fraud comparison is found in both S-PLUS and in a freeware version in R though S-PLUS does not contain an implementation of Random Forest,[44] while R does. The S-PLUS GLM (generalized linear models) was used for logistic regression. The naïve Bayes, Tree, and Ensemble Tree procedures from Insightful Miner are used here in the fraud comparison.

Many of the products used in this comparison contain the capability of ranking variables in importance to the model. This capability is useful in variable selection and as an aid in understanding and interpreting the model. Because the methods for ranking variables and the results of the ranking are significant additional topics, they are outside the scope of this paper.[45]

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 in Section 7.

[44]For a discussion of the Random Forest freeware in R see Liaw and Wiener 2003.

[45]See Derrig and Francis 2006 for a discussion of the importance ranking of the 21 variables used here. Each independent variable may have a differenent "importance" level depending on the target variable.

Table 5. 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	0.699	0.676	0.677
Lower Bound	0.641	0.692	0.669	0.669
Upper Bound	0.657	0.707	0.684	0.685

Table 6. 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	0.697	0.670	0.677
Lower Bound	0.643	0.676	0.660	0.667
Upper Bound	0.665	0.697	0.681	0.687

6. Comparing the models

Tables 5–8 show the values of AUROC for each of eight model/software combinations in predicting a decision to investigate with an IME (Table 5), a favorable IME outcome (Table 6), the decision to refer for special investigation SIU (Table 7), and favorable SIU investigation outcome (Table 8). For all models, the AUROC in the comparison used sample test data. Upper and lower bounds for the "true" AUROC value are shown as the AUROC value ± two standard deviations (95% confidence level). TreeNet and Random Forest both do well with AUROC values 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

Table 7. 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	0.667	0.615	0.612
Lower Bound	0.530	0.658	0.605	0.603
Upper Bound	0.548	0.677	0.625	0.621

Table 8. Area under the ROC curve—SIU favorable

	CART Tree	S-PLUS Tree	Iminer Tree	TreeNet
AUROC	0.598	0.603	0.547	0.678
Lower Bound	0.584	0.589	0.555	0.667
Upper Bound	0.612	0.617	0.575	0.689
	Iminer Ensemble	Random Forest	Iminer Naïve Bayes	Logistic
AUROC	0.575	0.643	0.607	0.610
Lower Bound	0.530	0.630	0.593	0.596
Upper Bound	0.548	0.657	0.625	0.623

above the benchmarks on SIU and IME decisions but below on SIU and IME favorable.

We note that, in general, the model scores for SIU 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.

Figures 8 to 11 show the ROC curves for TreeNet compared to the Logistic for both IME and SIU Decisions.[46] 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.

Tables 9 and 10 display the relative performance of the model/software combinations according to AUROC values and their ranks. With naïve Bayes and Logistic as the benchmarks,

Figure 8. TreeNet ROC Curve—IME AUROC = 0.701

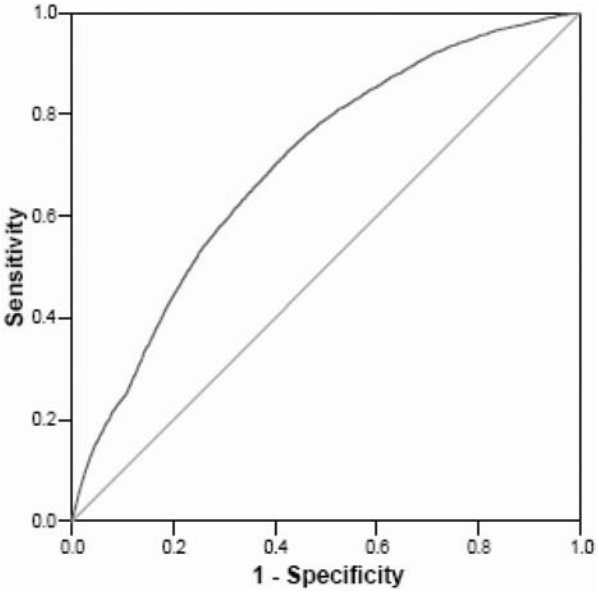
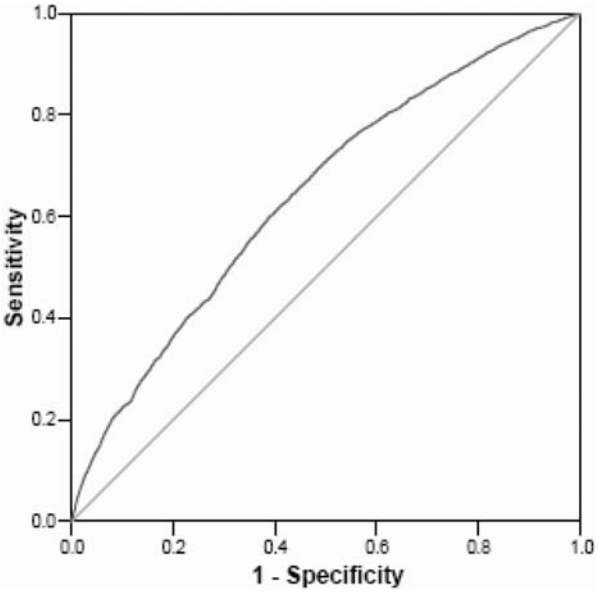


Figure 9. TreeNet ROC Curve—SIU AUROC = 0.643



TreeNet, and Random Forest do better than the benchmarks, while CART, Iminer Tree, and Iminer Ensemble do worse.

Finally, Figures 12A and 12B show the relative performance in a graph. 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

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

Figure 10. Logistic ROC Curve—IME AUROC = 0.677

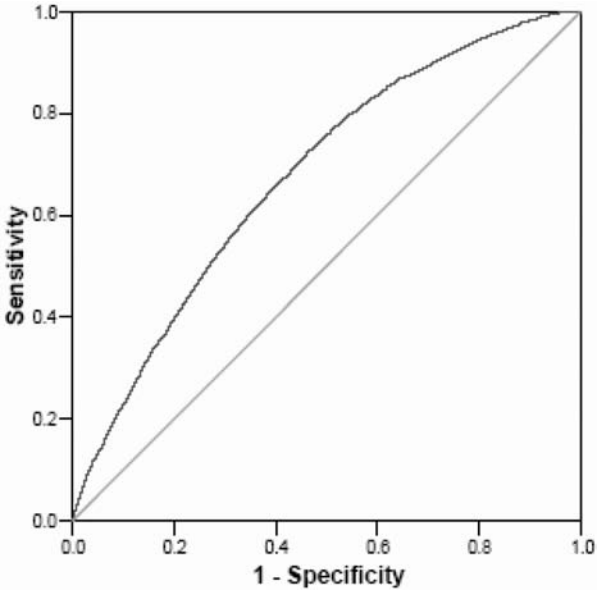
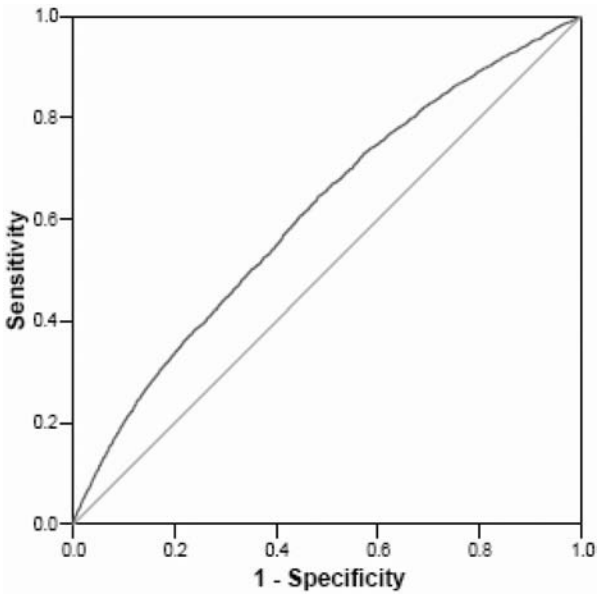


Figure 11. Logistic ROC Curve—SIU AUROC = 0.612



by positions farther to the right and closer to the top of the square. This graph clearly shows that TreeNet and Random Forest procedures do better than the other tree procedures and the benchmarks.

7. Conclusion

The focus of this paper has been to

- introduce a class of data mining techniques with potential applications in insurance,
- compare a number of different techniques and software implementations of those techniques, and
- illustrate applications of the techniques in insurance to fraud modeling for claims.

Insurance data often involve both large volumes of information and nonlinearity and complexity 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 the principal advantages of the efficient handling of large data sets and the fitting of nonlinear functions to that data. In this paper we illustrate the use of software implementations of six classification and regression tree methods together with benchmark procedures of naïve Bayes and logistic regression. Those eight model/software combinations are applied to closed claim data arising in the Detail Claim Database (DCD) of auto injury liability claims in Massachusetts. Twenty-one variables were selected to use in the explana-

Table 9. Ranking of methods by AUROC—decision

Method	SIU AUROC	SIU Rank	IME Rank	IME AUROC
Random Forest	0.667	1	2	0.699
TreeNet	0.643	2	1	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 10. Ranking of methods by AUROC—favorable

Method	SIU AUROC	SIU Rank	IME Rank	IME AUROC
TreeNet	0.678	1	2	0.683
Random Forest	0.643	2	1	0.697
S-PLUS Tree	0.603	5	5	0.664
Logistic	0.610	3	3	0.677
Iminer Naïve Bayes	0.607	4	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

Figure 12A. Plot of AUROC for SIU vs IME decision

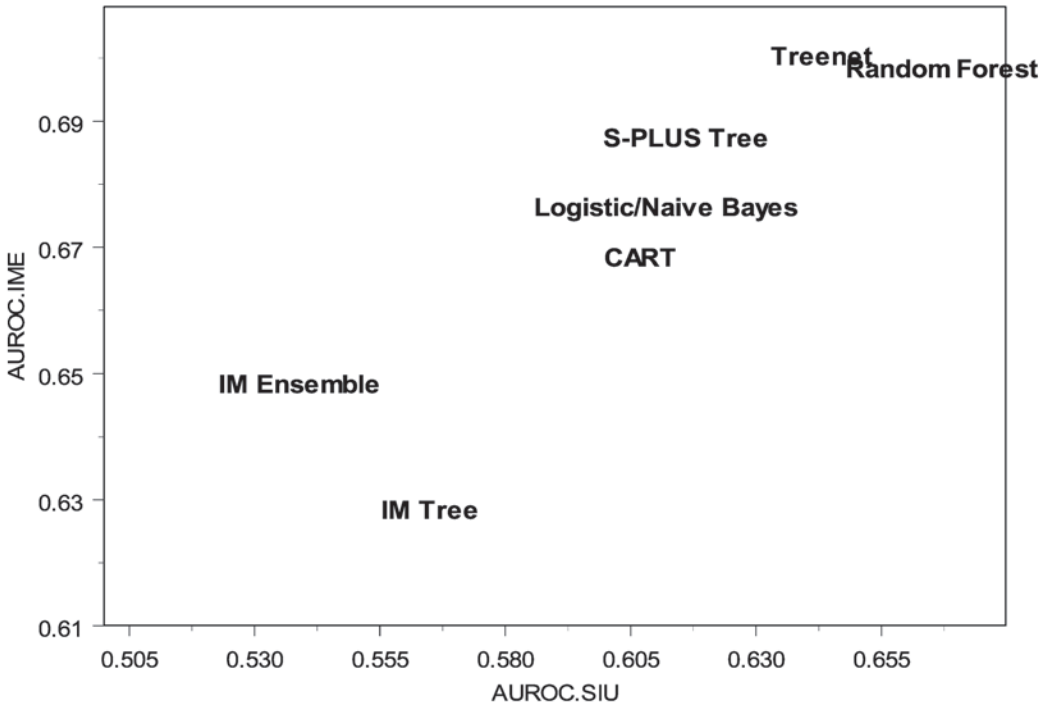
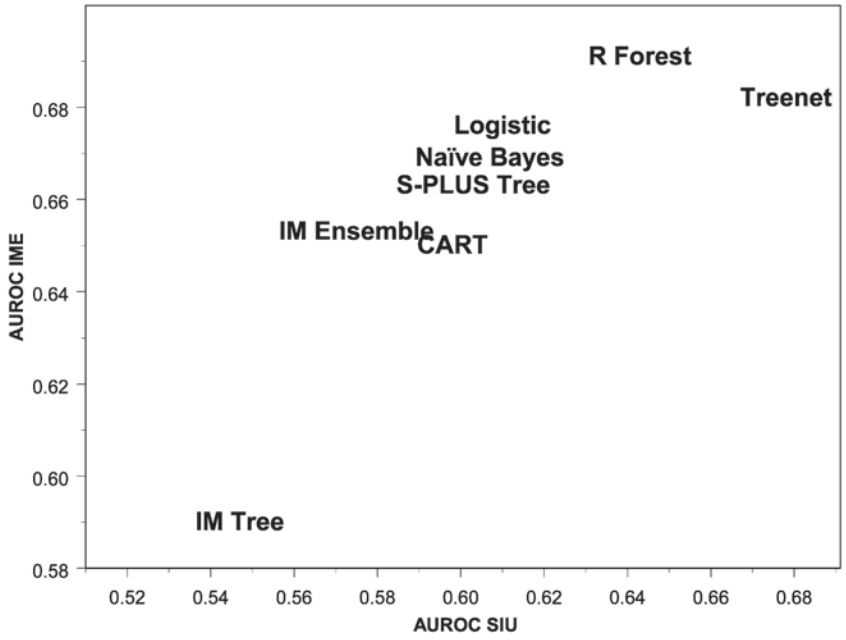


Figure 12B. Plot of AUROC for SIU vs IME favorable



tory 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, and outright fraud. Of course, there is

also an interest in modeling the conditions under which the investigation will be successful. All our models are *explanatory*, relating the key portions from the closed claim data with the investigation variables. Applying these techniques to produce a real-time *predictive* model must recognize that closed claim data arrives at different points in time. The explanatory models will at least guide the process of choosing informative features for real-time predictive models. It should be noted that the topic of insurance fraud and its drivers is an area of active research. For instance, the December 2002 issue of the *Journal of Risk and Insurance* was devoted entirely to fraud research and the Web Site *www.derrig.com* lists many studies at the IFRR tab. That said, the focus of this paper is the comparison of techniques applied to a large complex insurance database, rather than specific findings related to fraud and abuse investigations.

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 at the 95% level 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 that some software/methods can improve on the predictive ability of the logistic model. Straightforward applications of TreeNet and Random Forest do significantly better than the benchmark naïve Bayes and logistic methods, while Iminer Tree and Iminer Ensemble do significantly worse. That some model/software combinations do better than the benchmarks is due to the relative size and richness of this data set and/or the types of independent variables at hand compared to the Viaene data (2002).

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 (with and without various bells and whistles) that implement these and other data mining techniques.

References

AIB, *See* Automobile Insurers Bureau of Massachusetts.

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

Breiman, L., J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, New York: Chapman and Hall, 1993.

Breiman, L., "Random Forests," *Machine Learning* 45, 2001, pp. 5–32.

Brockett, P. L., R. A. Derrig, L. L. Golden, A. Levine, and M. Alpert, "Fraud Classification Using Principal Component Analysis of RIDITs," *Journal of Risk and Insurance* 69, 2002, pp. 341–371.

Derrig, R. A., and K. Ostaszewski, "Fuzzy Techniques of Pattern Recognition in Risk and Claim Classification," *Journal of Risk and Insurance* 62, 1995, pp. 447–482.

Derrig, R. A., "Insurance Fraud," *Journal of Risk and Insurance* 69, 2002, pp. 271–287.

Derrig, R. A., and L. Francis, "Comparison of Methods and Software for Modeling Nonlinear Dependencies: A Fraud Application," International Congress of Actuaries, June 2006, Paris.

Derrig, R. A., and H. I. Weisberg, "AIB PIP Claim Screening Experiment Final Report—Understanding and Improving the Claim Investigation Process," Massachusetts Division of Insurance, DOI, R98-4, July 1998.

Derrig, R. A., and H. I. Weisberg, "Determinants of Total Compensation for Auto Bodily Injury Liability under No-Fault: Investigation, Negotiation and the Suspicion of Fraud," *Insurance and Risk Management* 71, 2004, pp. 633–662.

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," *Data Mining and Decision Support*, eds. D. Mlandenic, Nada Lavrac, Marko Bohanec, and Steve Moyle, Boston: Kluwer Academic, 2003.

Fox, J., *An R and S-PLUS Companion to Applied Regression*, Thousand Oaks, CA: SAGE Publications, 2002.

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

Francis, L. A., "A Comparison of TreeNet and Neural Networks in Insurance Fraud Prediction," paper presented at CART Data Mining Conference, 2005, San Francisco, CA.

Friedman, J., "Greedy Function Approximation: The Gradient Boosting Machine," *Annals of Statistics* 29, 2001, pp. 1189–1232.

Hadidi, N., "Classification Ratemaking Using Decision Trees," Casualty Actuarial Society *Forum*, Winter 2003, pp. 253–283.

Harrell, F., *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression and Survival Analysis*, New York: Springer, 2001.

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

Hosmer, D. W., and S. Lemshow, *Applied Logistic Regression*, New York: Wiley, 1989.

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

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

IRC, *See* Insurance Research Council.

Jaccard, J., and R. Turrisi, *Interaction Effects in Multiple Regression*, Thousand Oaks, CA: SAGE, 2003.

Kantardzic, M., *Data Mining*, New York: Wiley, 2003.

Liaw, A., and M. Wiener, "Classification and Regression by Random Forest," *R News* 2/3, 2003, pp. 18–22.

Marzban, C., *A Comment on the ROC Curve and the Area Under it as Performance Measures*, Center for Analysis and Prediction of Storms, Norman, OK: University of Oklahoma, 2004.

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

Venebles, W., and B. Riplye, *Modern Applied Statistics with S-PLUS* (3rd ed.), New York: Springer, 1999.

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* 16, 2004, pp. 612–620.

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* 69, 2002, pp. 373–421.

Weisberg, H. I., and R. A. Derrig, "Methodes Qualitatives Pour la Detection des Demandes d'Indemnisation Frauduleuses," *RISQUES* 35, July–September 1998, pp. 75–101.

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