Martian Chronicles: Is MARS better than Neural Networks?

by Louise Francis, FCAS, MAAA

Abstract:
A recently developed data mining technique, Multivariate Adaptive Regression Splines (MARS) has been hailed by some as a viable competitor to neural networks that does not suffer from some of the limitations of neural networks. Like neural networks, it is effective when analyzing complex structures which are commonly found in data, such as nonlinearities and interactions. However, unlike neural networks, MARS is not a "black box", but produces models that are explainable to management.

This paper will introduce MARS by showing its similarity to an already well-understood statistical technique: linear regression. It will illustrate MARS by applying it to insurance fraud data and will compare its performance to that of neural networks.

Acknowledgements:
The author thanks the Automobile Insurers Bureau of Massachusetts for supplying the data used in this paper and wishes to acknowledge the efforts of Richard Derrig in providing insight into the data and approaches to analyzing it. The author also acknowledges the following people who reviewed this paper and provided many constructive suggestions: Jane Taylor, Francois Morrisette, Christopher Yaure, Patricia Francis-Lyon and Virginia Lambert.

Martian Chronicles: Is MARS better than Neural Networks?

The discipline of artificial intelligence has contributed a number of promising techniques to the analyst's toolkit. The techniques have names such as "machine learning", "genetic algorithms" and "neural networks". These techniques are collectively known as data mining. Data mining uses computationally intensive techniques to find patterns in data. When data mining tools are applied to data containing complex relationships they can identify relationships not otherwise apparent. These complexities have been a challenge for traditional analytical procedures such as linear regression.

The casualty actuarial literature contains only a few papers about data mining techniques. Speights *et al.* (Speights *et al.*, 1999) and Francis (Francis, 2001) introduced the neural network procedure for modeling complex insurance data. Hayward (Hayward, 2002) described the use of data mining techniques in safety promotion and better matching of premium rates to risk. The methods discussed by Hayward included exploratory data analysis using pivot tables and stepwise regression.

In this paper, a new technique, MARS, which has been proposed as an alternative to neural networks (Steinberg, 2001), will be introduced. The name MARS, coined for this technique by its developer, Freidman, (Hastie, *et al.*, 2001), is an acronym for Multivariate Adaptive Regression Splines. The technique is a regression based technique which allows the analyst to use automated procedures to fit models to large complex databases. Because the technique is regression based, its output is a linear function that is readily understood by analysts and can be used to explain the model to management. Thus, the technique does not suffer from the "black box" limitation of neural networks. However, the technique addresses many of the same data complexities addressed by neural networks.

Neural networks are one of the more popular data mining approaches. These methods are among of the oldest data mining methods and are included in most data mining software packages. Neural networks have been shown to be particularly effective in handling some complexities commonly found in data. Neural networks are well known for their ability to model nonlinear functions. The research has shown that a neural network with a sufficient number of parameters can model any continuous nonlinear function accurately.[1] Francis (Francis, 2001) also showed that neural networks are valuable in fitting models to data containing interactions. Neural networks are often the tools of choice when predictive accuracy is required. Berry and Linoff (Berry and Linoff, 1997) suggest that neural networks are popular because of their proven track record.

Neural networks are not ideal for all data sets. Warner and Misra presented several examples where they compared neural networks to regression (Warner and Misra, 1996). Their research showed that regression outperformed neural networks when the functional relationship between independent and dependent variables was known. Francis (Francis,

[1] A more technical description of the property is that with a sufficient number of nodes in the neural network's hidden layer, the neural network can approximate any deterministic nonlinear continuous function.

2001) showed that when the relationship between independent and dependent variables was linear, classical techniques such as regression and factor analysis outperformed neural networks.

Perhaps the greatest disadvantage of neural networks is the inability of users to understand or explain them. Because the neural network is a very complex function, there is no way to summarize the relationships between independent and dependent variables with functions that can be interpreted by data analysts or management. Berry and Linoff (Berry and Linoff, 1997) state that "Neural networks are best approached as black boxes with mysterious inner workings, as mysterious as the origins of our own consciousness". More conventional techniques such as linear regression result in simple mathematical functions where the relationship between predictor and target variables is clearly described and can be understood by audiences with modest mathematical expertise. The "black box" aspect of neural networks is a serious impediment to more widespread use.

Francis (Francis, 2001) listed several complexities found in actual insurance data and then showed how neural networks were effective in dealing with these complexities. This paper will introduce MARS and will compare and contrast how MARS and neural networks deal with several common data challenges. Three challenges that will be addressed in this paper are:

1) Nonlinearity: Traditional actuarial and statistical techniques often assume that the functional relationship between the independent variables and the dependent variable is linear or some transformation of the data exists that can be treated as linear.
2) Interactions: The exact form of the relationship between a dependent and independent variable may depend on the value of one or more other variables.
3) Missing data: Frequently data has not been recorded on many records of many of the variables that are of interest to the researcher.

The Data

This paper features the application of two data mining techniques, neural networks and MARS, to the fraud problem. The data for the application was supplied by the Automobile Insurers Bureau of Massachusetts (AIB). The data consists of a random sample of 1400 closed claims that were collected from PIP (personal injury protection or no-fault coverage) claimants in Massachusetts in 1993. The database was assembled with the cooperation of ten large insurers. This data has been used by the AIB, the Insurance Fraud Bureau of Massachusetts (IFB) and other researchers to investigate fraudulent claims or probable fraudulent claims (Derrig *et al.*, 1994, Weisberg and Derrig, 1995, Viaene *et al.,* 2002). While the typical data mining application would use a much larger database, the AIB PIP data is well suited to illustrating the use of data mining techniques in insurance. Viaene *et al.* used the AIB data to compare the performance of a number of data mining and conventional classification techniques (Viaene *et al.*, 2002).

Two key fraud related dependent variables were collected in the study: an overall assessment (ASSESS) of the likelihood the claim is fraudulent or abusive and a suspicion score (SUSPICION). Each record in the data was assigned a value by an expert. The value indicates the expert's subjective assessment as to whether the claim was legitimate or whether fraud or abuse was suspected. Experts were asked to classify suspected fraud or abuse claims into the following categories: exaggerated damages, opportunistic fraud or planned fraud. As shown in Table 1, the assessment variable can take on 5 possible values. In addition, each claim was assigned a score from 0 (none) to 10 (very high) indicating the expert's degree of suspicion that the claim was abusive or fraudulent. Weisberg and Derrig (Weisberg and Derrig, 1993) found that more serious kinds of fraud, such as planned fraud were associated with higher suspicion scores than "softer" fraud such as exaggeration of damages. They suggest that the suspicion score was able to measure the range of "soft" versus "hard" fraud.

The database contains detailed objective claim information on each claim in the study. This includes information about the policy inception date, the date the accident occurred, the date it was reported, the paid and incurred loss dollars, the injury type, payments to health care providers and the provider type. The database also contains "red flag" or fraud indicator variables. These variables are subjective assessments of characteristics of the claim that are believed to be related to the likelihood of fraud or abuse. More information on the variables in the model is supplied below in the discussion of specific models.

Table 1

Assessment Variable

Value	Assessment	Percent of Data
1	Probably legitimate	64%
2	Excessive treatment only	20%
3	Suspected opportunistic fraud, no injury	3%
4	Suspected opportunistic fraud, exaggerated injury	12%
5	Suspected planned fraud	1%

We may use the more inclusive term "abuse" when referring to the softer kinds of fraudulent activity, as only a very small percentage of claims meet the strict standard of criminal fraud (Derrig, 2002). However, misrepresentation and exaggeration of the nature and extent of the damages, including padding of the medical bills so that the value of the claim exceeds the tort threshold, occur relatively frequently. While these activities are often thought of as fraud, they do not meet a legal definition of fraud. Therefore, they will be referred to as abuse. Overall, about one third of the claims were coded as probable abuse or fraud claims.

Nonlinear Functions

The relationships encountered in insurance data are often nonlinear. Classical statistical modeling methods such as linear regression have had a tremendous impact on the analysis and modeling of data. However, traditional statistical procedures often assume that the relationships between dependent and independent variables are linear. Traditional modeling also allows linear relationship that result from a transformation of dependent or independent variables, so some nonlinear relationships can be approximated. In addition, there are techniques specifically developed for fitting nonlinear functions such as nonlinear regression. However, these techniques require that theory or experience specify the "true" form of the nonlinear relationships. Data mining techniques such as neural networks and MARS do not require that the relationships between predictor and dependent variables be linear (whether or not the variables are transformed). Both neural networks and MARS are also considered nonparametric because they require no assumptions about the form of the relationship between dependent and independent variables.

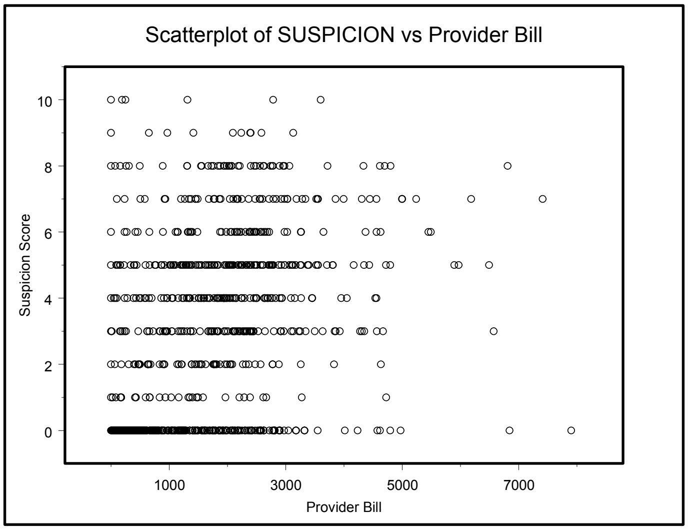
For this illustration, a dependent variable that is not categorical (i.e. values have a meaningful order) was selected. The selected dependent variable was SUSPICION. Unlike the ASSESS variable, the values on the SUSPICION variable have a meaningful range, with higher values associated with suspicion of more serious fraud.

To illustrate methods of fitting models to nonlinear curves, a variable was selected which 1) had a significant correlation with the dependent variable, and 2) displayed a highly nonlinear relationship. Illustrating the techniques is the objective of this example. The data used may require significant time to collect and may therefore not be practical for an application where the objective is to predict abuse and fraud (which would require data that is available soon after the claim is reported). Later in the paper, models for prospectively predicting fraud will be presented. The variable selected was the first medical provider's bill[2]. A medical provider may be a doctor, a clinic, a chiropractor or a physical therapist. Prior published research has indicated that abusive medical treatment patterns are often key drivers of fraud (Derrig *et al.*, 1994, Weisberg and Derrig, 1995). Under no-fault laws, claimants will often deliberately run the medical bills up high enough to exceed tort thresholds. In this example the relationship between the first provider's medical bill and the value of the suspicion score will be investigated. The AIB fraud database contains the medical bills submitted from the top two health care providers. If more costly medicine is delivered to suspicious claims than non-suspicious claims, the provider bills should be higher for the suspicious claims.

Figure 1 presents a scatterplot of the relationship between SUSPICION and the provider bill. No relationship is evident from the graph. However, certain nonlinear relationships can be difficult to detect visually.

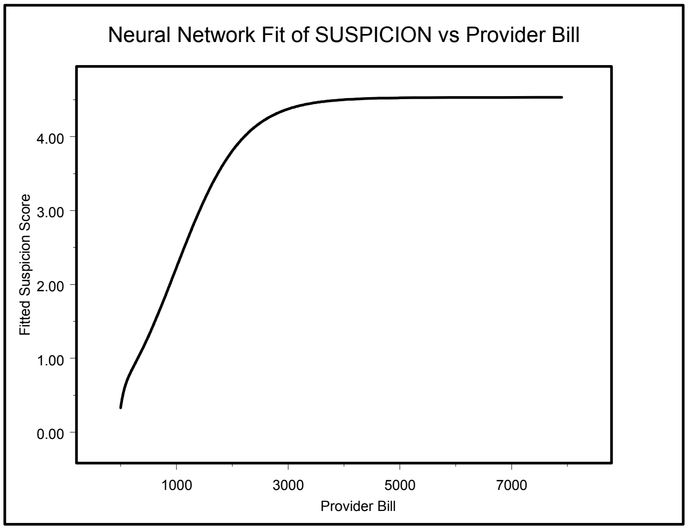
[2] Note that Massachusetts PIP covers only the first $8,000 of medical payments if the claimant has health insurance. Large bill amounts may represent data from claimants with no coverage. Bills may also exceed $8,000 even if payments are limited. However, the value of medical bills on some claims may be truncated because reimbursement is not expected.

Figure 1



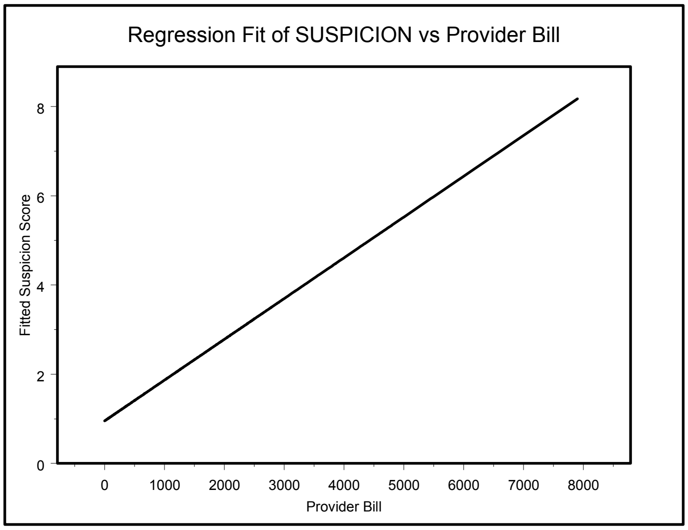
Neural networks will first be used to fit a curve to the data. A detailed description of how neural networks analyze data is beyond the scope of this paper. Several sources on this topic are Francis, Lawrence and Smith (Francis, 2001, Lawrence, 1994, Smith, 1996). Although based upon how neurons function in the brain, the neural network technique essentially fits a complex non-parametric nonlinear regression. A task at which neural networks are particularly effective is fitting nonlinear functions. The graph below displays the resulting function when the dependent variable SUSPICION is fit to the provider bill by a neural network. This graph displays a function that increases quickly at lower bill amounts and then levels off. Although the curve is flat over much of the range of medical bills, it should be noted that the majority of bills are below $2,000 (in 1993 dollars).

Figure 2



One of the most common statistical procedures for curve fitting is linear regression. Linear regression assumes the relationship between the dependent and independent variables is linear. Figure 3 displays the graph of a fitted regression line of SUSPICION on provider bill. The regression forces a linear fit to SUSPICION versus the payment amount. Thus, rather than a curve with a rapidly increasing trend line that levels off, a line with a constant slope is fitted. If the relationship is in fact nonlinear, this procedure is not as accurate as that of the neural network.

Figure 3

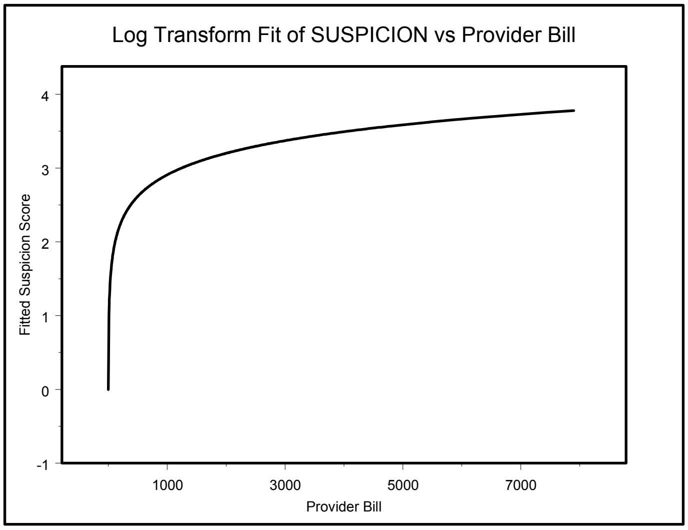


When the true relationship between a dependent and independent variable is nonlinear, various approaches are available when using traditional statistical procedures for fitting the curve. One approach is to apply a nonlinear transformation to the dependent or independent variable. A linear regression is then fit to the transformed variables. As an example, a log transform was applied to the provider bill variable in the AIB data. The regression fit was of the form:

$$Y = B_o + B_1 \ln(X)$$

That is, the dependent variable, the suspicion score, is assumed to be a linear function of the natural log of the independent variable, provider bill. Figure 4 displays the curve fit using the logarithmic transformation.

Figure 4

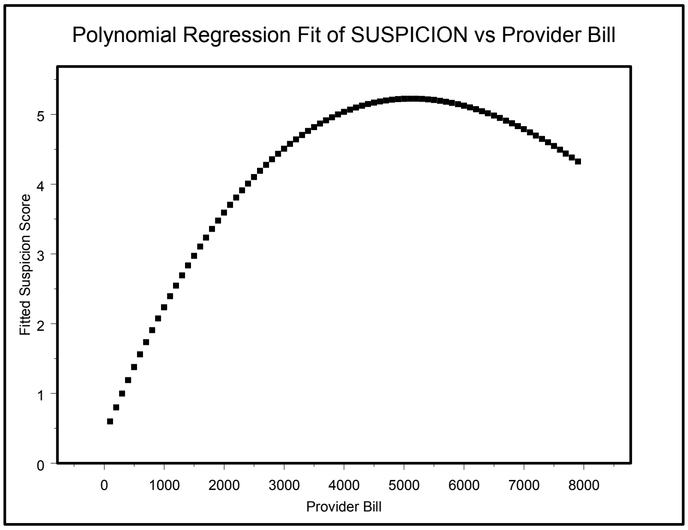


Another procedure which is used in classical linear regression to approximate nonlinear curves is polynomial regression. The curve is approximated by the function:

$$Y = B_o + B_1X + B_2X^2 + \ldots + B_nX^n$$

Generally, low order polynomials are used in the approximation. A cubic polynomial (including terms up to provider bill raised to the third power) was used in the fit. Figure 5 displays a graph of a fitted polynomial regression.

Figure 5



The use of polynomial regression to approximate functions is familiar to readers from its use in Taylor series expansions for this purpose. However, the Taylor series expansion is used to approximate a function near a point, rather than over a wide range. When evaluating a function over a range, the maximums and inflection points of the polynomial may not exactly match the curves of the function being approximated.

The neural network model had an R^2 (coefficient of determination) of 0.37 versus 0.25 for the linear model and 0.26 for the log transform. The R^2 of the polynomial model was comparable to that of the neural network model. However, the fit was influenced strongly by a small number of claims with large values. Though not shown in the graph, at high values for the independent variable the curve declines below zero and then increases again. This unusual behavior suggests that the fitted curve may not approximate the "true" relationship between provider bill and suspicion score well at the extremes of the data and may perform poorly on new claims with values outside the range of the data used for fitting.

Table 2 below shows the values of SUSPICION for ranges of the provider bill variable. The table indicates that SUSPICION increases rapidly at low bill amounts and then levels off at about $3,000.

Table 2

Suspicion Scores by Provider Bill

Provider Bill	Number of Claims	Mean Suspicion Score
$0	444	0.3
1 - 1,000	376	1.1
1,001 - 2,000	243	3.0
2,001 - 3,000	227	4.2
3,001 - 4,000	60	4.6
4,001 - 5000	33	4.2
5,001 - 6,000	5	5.8
6,001 - 7,000	12	4.3

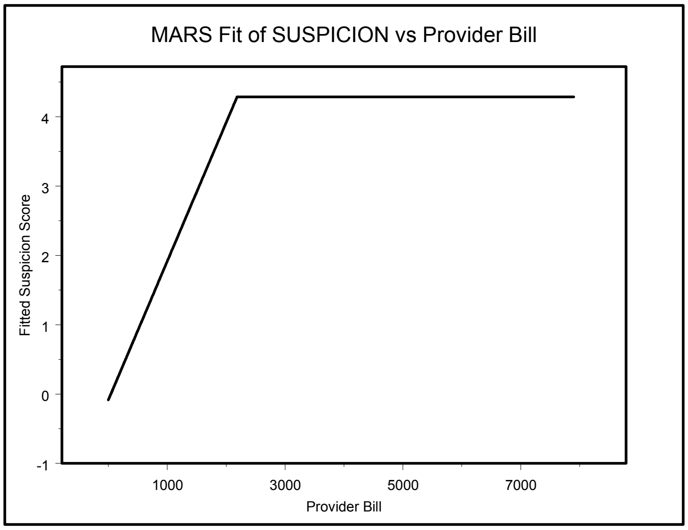
The examples illustrate that traditional techniques which require specific parametric assumptions about the relationship between dependent and independent variables may lack the flexibility to model nonlinear relationships. It should be noted, however, that Francis (Francis, 2001) presented examples where traditional techniques performed as well as neural networks in fitting nonlinear functions. Also, when the true relationship between the dependent and independent variables is linear, classical statistical methods are likely to outperform neural networks.

MARS and Nonlinear Functions

The MARS approach to fitting nonlinear functions has similarities to polynomial regression. In its simplest form MARS fits piecewise linear regressions to the data. That is, MARS breaks the data into ranges and allows the slope of the line to be different for the different ranges. MARS requires the function fit to be continuous, thus there are no jump points between contiguous ranges.

To continue the previous example, a function was fit by MARS. The graph below displays the MARS fitted function. It can be seen that the curve is broken into a steeply sloping line, which then levels off much the way the neural network fitted function did.

Figure 6



MARS uses an optimization procedure that fits the best piecewise regression. Simpler functions may adequately approximate the relationship between predictor and dependent variables and are favored over more complex functions. From the graph, it can be seen that the best MARS regression had two pieces:

1) The curve has a steep slope between bill amounts of $0 and $2,185
2) The curve levels off at bill amounts above $2,185

The fitted regression model can be written as follows:

BF1 = max(0, 2185 – X)
Y = 4.29 - 0.002 * BF1

where

Y is the dependent variable (Suspicion score)
X is the provider bill

The points in the data range where the curves change slope are known as knots. The impact of knots on the model is captured by basis functions. For instance BF1 is a basis function. Basis functions can be viewed as similar to dummy variables in linear regression. Dummy variables are generally used in regression analysis when the predictor variables are categorical. For instance, the Provider bill variable can be converted into a categorical variable by using amount ranges for the categories. We could have the following categories:

Range 1	\$0 - \$2,185	Dummy Variable = 1
Range 2	> \$2,185	Dummy Variable = 0

A dummy variable is a binary indicator variable. It will have a value of 1 when the bill falls within the specified interval for the dummy. Here if the bill is \$1,000, D1 will be 1. When it is \$5,000 D1 will be 0.

A regression with dummy variables has the form:

$Y = B_0 + B_1*D1 + B_2 * D2 + B_3*D3+ \ldots + B_n * Dn$

Since in this simple example there is only one dummy variable, the model is:

$Y = B_0 + B_1*D1$

The constant B_0 captures the effect of the first or base group (greater than \$2185). The dummy variable D1 captures the effect of its bill group relative to the base group. The coefficients for the above model when fitted to the AIB data were:

$Y = 4.28 - 2.89*D1$

This regression function indicates that the mean suspicion score is 4.28 for bills greater than \$2,185 and 1.39 for bills less than \$2,185. However, the use of categorical dummy variables (as opposed to basis functions) creates jumps in the level of the dependent variable, rather than a linear curve, when the range changes.

Basis Functions and Dummy Variables

Each basis function is a combination of a dummy variable with a continuous variable. In the regression function between suspicion score and provider bill:

$BF1 = \max(0, 2185 – X)$

$Y = 4.287 - 0.002 * BF1$

BF1 can be rewritten as:

$BF1 = D1*(2185 - X)$

where D1 is a dummy variable, which takes on the value of 0 if the provider bill is greater than or equal to \$2,185 and 1 if it is less than that value.

Finding the Knots

As mentioned above, a knot is the point in a range at which the slope of the curve changes. Both the number of knots and their placement are unknown at the beginning of the process. A stepwise procedure is used to find the best points to place the spline knots. In its most general form each value of the independent variable is tested as a possible point for placement of a knot. The model initially developed is overfit. A statistical criterion that tests for a significant impact on a goodness of fit measure is used to remove knots. Only those that have a significant impact on the regression are retained. The statistical criterion, generalized cross-validation, will be described later in the paper.

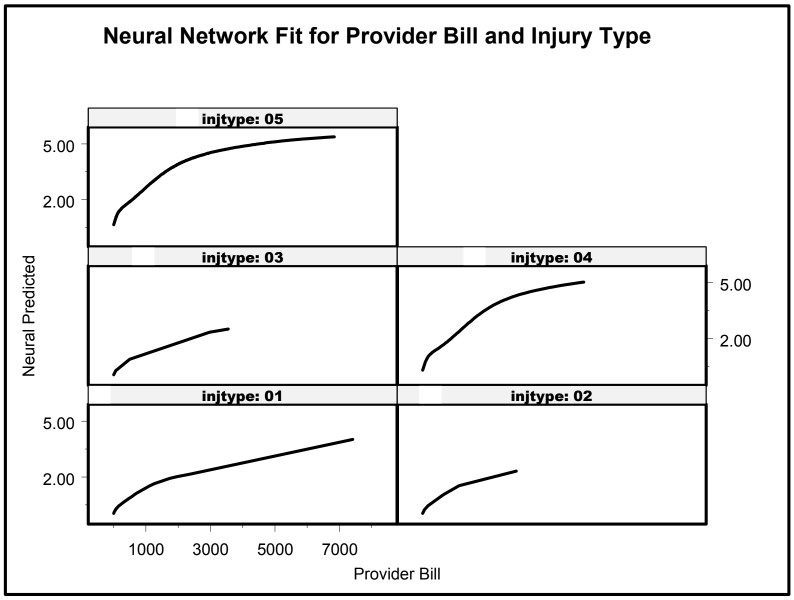
Fitting Smooth Curves

The above discussion describes spline functions which are piecewise linear regressions. For such regressions there is a break in the slope of the curve at each knot. A smooth curve could be created by allowing for higher order terms in the regression, i.e. quadratic or cubic terms could be included. Often, when fitting smoothing splines the curve is a cubic curve. For cubic splines, there is a requirement that the first and second derivatives are continuous at the knot points. For the remainder of this paper we will use piecewise linear splines. Although cubic splines produce smoother curves, they do not, in general, (Steinberg, 1999) significantly improve the fit of the model and are more difficult to parameterize.

Functions with Interaction Terms

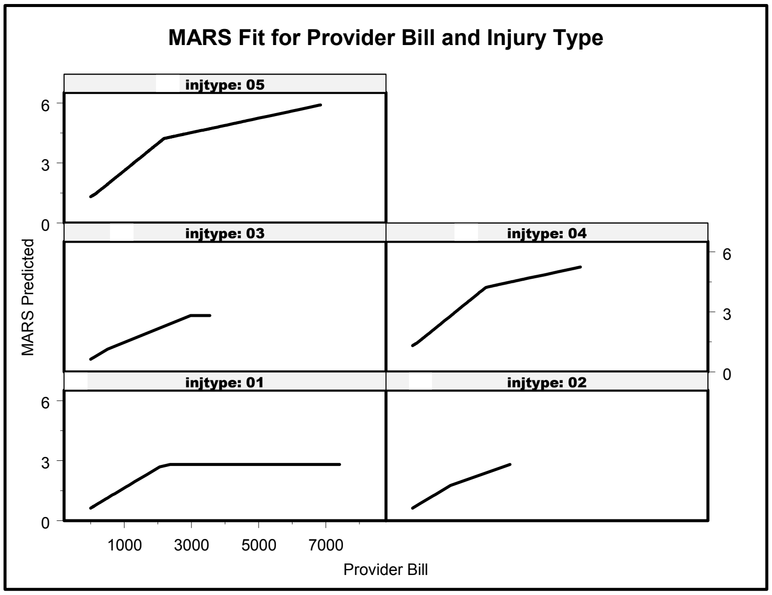
The illustrations shown so far demonstrate MARS's capability for modeling nonlinear relationships. Another complication that occurs when working with insurance data is known as "interactions". The relationship between a predictor variable and the target variable may depend on the value of a second variable. For instance, the relationship between the medical provider bill and the suspicion score may depend on the injury type. This hypothesis is supported by the results of fitting a neural network model for SUSPICION to provider bill and injury type (shown in Figure 7). (For presentation purposes, only some of the injury types are shown). It can be seen that the curves for injury type 4 (neck sprain), and type 5 (back sprain) increase faster than those of the other injury types and ultimately plateau at higher levels.

Figure 7



A MARS curve was fit to the fraud interaction data. The results of the fit are shown below:

Figure 8



It can be seen that, as with the neural network, injury type 4 (neck sprain), and type 5 (back sprain) increase faster and have higher scores than the other injury types. The MARS fitted function was:

BF1 = max(0, 2185 - X)
BF2 = (INJTYPE = 4 OR INJTYPE = 5)
BF3 = max(0, X - 159) * BF2
Y = 2.815 - 0.001 * BF1 + 0.685 * BF2 + .360E-03 * BF3

where
X is the provider bill
INJTYPE is the injury type

There are three basis functions in the model. Basis function BF1 splits the provider bill into the range $0 to $2,185 and greater than $2,185. It is like the first basis function in the previous model of SUSPICION and provider bill. Basis function BF2 is a categorical dummy variable, based on the value of injury type. If the injury type is 4 or 5, it takes on a value of 1, otherwise it is 0. In the model, the coefficient of BF2 is 0.685. Thus, the regression constant value is increased by 0.685 if the injury is a sprain. Basis function BF3 captures the interaction between injury type and provider bill and increases the slope of the curve for sprains.

To create the BF2 basis function, MARS searches all the categories of injury type. By recursive partitioning, or sequential splitting of the categories into two distinct groups, it groups together those categories with a similar effect on the dependent variable into basis functions. When there is more than one categorical variable, the procedure is performed on each one. Only those basis functions with a significant effect on the target variable, as determined by the improvement in the R^2, are included in the final model.

Similarly, an automated search procedure is used to create basis functions that specify interaction effects. Combinations of predictors are tested two at a time for two-way interaction[3]. New basis functions may be created to capture the interaction effect. Thus, a different combination of the injury types than those in BF2 could be associated with the interaction of injury type and provider bill. For this model the injury types were the same for BF2 and BF3.

This example illustrates one advantage of MARS over other data mining techniques such as neural networks. MARS groups together related categories of nominal variables. Many insurance categorical variables have many different levels[4]. For instance, while the injury type variable in the AIB data has only 15 levels, injury type data often has hundreds or even thousands of possible values. Increasingly, the insurance industry is shifting to the use of ICD9[5] codes for injuries. There are in excess of 15,000 possible

[3] Higher order interactions, such as three way and four way interactions are permissible. However, high order interactions are unlikely to be statistically significant in a database of this size.
[4] Note that another data mining technique, Decision Trees (also know as CART) can also group together categories with similar impacts on the dependent variable.
[5] ICD9 codes are the codes used by medical providers and health insurers to classify injuries and illnesses. The definition of the classes is standardized and there is widespread use of these codes.

values for ICD9, many of which are related to similar illnesses or injuries. A procedure that can group together codes with a similar impact on the dependent variable is very handy when so many values are available. The neural network procedure turns each of the possible values of a categorical variable into a binary dummy variable when fitting a model. Many of these categories contain a tiny fraction of the data, thus the parameters fitted to the categories of the categorical variables may be very unstable. Collapsing the categories into a smaller number, with each group having a similar impact on the dependent variable (perhaps when interacting with another variable) significantly reduces the number of parameters in the model.

Missing Data

Missing data occurs frequently when working with large databases. The software commonly used for applying statistical models (including neural networks) typically applies very crude rules when data is missing. Such rules include elimination of records where any value on any variable is missing and substitution of the mean of a variable for the missing value on that variable. More sophisticated methods for addressing missing values, such as data imputation and the expectation maximization (EM) algorithm, have been developed. However, these methods are not widely available in the major statistical software packages. Two significant problems occur with missing data.

1. Because many statistical packages eliminate any record with a missing value on any variable, a lot of the data can be lost to the analysis.
2. In order for the analysis to be valid, the analyst must assume that value of both the dependent and predictor variables is independent of the presence of missing values.

MARS handles missing data by creating a basis function for any variable with missing data. This variable has a value of one when the data is missing on a given variable and zero otherwise. The search procedure can then determine if an interaction between missing data basis functions and other variables in the data are significant in predicting the dependent variable. Thus, other variables can act as surrogates for the missing variable.

Neural networks were not developed with the treatment of missing data in mind. Many neural network software products automatically eliminate from the model any record with a missing value for any variable in the model. Nevertheless there are procedures that can be used to deal with this challenge. One approach is to assign a constant value to data missing on a variable. This value is often the mean for that variable, but this need not be the case. Because neural networks fit nonlinear functions, a value not in the range of the remainder of the data might be assigned to the missing data on a variable, allowing a different relationship between independent and dependent variable for this value than for the remainder of the data. In addition, a dummy variable can be constructed for each of the variables with missing data, and this can be used in the neural network model. Unfortunately, the software available for fitting neural networks does not provide an automated approach to addressing the missing data issue so significant additional programming effort may be required.

To illustrate the handling of missing data, suspicion score is modeled as a function of total provider medical bill and health insurance. The total provider medical bill is the sum of the bills from all providers. Health insurance is a categorical variable with values of yes (claimant has health insurance), no (claimant does not have health insurance) and unknown (missing). The table below shows the distribution of each of these values in the data. The variables in this example were selected because they provided a good illustration of the handling of missing values. That is, the health insurance variable had a significant number of missing cases (see table below) and the total medical bill's influence on the dependent variable is impacted by the presence/absence of missing values on this variable.

Table 3

Health Insurance

Value	Frequency	Percent	Cumulative Percent
No	457	32.6	32.6
Missing	208	14.9	47.5
Yes	735	52.5	100.0
Total	1400	100.0	

The following MARS model was fit:

```
BF1 = max(0, MP_BILL - 2885)
BF2 = max(0, 2885 - MP_BILL )
BF3 = (HEALTHIN ≠  MISSING)
BF4 = (HEALTHIN = MISSING)
BF5 = (HEALTHIN = N)
BF7 = max(0, MP_BILL - 2262) * BF5
BF8 = max(0, 2262 - MP_BILL ) * BF5
BF9 = max(0, MP_BILL - 98) * BF4
BF10 = max(0, 98 - MP_BILL ) * BF4
BF11 = max(0, MP_BILL - 710) * BF3
BF13 = max(0, MP_BILL - 35483)
BF15 = BF3 * BF2

Y = -0.754 - 0.002 * BF1 + 0.967 * BF3 + 1.389 * BF5 - .808E-04 * BF7
        - .624E-03 * BF8 + 0.001 * BF9 + 0.016 * BF10
        + 0.001 * BF11 + .114E-03 * BF13 + .376E-03 * BF15
```

where:
MP_BILL is the total provider medical bill
HEALTHIN is the health insurance variable
BF1 – BF15 are the basis functions
Y is the dependent variable, suspicion score

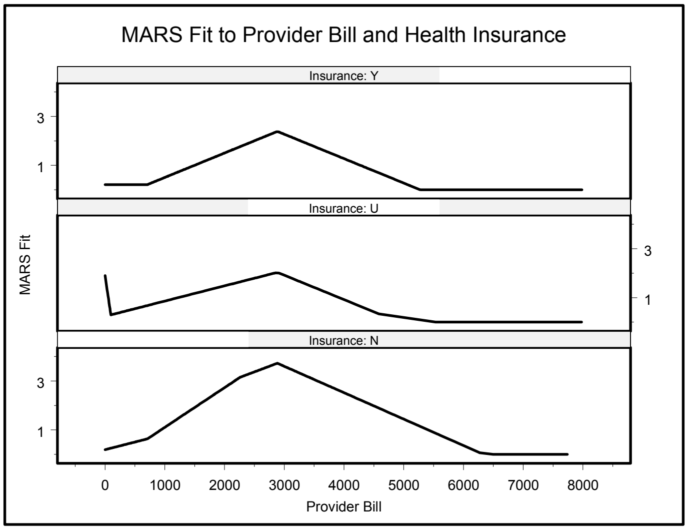
Note that there is no BF6, BF12 or BF14. Variables BF6, BF12 and BF14 were created by MARS, but as they were not found to be significant, they were not included in the final model.

The MARS model created two basis functions for the missing values, one for the presence of missing values and one for the absence. It can be seen that the shape and level of the curve depends on both the value of the health insurance variable and whether it is missing. Basis functions BF3 and BF4 are the dummy variables denoting missing/not-missing values on the health insurance variable. If the health insurance information is missing, BF4 is one. If the information is not missing, BF3 is one. The model indicates that the overall score is raised by .967 if health insurance information is present. Basis functions BF9 and BF10 are the interactions of a missing value on health insurance with provider bill. Basis functions BF11 and BF15 are the interaction of health insurance not missing with total provider bill. Thus, when the provider bill is less than $98 and the health insurance information is missing, the curve's slope is increased by 0.016. This causes the suspicion score to spike at low provider bill values. BF11 indicates that the slope of the curve increases by .001 for values above $710 and BF15 indicates that the slope of the curve increases by 0.00038 up to bill values of $2,885, when health insurance information is present.

Figure 9 displays the curve fit by MARS.[6] The top graph is curve for health insurance (i.e. equal to "yes"), the middle curve is the curve for health insurance unknown (missing) and the bottom graph is the curve for no health insurance. The figure shows that suspicion scores are on average highest when the claimant does not have health insurance and lowest when the information about health insurance is missing. The graphs show that suspicion scores for all categories decline after values of about $3,000.

[6] In the graph, suspicion scores of less than one were censored to have a value of zero.

Figure 9



A neural network was fit to the data using the dummy variable approach described above. That is, a dummy variable was created for the presence or absence of a value on the health insurance variable. Figure 10 shows a comparison of the MARS and the neural network fitted values. The curves fit by the neural network did not vary much over the different values of the health insurance variable. Moreover, for health insurance missing and health insurance equal to 'Y' the neural network scores are above the MARS scores for provider bills greater than about $1,000. In addition, the MARS model suspicion scores decline at high bill amounts, but they do not for the neural network model. Table 4 presents average suspicion scores by bill amount categories for each of the values on the health insurance variable. This table indicates that suspicion scores are higher for claimants with health insurance information, and are highest for claimants with no health insurance. The table also indicates that the suspicion score declines at higher bill amounts, but the decline in the data seems to occur later than the MARS model indicates.

Figure 10

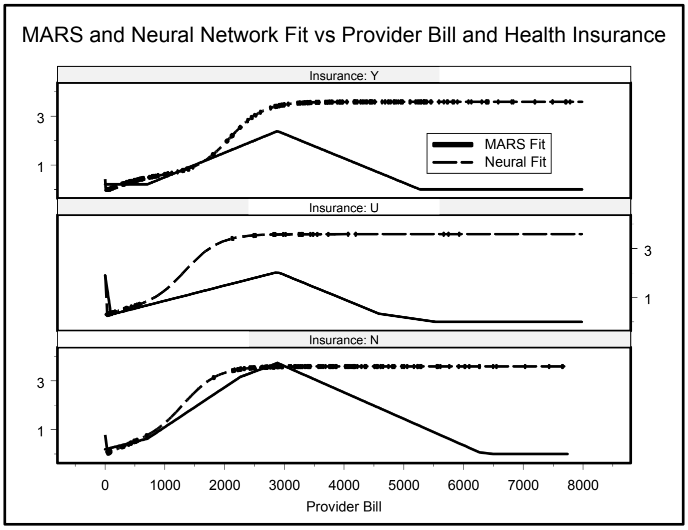


Table 4

		Suspicion Scores by Health Insurance Category		
Total Provider Bill	**Claim Count**	**Y**	**U**	**N**
$0	65	0.3	2.1	0.7
1 - 1,000	532	0.3	0.5	0.4
1,001 - 2,000	140	1.2	3.1	1.9
2,001 - 3,000	268	2.9	3.0	4.5
3,001 - 4,000	149	3.1	2.9	4.2
4,001 - 5000	85	3.4	-	4.8
5,001 - 6,000	54	3.0	2.5	3.4
6,001 - 7,000	25	4.4	-	5.1
7.001 - 8,000	18	2.6	-	4.5
8,001 - 9,000	12	2.8	4.0	-
9,001 - 11,000	13	3.1	-	2.7
> 11,000	39	1.0	-	2.5
Total	1,400	1.6	1.5	2.8

Both the MARS model and the neural network model had similar R^2 (around 0.37). The MARS software uses a statistical procedure to assess the significance of variables and rank them in order of importance. This procedure is described in a later section of this paper. The MARS procedure found the health insurance variable to be significant, but much less significant than the provider bill variable. By visual inspection, it appears that the neural network procedure found no meaningful difference in suspicion score by health insurance category. A more formal neural network procedure for assessing the importance of variables will be discussed in the next section of the paper.

A simple procedure for comparing the accuracy of two models is to hold out a portion of the data for testing. Data is separated into training and test data. The model is fit using the training data and its accuracy is tested using the test data to determine how well the dependent variable was predicted on data not used for fitting. This test is relatively straightforward to perform. In the next section of the paper a more computationally intensive procedure will be presented.

To compare the neural network and MARS models, two thirds of the data was used for fitting and one third was used for testing. The neural network had an R^2 of 0.30 compared to 0.33 for the MARS model. The performance of the two models was also tested on subsets of the data containing only one value of the health insurance variable (i.e., health insurance missing, health insurance equal to yes and health insurance equal to no). MARS outperformed the neural network model on health insurance missing (R^2 = .26 versus R^2 = 0) and health insurance equal to no (R^2 = .31 versus R^2= .25). The neural network outperformed MARS on health insurance equal to yes (R^2= .43 versus R^2 = .32).

This example suggests that MARS more accurately modeled the effect of the health insurance variable and the effect of a missing value for this variable on the dependent variable than did the neural network model. However, it would be desirable to assess the significance of the differences in the accuracy of the overall fit.

The square root of R^2 is the correlation coefficient, which can be used in a test of significance. The distribution of a transform of the correlation coefficient can be approximated by a normal distribution[7]:

$$Z = \frac{1}{2}\ln\frac{1+r}{1-r}$$

$$\mu_z = Z,$$

$$\sigma_z = \frac{1}{n-3}$$

where r is the correlation coefficient and n is the sample size.

The normal approximation was used to compute confidence intervals for each of the correlations. As shown in Table 5, the 95% confidence intervals around the Z statistic computed from the two correlations overlapped, suggesting that the difference between the fits of the two models is not statistically significant.

Table 5

Confidence Intervals for Correlation Coefficient

Model	R^2	r	Z	sd	Lower 95% CI	Upper 95% CI
MARS	0.33	0.57	0.65	0.05	0.56	0.74
Neural Network	0.30	0.55	0.62	0.05	0.52	0.71

This example illustrates one of the great strengths of MARS: its automated procedures for handling missing data. While missing data was not a major issue with the AIB database, as most of the variables were fully populated, it is a common problem with most insurance databases. One possible use for MARS is to create basis functions for variables having missing values. These basis functions could then be used by other procedures such as neural networks.

A More Complex Model

The models presented thus far have been relatively simple one and two variable models. In this section of the paper, the results of a more complex model will be presented. The variables used in the model are described below.

This section will present an example where MARS and neural networks are used for classification. The dependent variable for this model is ASSESS, the expert's assessment of the likelihood that the claim is a fraud or abuse claim. This variable was converted to a binary dependent variable. The two categories were the value 1 (probably legitimate) versus 2 through 5 (the various kinds of suspected fraud or abuse). Thus, if a claim is other than probably legitimate, it is treated as a suspected abuse claim.

[7] This formula is from Miller and Wichern (Miller and Wichern, 1977, pp. 213 – 214).

MARS can perform regressions on binary variables. When the dependent variable is binary, MARS is run in binary mode. In binary mode, the dependent variable is converted into a 0 (legitimate) or a 1 (suspected fraud or abuse). Ordinary least squares regression is then performed regressing the binary variable on the predictor variables. Logistic regression is a more common procedure when the dependent variable is binary. Suppose that the true target variable is the probability that a given claim is abusive, and this probability is denoted *p(**x**)*. The model relating *p(**x**)* to the a vector of independent variables **x** is:

$$\ln(\frac{p}{1-p};\mathbf{x}) = B_0 + B_1X_1 + \ldots + B_nX_n$$

where the quantity ln(*p*(**x**)*/(1-p*(***x***))) is known as the logit function or log odds. Logistic regression can be used to produce scores that are between zero and one, consistent with viewing the score as a probability. Binary regressions can produce predicted values which can be less than zero and greater than one. One solution to this issue is to truncate the predicted values at zero and one. Another solution is to add the extra step of fitting a logistic regression to the data using the MARS predicted value as the independent variable and the binary assessment variable as the dependent variable. The fitted probabilities from the logistic regression can then be assigned as a score for the claim. The neural network model was also run in binary mode and also produced fitted values which were less than zero or greater than one. In this analysis, logistic regression was applied to the results of both the MARS and neural network fits to convert the predicted values into probabilities.

Variables in the Model

There are two categories of predictor variables that were incorporated into the models described in this section. The first category is red flag variables. These are primarily subjective variables that are intended to capture features of the accident, injury or claimant that are believed to be predictive of fraud or abuse. Many red flag variables represent accumulated industry wisdom about which indicators are likely to be associated with fraud or abuse. The information recorded in these variables represents an expert's subjective assessment of fraud indications, such as "the insured felt set up, denied fault". These variables are binary, that is, they are either true or false. Such red flag variables are often used to target certain claims for further investigation. The data for these red flag variables is not part of the claim file; it was collected as part of the special effort undertaken in assembling the AIB database for fraud research.

The red flag variables were supplemented with claim file variables deemed to be available early in the life of a claim and therefore of practical value in predicting fraud and abuse.

The variables selected for use in the full model are the same as those used by Viaene *et al*. (Viaene *et. al*., 2002) in their comparison of statistical and data mining methods. While a much larger number of predictor variables is available in the AIB data for modeling fraud, the red flag and objective claim variables selected for incorporation into their models by Viaene *et al*. were chosen because of early availability. Therefore they are likely to be useful in predicting fraud and abuse soon enough in the claim's lifespan for effective mitigation efforts to lower the cost of the claim. Tables 6 and 7 present the red flag and claim file variables.

Table 6

Red Flag Variables

Subject	**Indicator Variable**	**Description**
Accident	ACC01	No report by police officer at scene
	ACC04	Single vehicle accident
	ACC09	No plausible explanation for accident
	ACC10	Claimant in old, low valued vehicle
	ACC11	Rental vehicle involved in accident
	ACC14	Property Damage was inconsistent with accident
	ACC15	Very minor impact collision
	ACC16	Claimant vehicle stopped short
	ACC19	Insured felt set up, denied fault
Claimant	CLT02	Had a history of previous claims
	CLT04	Was an out of state accident
	CLT07	Was one of three or more claimants in vehicle
Injury	INJ01	Injury consisted of strain or sprain only
	INJ02	No objective evidence of injury
	INJ03	Police report showed no injury or pain
	INJ05	No emergency treatment was given
	INJ06	Non-emergency treatment was delayed
	INJ11	Unusual injury for auto accident
Insured	INS01	Had history of previous claims
	INS03	Readily accepted fault for accident
	INS06	Was difficult to contact/uncooperative
	INS07	Accident occurred soon after effective date
Lost Wages	LW01	Claimant worked for self or a family member
	LW03	Claimant recently started employment

Table 7

Claim Variables Available Early in Life of Claim

Variable	**Description**
AGE	Age of claimant
POLLAG	Lag from policy inception to date of accident[8]
RPTLAG	Lag from date of accident to date reported
TREATLAG	Lag from date of accident to earliest treatment by service provider
AMBUL	Ambulance charges
PARTDIS	The claimant partially disabled
TOTDIS	The claimant totally disabled
LEGALREP	The claimant represented by an attorney

[8] POLLAG, RPTLAG and TRTLAG are continuous variables.

One of the objectives of this research is to investigate which variables are likely to be of value in predicting fraud and abuse. To do this, procedures are needed for evaluating the importance of variables in predicting the target variable. Below, we present some methods that can be used to evaluate the importance of the variables.

Evaluating Variable Importance

A procedure that can be used to evaluate the quality of the fit when fitting complex models is generalized cross-validation (GCV). This procedure can be used to determine which variables to keep in the model, as they produce the best fit, and which to eliminate. Generalized cross-validation can be viewed as an approximation to cross-validation, a more computationally intensive goodness of fit test described later in this paper.

$$GCV = \frac{1}{N}\sum_{i=1}^{N}[\frac{y_i - \hat{f}(x_i)}{1-k/N}]^2$$

where N is the number of observations
y is the dependent variable
x is the independent variable(s)
k is the effective number of parameters or degrees of freedom in the model.

The effective degrees of freedom is the means by which the GCV error functions puts a penalty on adding variables to the model. The effective degrees of freedom is chosen by the modeler. Since MARS tests many possible variables and possible basis functions, the effective degrees of freedom used in parameterizing the model is much higher than the actual number of basis function in the final model. Steinberg states that research indicates that k should be two to five times the number of basis functions in the model, although some research suggests it should be even higher (Steinberg, 2000).

The GCV can be used to rank the variables in importance. To rank the variables in importance, the GCV is computed with and without each variable in the model.

For neural networks, a statistic known as the sensitivity can be used to assess the relative importance of variables. 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. Potts (Potts, 2000) and Francis (Francis, 2001) described a procedure for computing this statistic. Many of the major data mining packages used for fitting neural networks supply this statistic or a ranking of variables based on the statistic. Statistical procedures for testing the significance of variables are not well developed for neural networks. One approach is to drop the least important variables from the model, one at a time and evaluate whether the fit deteriorates on a sample of claims that have been held out for testing. On a large database this approach can be time consuming and inefficient, but it is feasible on small databases such as the AIB database.

Table 8 displays the ranking of variable importance from the MARS model. Table 9 displays the ranking of importance from the neural network model. The final model fitted by MARS uses only the top 12 variables in importance. These were the variables that were determined to have made a significant contribution to the final model. Only variables included in the model, i.e., found to be significant are included in the tables.

Table 8

MARS Ranking of Variables

Rank	Variable	Description
1	LEGALREP	Legal Representation
2	TRTMIS	Treatment lag missing
3	ACC04	Single vehicle accident
4	INJ01	Injury consisted of strain or sprain only
5	AGE	Claimant age
6	PARTDIS	Claimant partially disabled
7	ACC14	Property damage was inconsistent with accident
8	CLT02	Had a history of previous claims
9	POLLAG	Policy lag
10	RPTLAG	Report lag
11	AMBUL	Ambulance charges
12	ACC15	Very minor impact collision

The ranking of variables as determined by applying the sensitivity test to the neural network model is shown below.

Table 9

Neural Network Ranking of Variables

Rank	Variable	Description
1	LEGALREP	Legal Representation
2	TRTMIS	Treatment lag missing
3	AMBUL	Ambulance charges
4	AGE	Claimant age
5	PARTDIS	Claimant partially disabled
6	RPTLAG	Report lag
7	ACC04	Single vehicle accident
8	POLLAG	Policy lag
9	CLT02	Had a history of previous claims
10	INJ01	Injury consisted of strain or sprain only
11	ACC01	No report by police officer at scene
12	ACC14	Property damage was inconsistent with accident

Both the MARS and the neural network find the involvement of a lawyer to be the most important variable in predicting fraud and abuse. Both procedures also rank as second a missing value on treatment lag. The value on this variable is missing when the claimant has not been to an outpatient health care provider, although in over 95% of these cases, the claimant has visited an emergency room.[9] Note that both medical paid and total paid for this group is less than one third of the medical paid and total paid for claimants who visited a provider. Thus the TRTMIS (treatment lag missing) variable appears to be a surrogate for not using an outpatient provider. The actual lag in obtaining treatment is not an important variable in either the MARS or neural network models.

Explaining the Model

Below are the formulas for the model fit by MARS. Again note that some basis functions created by MARS were found not to be significant and are not shown. To assist with interpretation, Table 10 displays a description of the values of some of the variables in the model.

```
 BF1 = (LEGALREP = 1)
 BF2 = (LEGALREP = 2)
 BF3 = ( TRTLAG = missing)
 BF4 = ( TRTLAG ≠ missing)
 BF5 = ( INJ01 = 1) * BF2
 BF7 = ( ACC04 = 1) * BF4
 BF9 = ( ACC14 = 1)
 BF11 = ( PARTDIS = 1) * BF4
 BF15 = max(0, AGE - 36) * BF4
 BF16 = max(0, 36 - AGE ) * BF4
 BF18 = max(0, 55 - AMBUL ) * BF15
 BF20 = max(0, 10 - RPTLAG ) * BF4
 BF21 = ( CLT02 = 1)
 BF23 = POLLAG * BF21
 BF24 = ( ACC15 = 1) * BF16

 Y = 0.580 - 0.174 * BF1 - 0.414 * BF3 + 0.196 * BF5 - 0.234 * BF7
+ 0.455 * BF9 + 0.131 * BF11 - 0.011 * BF15 - 0.006 * BF16 +
.135E-03 * BF18 - 0.013 * BF20 + .286E-03 * BF23 + 0.010 * BF24
```

[9] Because of the strong relationship between a missing value on treatment lag and the dependent variable, and the high percentage of claims in this category which had emergency room visits, an indicator variable for emergency room visits was tested as a surrogate. It was found not to be significant.

Table 10

Description of Categorical Variables

Variable	Value	Description
LEGALREP	1	No legal representation
	2	Has legal representation
INJ01	1	Injury consisted of strain or sprain only
	2	Injury did not consist of strain or sprain only
ACC04	1	Single vehicle accident
	2	Two or more vehicle accident
ACC14	1	Property damage was inconsistent with accident
	2	Property damage was consistent with accident
PARTDIS	1	Partially disabled
	2	Not partially disabled
CLT02	1	Had a history of previous claims
	2	No history of previous claims
ACC15	1	Was very minor impact collision
	2	Was not very minor impact collision

The basis functions and regression produced by MARS assist the analyst in understanding the impact of the predictor variables on the dependent variable. From the formulae above, it can be concluded that

1) when a lawyer is not involved (LEGALREP = 1), the probability of fraud or abuse declines by about 0.17
2) when the claimant has legal representation and the injury is consistent with a sprain or strain only, the probability of fraud or abuse increases by 0.2
3) when the claimant does not receive treatment from an outpatient health care provider (TRTLAG = missing), the probability of abuse declines by 0.41
4) a single vehicle accident where the claimant receives treatment from an outpatient health care provider (treatment lag not missing) decreases the probability of fraud by 0.23
5) if property damage is inconsistent with the accident, the probability of fraud or abuse increases by 0.46
6) if the claimant is partially disabled and receives treatment from an outpatient health care provider the probably of fraud or abuse is increased by 0.13

Of the red flag variables, small contributions were made by the claimant having a previous history of a claim[10] and the accident being a minor impact collision. Of the objective continuous variables obtained from the claim file, variables such as claimant age, report lag and policy lag have a small impact on predicting fraud or abuse.

Figures 11 and 12 display how MARS modeled the impact of selected continuous variables on the probability of fraud and abuse. For claims receiving outpatient health

[10] This variable only captures history of a prior claim if it was recorded by the insurance company. For some companies participating in the study, it was not recorded.

care, report lag has a positive impact on the probability of abuse, but its impact reaches its maximum value at about 10 days. Note the interaction between claimant age and ambulance costs displayed in Figure 12. For low ambulance costs, the probability of abuse rises steeply with claimant age and maintains a relatively high probability except for the very young and very old claimants. As ambulance costs increase, the probability of fraud or abuse decreases, and the decrease is more pronounced at lower and higher ages. Ambulance cost appears to be acting as a surrogate for injury severity.

Figure 11

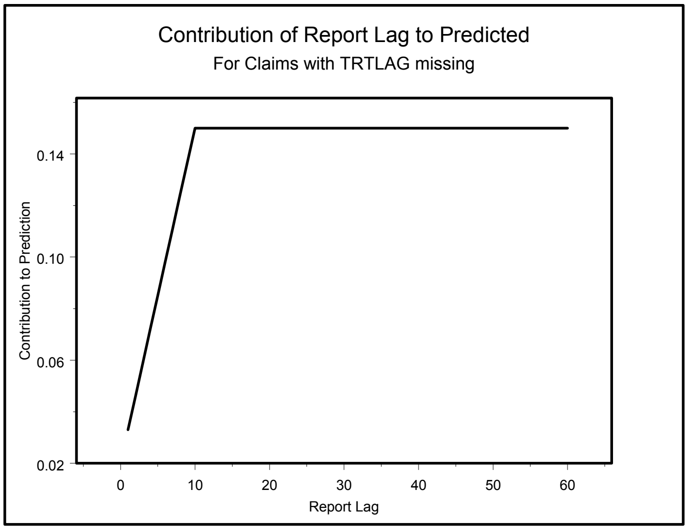
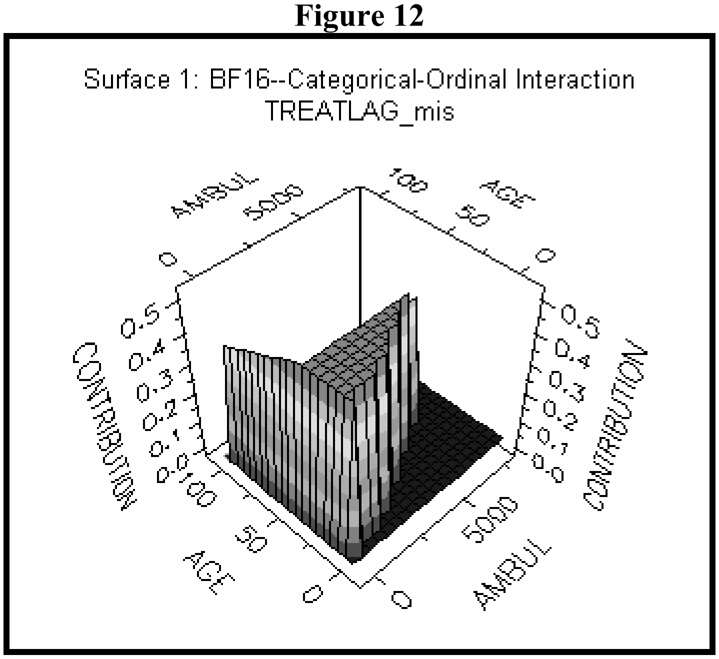


Figure 12



This section on explaining the model illustrates one of the very useful qualities of MARS as compared to neural networks: the output of the model is a formula which describes the relationships between predictor and dependent variables and which can be used to explain the model to management. To some extent, the sensitivity measure assists us in understanding the relationships fit by the neural network model, as it provides a way to assess the importance of each of the variables to the prediction. However, the actual functional relationships between independent and dependent variables are not typically available and the model can be difficult to explain to management.[11]

Evaluating the Goodness of the Fit and Comparing the Accuracy
One approach for testing the accuracy of models that is commonly used in data mining applications is to have separate training and testing samples. This approach was used in the previous example. Typically one half to one third of the data is held out for testing. However, when the database used for modeling is small, the analyst may not want to lose a large portion of the data to testing. Moreover, as the testing is performed on a relatively small sample, the goodness of fit results may be sensitive to random variation in the subsets selected for training and testing. An alternative procedure that allows more of the data to be used for fitting and testing is cross-validation. Cross-validation involves iteratively holding out part of the sample, fitting the model to the remainder of the sample and testing the goodness of the fitted model on the held out portion. For instance, the sample may be divided into 4 groups. Three of the groups are used to fit the model and one is used for testing. The process is repeated four times, and the goodness of fit statistics for the four test samples are averaged. As the AIB database is relatively small for a data mining application, this is the procedure used. Testing was performed using four fold cross-validation.

Both a MARS model and a neural network model were fit to four samples of the data. Each time the fitted model was used to predict the probability of fraud or abuse for one quarter of the data that was held out. The predictions from the four test samples were then combined to allow comparison of the MARS and neural network procedures.

Table 11 presents some results of the analysis. This table presents the R^2 of the regression of ASSESS on the predicted value from the model. The table shows that the neural network R^2 was higher than that of MARS. The table also displays the percentage of observations whose values were correctly predicted by the model. The predictions are based only on the samples of test claims. The neural network model correctly predicted 79% of the test claims, while MARS correctly predicted 77% of the test claims.

Table 11

Four Fold Cross-validation

Technique	**R^2**	**Percent Correct**
MARS	0.35	0.77
Neural Network	0.39	0.79

[11] Plate (2000) and Francis (2001) present a method to visualize the relationships between independent and dependent variables, The technique is not usually available in data mining software.

Tables 12 and 13 display the accuracy of MARS and the neural network in classifying fraud and abuse claims.[12] A cutoff point of 50% was used for the classification. That is, if the model's predicted probability of a 1 on ASSESS exceeded 50%, the claim was deemed an abuse claim. Thus, those claims in cell Actual =1 and Predicted=1 are the claims assessed by experts as probably abusive which were predicted to be abusive. Those claims in cell Actual=1, Predicted =0, are the claims assessed as probable abuse claims which were predicted by the model to be legitimate.

Table 12

MARS Predicted * Actual

Predicted	**Actual**		
	0	1	Total
0	738	160	898
1	157	344	501
Total	895	505	

Table 13

Neural Network Predicted * Actual

Predicted	**Actual**		
	0	1	Total
0	746	127	873
1	149	377	526
Total	895	505	

Table 14 presents the sensitivity and specificity of each of the models. The sensitivity is the percentage of events (in this case suspected abuse claims) that were predicted to be events. The specificity is the percentage of nonevents (in this case claims believed to be legitimate) that were predicted to be nonevents. Both of these statistics should be high for a good model. The table indicates that both the MARS and neural network models were more accurate in predicting nonevent or legitimate claims. The neural network model had a higher sensitivity than the MARS model, but both were approximately equal in their specificities. The neural network's higher overall accuracy appears to be a result of its greater accuracy in predicting the suspected fraud and abuse claims. Note that the sensitivity and specificity measures are dependent on the choice of a cutoff value. Thus, if a cutoff lower than 50% were selected, more abuse claims would be accurately predicted and fewer legitimate claims would be accurately predicted.

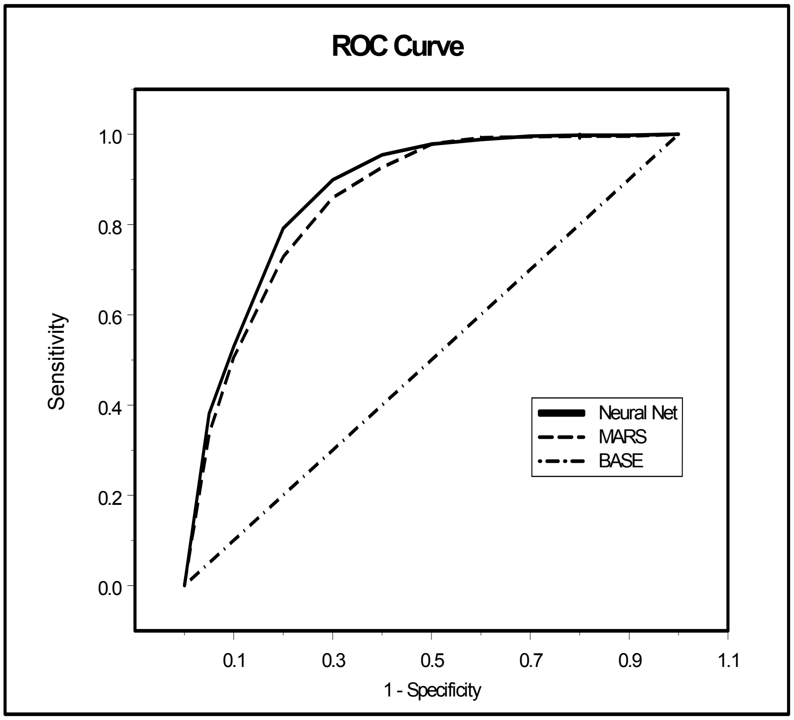
Table 14

Model	**Sensitivity**	**Specificity**
MARS	68.3	82.5
Neural Network	74.8	83.4

[12] These tables are often referred to as confusion matrices

A common procedure for visualizing the accuracy of models used for classification is the receiver operating characteristics (ROC) curve. This is a curve of sensitivity versus specificity (or more accurately 1.0 minus the specificity) over a range of cutoff points. 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. 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", the more accurate the model. It can be seen from the graph of the ROC curve that both the MARS and neural network models have significant "lift" but the neural network model has more "lift" than the MARS model.

Figure 13



A statistic that summarizes the predictive accuracy of a model as measured by an ROC curve is the area under the ROC curve (AUROC). A curve that rises quickly has more area under the ROC curve. Table 15 displays the AUROC for both models, along with their standard deviations and 95% confidence intervals. As the lower bound of the confidence interval for the neural network is below the higher bound of the confidence interval for MARS, it can be concluded that differences between the MARS model and the neural network model are not statistically significant.

Table 15

Statistics for Area Under the ROC Curve

Test Result Variables	**Area**	**Std Error**	**Asymptotic Sig**	**Lower 95% Bound**	**Upper 95% Bound**
MARS Probability	0.85	0.01	0.000	0.834	0.873
Neural Probability	0.88	0.01	0.000	0.857	0.893

Summary of Comparison

The ROC curve results suggest that in this analysis the neural network enjoyed a modest though not statistically significant advantage over MARS in predictive accuracy. It should be noted that the database used for this study was quite small for a data mining application and may produce results that do not generalize to larger applications. Steinberg (Steinberg, 2001) reports that on other applications MARS equaled or exceeded the performance of neural networks. It should also be noted that some of the key comparative strengths of MARS such as its ability to handle missing data were not a significant factor in the analysis, as all but one of the variables were fully populated. [13] In addition, MARS's capability of clustering levels of categorical variables together was not relevant to this analysis, as no categorical variable had more than two levels.

A practical advantage that MARS enjoys over neural networks is the ease with which results can be explained to management. Thus, one potential use for MARS is to fit a model using neural networks and then apply MARS to the fitted values to understand the functional relationships fitted by the neural network model. The results of such an exercise are shown below:

BF1 = (LEGALREP = 1)
BF2 = (LEGALREP = 2)
BF3 = (TRTLAG ≠ missing)
BF4 = (TRTLAG = missing)
BF5 = (INJ01 = 1)
BF7 = (ACC04 = 1) * BF3
BF8 = (ACC04 = 2) * BF3
BF9 = (PARTDIS = 1) * BF8
BF11 = max(0, AMBUL - 182) * BF2
BF12 = max(0, 182 - AMBUL) * BF2
BF13 = (ACC14 = 1) * BF3
BF15 = (CLT02 = 1) * BF3
BF17 = max(0, POLLAG - 21) * BF3
BF19 = max(0, AGE - 41) * BF3
BF20 = max(0, 41 - AGE) * BF3

[13] One of the claims was missing data on the AGE variable, and this claim was eliminated from the neural network analysis and from comparisons of MARS the neural network model. Had more claims been missing the AGE variable, we would have modeled it in the neural network.

BF21 = (INS06 = 1)
BF23 = max(0, RPTLAG - 24) * BF8
BF24 = max(0, 24 - RPTLAG) * BF8
BF25 = BF1 * BF4
BF27 = (ACC15 = 1) * BF8
BF29 = (INJ03 = 1) * BF2

Y = 0.098 - 0.272 * BF1 + 0.334 * BF3 + 0.123 * BF5 - 0.205 * BF7 + 0.145 * BF9 - .623E-04 * BF11 + .455E-03 * BF12 + 0.258 * BF13 + 0.100 * BF15 + .364E-03 * BF17- 0.004 * BF19 - 0.001 * BF20 + 0.152 * BF21 + .945E-03 * BF23 - 0.002 * BF24 + 0.135 * BF25 + 0.076 * BF27 - 0.073 * BF29

This model had an R^2 of 0.9. Thus, it was able to explain most of the variability in the neural network fitted model. Though the sensitivity test revealed that LEGALREP is the most significant variable in the neural network model, its functional relationship to the probability of fraud is unknown using standard neural network modeling techniques. As interpreted by MARS, the absence of legal representation reduces the probability of fraud by 0.272., even without interacting with other variables. LEGALREP also interacts with the ambulance cost variable, INJ03 (police report shows no injury) and no use of a health care provider (treatment lag missing). The sensitivity measure indicated that the presence or absence of a value for treatment lag was the second most important variable. As stated earlier, this variable can be viewed as a surrogate for use of an outpatient health care provider. The use of an outpatient health care provider (TRTLAG ≠ missing) adds 0.334 to the probability of fraud or abuse, but this variable also interacts with the policy lag, report lag, claimant age, partial disability, ACC04, (single vehicle accident), ACC14 (property damage inconsistent with accident) and CLT02 (history of prior claims).

The MARS model helps the user understand not only the nonlinear relationships uncovered by the neural network model, but also describes the interactions which were fit by the neural network.

A procedure frequently used by data mining practitioners when two or more approaches are considered appropriate for an application is to construct a hybrid model or average the results of the modeling procedures. This approach has been reported to reduce the variance of the prediction (Salford Systems, 1999). Table 16 displays the AUROC statistics resulting from averaging the results of the MARS and neural network models. The table indicates that the performance of the hybrid model is about equal to the performance of the neural network. (The graph including the ROC curve for the combined model is not shown, as the curve is identical to Figure 13 because the neural network and combined curves cannot be distinguished.) Salford Systems (Salford Systems, 1999) reports that the accuracy of hybrid models often exceeds that of its components, but usually at least equals that of the best model. Thus, hybrid models that combine the results of two techniques may be preferred to single technique models because uncertainty about the accuracy of the predicted values on non-sample data is reduced.

Table 16

Statistics for Area Under the ROC Curve					
Test Result Variables	Area	Std Error	Asymptotic Sig	Lower 95% Bound	Upper 95% Bound
MARS Probability	0.853	0.01	0.000	0.834	0.873
Neural Probability	0.875	0.01	0.000	0.857	0.893
Combined Probability	0.874	0.01	0.000	0.857	0.892

Using Model Results

The examples in this paper have been used to explain the MARS technique and compare it to neural networks. The final example in this paper has been a fraud and abuse application that used information about the PIP claim that would typically be available shortly after the claim is reported to predict the likelihood that the claim is abusive or fraudulent. The results suggest that a small number of variables, say about a dozen, are effective in predicting fraud and abuse. Among the key variables in importance for both the neural network model and MARS are use of legal representation, use of an outpatient health care provider (as proxied by TRTLAG missing) and involvement in a single vehicle accident. Due to the importance of legal representation, it would appear useful for insurance companies to record information about legal representation in computer systems, as not all companies have this data available.

The results of both the MARS and neural network analysis suggest that both claim file variables (present in most claims databases) and red flag variables (common wisdom about which variables are associated with fraud) are useful predictors of fraud and abuse. However, this and other studies support the value of using analytical tools for identifying potentially abusive claims. As pointed out by Derrig (Derrig, 2002), fraud models can help insurers sort claims into categories related to the need for additional resources to settle the claim efficiently. For instance, claims assigned a low score by a fraud and abuse model, can be settled quickly with little investigative effort on the part of adjusters. Insurers may apply increasingly greater resources to claims with higher scores to acquire additional information about the claimant/policyholder/provider and mitigate the total cost of the claim. Thus, the use of a fraud model is not conceived as an all or nothing exercise that classifies a claim as fraudulent or legitimate, but a graduated effort of applying increasing resources to claims where there appears to be a higher likelihood of material financial benefit from the expenditures.

Conclusion

This paper has introduced the MARS technique and compared it to neural networks. Each technique has advantages and disadvantages and the needs of a particular application will determine which technique is most appropriate.

One of the strengths of neural networks is their ability to model highly nonlinear data. MARS was shown to produce results similar to neural networks in modeling a nonlinear function. MARS was also shown to be effective at modeling interactions, another strength of neural networks.

In dealing with nominal level variables, MARS is able to cluster together the categories of the variables that have similar effects on the dependent variable. This is a capability not possessed by neural networks that is extremely useful when the data contain categorical variables with many levels such as ICD9 code.

MARS has automated capabilities for handling missing data, a common feature of large databases. Though missing data can be modeled with neural networks using indicator variables, automated procedures for creating such variables are not available in most standard commercial software for fitting neural networks. Moreover, since MARS can create interaction variables from missing variable basis functions and other variables, it can create surrogates for the missing variables. Thus, on applications using data with missing values on many variables, or data where the categorical variables have many values, one may want to at least preprocess the data with MARS to create basis functions for the missing data and categorical variables which can be used in other procedures.

A significant disadvantage of neural networks is that they are a “black box”. The functions fit by neural networks are difficult for the analyst to understand and difficult to explain to management. One of the very useful features of MARS is that it produces a regression like function that can be used to understand and explain the model; therefore it may be preferred to neural networks when ease of explanation rather than predictive accuracy is required. MARS can also be used to understand the relationships fit by other models. In one example in this paper MARS was applied to the values fit by a neural network to uncover the important functional relationships modeled by the neural network.

Neural networks are often selected for applications because of their predictive accuracy. In a fraud modeling application examined in this paper the neural network outperformed MARS, though the results were not statistically significant. The results were obtained on a relatively small database and may not generalize to other databases. In addition, the work of other researchers suggests that MARS performs well compared to neural networks. However, neural networks are highly regarded for their predictive capabilities. When predictive accuracy is a key concern, the analyst may choose neural networks rather than MARS when neural networks significantly outperform MARS. An alternative approach that has been shown to improve predictive accuracy is to combine the results of two techniques, such as MARS and neural networks, into a hybrid model.

This analysis and those of other researchers supports the use of intelligent techniques for modeling fraud and abuse. The use of an analytical approach can improve the performance of fraud detection procedures that utilize red flag variables or subjective claim department rules by 1) determining which variables are really important in predicting fraud, 2) assigning an appropriate weight to the variables when using them to predict fraud or abuse, and 3) using the claim file and red flag variables in a consistent manner across adjusters and claims.

References

Allison, Paul. *Missing Data*, Sage Publications, 2001

Berry, Michael J. A., and Linoff, Gordon, *Data Mining Techniques*, John Wiley and Sons, 1997

Brockett, Patrick L., Xiaohua Xia and Richard A. Derrig, 1998, "Using Kohonen's Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud", *Journal of Risk and Insurance*, June, 65:2.

Brockett, Patrick L., Richard A. Derrig, Linda L. Golden, Arnold Levine and Mark Alpert, "Fraud Classification Using Principal Component Analysis of RIDITs", *Journal of Risk and Insurance*, September, 2002, pp. 341-371.

Dhar, Vasant and Stein, Roger, *Seven Methods for Transforming Corporate Data Into Business Intelligence*, Prentice Hall, 1997

Derrig, Richard A., Herbert I. Weisberg and Xiu Chen, 1994, "Behavioral Factors and Lotteries Under No-Fault with a Monetary Threshold: A Study of Massachusetts Automobile Claims", *Journal of Risk and Insurance*, June, 1994, 61:2: 245-275.

Derrig, Richard A., and Krzysztof M. Ostaszewski, "Fuzzy Techniques of Pattern Recognition in Risk and Claim Classification", *Journal of Risk and Insurance*, September, 1995, 62:3: 447-482.

Derrig, Richard, "Patterns, Fighting Fraud With Data", *Contingencies*, Sept/Oct, 1999, pp. 40–49.

Derrig, Richard A., and Valerie Zicko, "Prosecuting Insurance Fraud: A Case Study of The Massachusetts Experience in the 1990s", Risk Management and Insurance Research, 2002.

Derrig, Richard, "Insurance Fraud", *Journal of Risk and Insurance*, September, 2002, pp. 271-287

Francis, Louise, "Neural Networks Demystified", *Casualty Actuarial Society Forum,* Winter 2001*, pp 253 - 320.*

Freedman, Roy S., Klein, Robert A. and Lederman, Jess, *Artificial Intelligence in the Capital Markets*, Probus Publishers 1995

Hastie, Trevor, Tibshirani, Robert, *Generalized Additive Models*, Chapman and Hall, 1990

Hastie, Trevor, Tibshirani, Robert and Freidman, Jerome, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, 2001

Hayward, Gregory, "Mining Insurance Data to Promote Traffic Safety and Better Match Rates to Risk", *Casualty Actuarial Society Forum,* Winter 2002*, pp. 31 – 56.*

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

Keefer, James, "Finding Causal Relationships By Combining Knowledge and Data in Data Mining Applications", Paper presented at Seminar on Data Mining, University of Delaware, April, 2000.

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

Little, Roderick and Rubin, Donald, *Statistical Analysis with Missing Data*, John Wiley and Sons, 1987

Marsh, Lawrence and Cormier, David, *Spline Regression Models*, Sage Publications, 2002

Martin, E. B. and Morris A. J., "Artificial Neural Networks and Multivariate Statistics", in *Statistics and Neural Networks: Advances at the Interface,* Oxford University Press, 1999, pp. 195 – 292

Miller, Robert and Wichern, Dean, *Intermediate Business Statistics*, Holt, Reinhart and Winston, 1977

Neal, Bradford, *Bayesian Learning for Neural Networks*, Springer-Verlag, 1996

Plate, Tony A., Bert, Joel, and Band, Pierre, "Visualizing the Function Computed by a Feedforward Neural Network", *Neural Computation*, June 2000, pp. 1337-1353.

Potts, William J.E., *Neural Network Modeling: Course Notes*, SAS Institute, 2000

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

Smith, Murry, *Neural Networks for Statistical Modeling*, International Thompson Computer Press, 1996

Speights, David B, Brodsky, Joel B., Chudova, Durya L., "Using Neural Networks to Predict Claim Duration in the Presence of Right Censoring and Covariates", *Casualty Actuarial Society Forum*, Winter 1999, pp. 255-278.

Steinberg, Dan, "An Introduction to MARS", Salford Systems, 1999

Steinberg, Dan, "An Alternative to Neural Networks: Multivariate Adaptive Regression Splines (MARS)", *PC AI*, January/February, 2001, pp. 38 – 41

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

Viaene, Stijn, Derrig, Richard, Baesens, Bart and Dedene Guido, "A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Fraud Detection", *Journal of Risk and Insurance*, September, 2002, pp. 373 – 421

Warner, Brad and Misra, Manavendra, "Understanding Neural Networks as Statistical Tools", *American Statistician*, November 1996, pp. 284 – 293

Weisberg, Herbert I., and Richard A. Derrig, "Fraud and Automobile Insurance: A Report on the Baseline Study of Bodily Injury Claims in Massachusetts", *Journal of Insurance Regulation*, June, 1991, 9:4: 497-541.

Weisberg, Herbert I., and Richard A. Derrig, "Quantitative Methods for Detecting Fraudulent Automobile Injury Claims", AIB Cost Containment/Fraud Filing (DOI Docket R95-12), Automobile Insurers Bureau of Massachusetts, July,1993, 49-82.

Weisberg, Herbert I., and Richard A. Derrig, "Identification and Investigation of Suspicious Claims", AIB Cost Containment/Fraud Filing (DOI Docket R95-12), Automobile Insurers Bureau of Massachusetts, July, 1995, 192-245, Boston.

Weisberg, Herbert I., and Richard A. Derrig, "Massachusetts Automobile Bodily Injury Tort Reform", *Journal of Insurance Regulation*, Spring, 1992, 10:3:384-440.

Zapranis, Achilleas and Refenes, Apostolos-Paul, *Principles of Neural Network Model Identification Selection and Adequacy*, Springer-Verlag, 1999