Using Data Mining Techniques for Fraud Detection

Solving Business Problems Using SAS® Enterprise Miner™ Software

FEDERAL DATA CORPORATION

A SAS Institute

Best Practices Paper

In conjunction with

Federal Data Corporation

SAS Institute

Table of Contents

Figures

Tables

Abstract

Data mining combines data analysis techniques with high-end technology for use within a process. The primary goal of data mining is to develop usable knowledge regarding future events. This paper defines the steps in the data mining process, explains the importance of the steps, and shows how the steps were used in two case studies involving fraud detection.[1]

The steps in the data mining process are:

- problem definition
- data collection and enhancement
- modeling strategies
- training, validation, and testing of models
- analyzing results
- modeling iterations
- implementing results.

The first case study uses the data mining process to analyze instances of fraud in the public sector health care industry. In this study, called "the health care case," the data contain recorded examples of known fraudulent cases. The objective of the health care case is to determine, through predictive modeling, what attributes depict fraudulent claims.

In the second case study, a public sector organization deploys data mining in a purchase card domain with the aim of determining what transactions reflect fraudulent transactions in the form of diverting public funds for private use. In this study, called "the purchase card case," knowledge of fraud does not exist.

Problem Definition

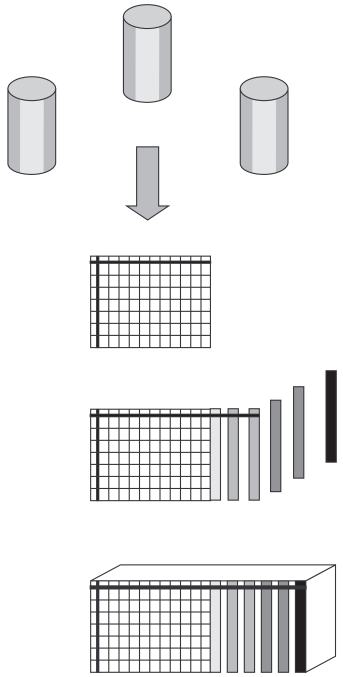
Defining the business problem is the first and arguably the most important step in the data mining process. The problem definition should not be a discussion of the implementation or efficacy of enabling technology that is data mining. Instead, the problem definition should state the business objective.

A proper business objective will use clear, simple language that focuses on the business problem and clearly states how the results are to be measured. In addition, the problem definition should include estimates of the costs associated with making inaccurate predictions as well as estimates of the advantages of making accurate ones.

[1] Due to the proprietary nature of the data used in this paper, the case studies focus on the data mining process instead of in-depth interpretation of the results.

Data Collection and Enhancement

Data mining algorithms are only as good as the data shown to them. Incomplete or biased data produce incomplete or biased models with significant blind spots. As Figure 1 illustrates, data collection itself involves four distinct steps:



1. **Define Data Sources**

Select from multiple databases. These data may include transaction databases, personnel databases, and accounting databases. Care should be taken though, as the data used to train data mining models must match the data on which models will be deployed in an operational setting. Thus, the tasks that need to be performed in data mining will have to be repeated when data mining models are deployed.

2. **Join and Denormalize Data**

This step involves joining the multiple data sources into a flat file structure. This step sometimes requires that decisions be made on the level of measurement (e.g. do some data get summarized in order to facilitate joining.)

3. **Enrich Data**

As data from disparate sources are joined, it may become evident that the information contained in the records is insufficent. For instance, it may be found that the data on vendors is not specific enough. It may be necessary to enrich data with external (or other) data.

4. **Transform Data**

Data transformations enable a model to more readily extract the valuable information from data. Examples of data transformations include aggregating records, creating rations, summarizing very granular fields, etc.

Figure 1: Data Collection and Associated Steps in the Data Mining Process

Defining the data sources should be a part of the details of the problem definition such as, “We want to get a comprehensive picture of our business by incorporating legacy data with transactions on our current systems.” Enriching data fills gaps found in the data during the join process. Finally, data mining tools can often perform data transformations involving aggregation, arithmetic operations, and other aspects of data preparation.

Modeling Strategies

Data mining strategies fall into two broad categories: supervised learning and unsupervised learning. Supervised learning methods are deployed when there exists a target variable[2] with known values and about which predictions will be made by using the values of other variables as input.

Unsupervised learning methods tend to be deployed on data for which there does not exist a target variable with known values, but for which input variables do exist. Although unsupervised learning methods are used most often in cases where a target variable does not exist, the existence of a target variable does not preclude the use of unsupervised learning.

Table 1 maps data mining techniques by modeling objective and supervised/unsupervised distinctions using four modeling objectives: prediction, classification, exploration, and affinity.

Modeling Objective	Supervised	Unsupervised
Prediction	Regression and Logistic regression Neural Networks Decision Trees Note: Targets can be binary, interval, nominal, or ordinal.	Not feasible
Classification	Decision Trees Neural Networks Discriminant Analysis Note: Targets can be binary, nominal, or ordinal.	Clustering (K-means, etc) Neural Networks Self-Organizing Maps (Kohonen Networks)
Exploration	Decision Trees Note: Targets can be binary, nominal, or ordinal.	Principal Components Clustering (K-means, etc)
Affinity	Not applicable	Associations Sequences Factor Analysis

Table 1: Modeling Objectives and Data Mining Techniques

Prediction algorithms determine models or rules to predict continuous or discrete target values given input data. For example, a prediction problem could attempt to predict the value of the S&P 500 Index given some input data such as a sudden change in a foreign exchange rate.

Classification algorithms determine models to predict discrete values given input data. A classification problem might involve trying to determine if transactions represents fraudulent behavior based on some indicators such as the type of establishment at which the purchase was made, the time of day the purchase was made, and the amount of the purchase.

[2] In some disciplines, the terms *field* and *variable* are synonymous.

Exploration uncovers dimensionality in input data. For example, trying to uncover groups of similar customers based on spending habits for a large, targeted mailing is an exploration problem.

Affinity analysis determines which events are likely to occur in conjunction with one another. Retailers use affinity analysis to analyze product purchase combinations.

Both supervised and unsupervised learning methods are useful for classification purposes. In a particular business problem involving fraud, the objective may be to establish a classification scheme for fraudulent transactions. Regression, decision trees, neural networks and clustering can all address this type of problem. Decision trees and neural networks build classification rules and other mechanisms for detecting fraud. Clustering can indicate what types of groupings in a given population (based on a number of inputs) are more at risk for exhibiting fraud.

Classification modeling tries to find models (or rules) that predict the values of one or more variables in a data set (target) from the values of other variables in the data set (inputs). After finding a good model, a data mining tool applies the model to new data sets that may or may not contain the variable(s) being predicted. When applying a model to new data, each record in the new data set receives a score based on the likelihood the record represents some target value. For example, in the health care case, fraud represents the target value. In this paper, case study 1—the health care case—uses classification modeling.

Exploration uses different forms of unsupervised learning. Clustering places objects into groups or clusters that are suggested by the data and are based on the values of the input variables. The objects in each cluster tend to have a similar set of values across the input fields, and objects in different clusters tend to be dissimilar. Clustering differs from artificial intelligence (AI) and online analytical processing (OLAP) in that clustering leverages the relationships in the data themselves to inductively uncover structure rather than imposing an analyst's structure on the data. In this paper, case study 2—the purchase card case—uses exploratory clustering.

Training, Validation, and Testing of Models

Model development begins by partitioning data sets into one set of data used to train a model, another data set used to validate the model, and a third used to test the trained and validated model.[3] This splitting of data ensures that the model does not memorize a particular subset of data. A model trains on one set of data, where it learns the underlying patterns in that data, then gets validated on another set of data, which it has never seen before. If the model does not perform satisfactorily in the validation phase (for example, it may accurately predict too few cases in the target field), it will be re-trained.

The training, validating, and testing process occurs iteratively. Models are repeatedly trained and validated until the tool reaches a time limit or an accuracy threshold. Data partitioning typically splits the raw data randomly between training and validation sets. In some instances, controlling the splitting of training and validation data is desirable. For instance, a credit card fraud case may necessitate controlled partitioning to avoid distributing the fraudulent transactions for one particular account between training and validation.

[3] Some texts and software products refer to the *test set* as the *validation set,* and the *validation set* as the *test set.*

After model training and validation, algorithm parameters can be changed in an effort to find a better model. This new model produces new results through training and validation. Modeling is an iterative process. Before settling on an acceptable model, analysts should generate several models from which the best choice can be made.

While training and validation are independent activities in the modeling process, they nonetheless are indirectly linked and could impact the generalizability of the developed model. During validation, the model indirectly "sees" the validation data, and tries to improve its performance in the validation sessions. The model may eventually memorize both the training data and indirectly the validation data making a third data set know as a test data set instrumental to delivering unbiased results.

The test data set is used at the very end of the model building process. The test data set must have a fully populated target variable. The test data set should only be used once, to evaluate or compare models' performance, not to determine how the model should be re-trained.

Analyzing Results

Diagnostics used in model evaluation vary in supervised and unsupervised learning. For classification problems, analysts typically review gain, lift and profit charts, threshold charts, confusion matrices, and statistics of fit for the training and validation sets, or for the test set. Business domain knowledge is also of significant value to interpreting model results.

Clustering models can be evaluated for overall model performance or for the quality of certain groupings of data. Overall model diagnostics usually focus on determining how capable the model was in dividing the input data into discrete sets of similar cases. However, analysts may also determine the adequacy of certain individual clusters by analyzing descriptive statistics for key fields in the cluster vis-à-vis remaining data. For instance, if an analyst seeks patterns of misuse of a purchase card, clusters with high concentrations of questionable purchases may be targeted for investigations.

Linking Techniques to Business Problems

Business problems can be solved using one or more of the data mining techniques listed in Table 1. Choice of the data mining technique to deploy on a business problem depends on business goals and the data involved. Rarely does a data mining effort rely on a single algorithm to solve a particular business problem. In fact, multiple data mining approaches are often deployed on a single problem.

Table 2 displays some uses of data mining in the different modeling objectives

Modeling Objective	Supervised (Target Field Data Exists)	Unsupervised (No Target Field Data Exists)
Prediction	Attrition/Retention Cash Used in ATM Cost of Hospital Stay Fraud Detection Campaign Analysis	Not feasible
Classification	Segmentation Brand Switching Charge Offs Fraud Detection Campaign Analysis	Segmentation Attrition/Retention
Exploration	Segmentation Attrition/Retention Scorecard Creation Fraud Detection Campaign Analysis	Segmentation Profiling
Affinity	Not applicable	Cross-sell/Up-sell Market Basket Analysis

Table 2: Use of Data Mining by Modeling Objective and Learning Method

Case Study 1: Health Care Fraud Detection

Problem Definition

A public sector health care organization has begun to track fraudulent claims. In most cases, the organization identifies fraudulent claims by receiving tips from concerned parties then investigating those tips. In this case study, the business question centers on identifying patterns that produce fraudulent claims for health care among program beneficiaries.

Data Collection and Enhancement

A public sector organization has collected data in which there exists a field indicating fraudulent activity. Some 15 input fields exist as inputs that predict fraud. Figure 2 shows how SAS® Enterprise Miner ™ displays a table view of the health care fraud data.

Most of the records in Figure 2 have missing data for some fields; however, missing data actually represent important data values. The data came directly from a transaction processing system that recorded “missing” as the default value. For instance, if person was born in the U.S., then no country of birth value is entered for that record. In order to become useful, this transaction data must be cleansed and transformed. For example, “missing” values for the country of birth variable are reset to “U.S.” prior to analysis.

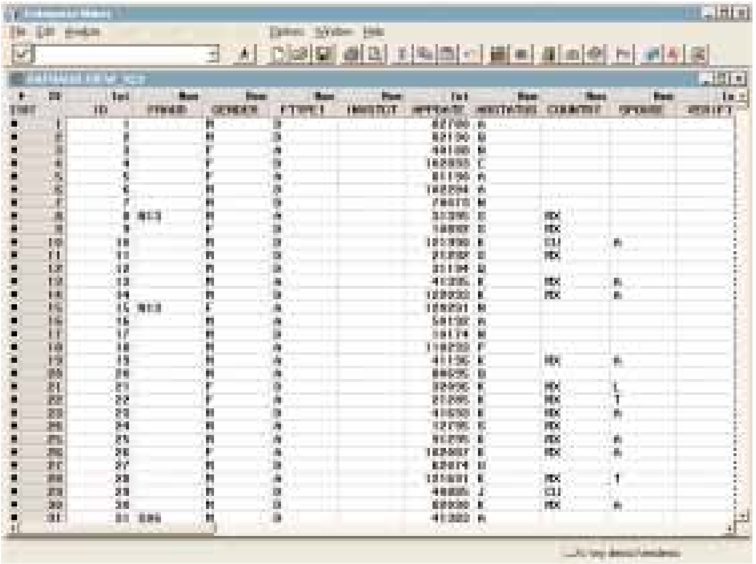


Figure 2: Missing Health Care Fraud Data

Modeling Strategies

In choosing a modeling strategy for the data in this case study, the following factors come into play:

- amount of missing data and how it is handled
- measurement level of the input variables
- percentage of data representing the target event (fraud)
- goal of the analysis – understanding predictive factors versus making good predictions.

The level of measurement of the target field is binary and the inputs are primarily nominal with missing values that will be assigned to their own class before analysis. There is one interval variable with no missing values. So our choice of modeling strategy is not highly restricted yet. In fact, regression, decision trees, or neural networks may be appropriate for the problem at hand. However, training a neural network will be slow with the nominal inputs.

The goals of the analysis are to understand how the input factors relate to predicting fraud and to develop rules for identifying new cases to investigate. Because neural networks provide little if any feedback on how the inputs relate to the target, they may be inappropriate for this analysis. Thus for this exercise, regression and decision trees were used. This leads to the last consideration – the percentage of fraudulent records in the data. To understand the factors in choosing a modeling strategy, you must understand how the model algorithms work.

Decision trees will try to group all the nominal values of an input into smaller groups that are increasingly predictive of the target field. For example, for an input with 8 nominal levels, the decision tree might automatically create two groups with 3 and 5 levels each so that the 3-level grouping contains the majority of fraudulent records. Because the tree puts the data into a large group and then tries to split the large groups, the decision tree has most of the data available to work with from the start. Data do get scarce as the tree grows, because each new split subsets the data for further modeling.

In terms of data scarcity, regression works differently than decision trees. All levels of all inputs are used to create a contingency table against the target. If there are not enough data in a cell of the table, then the regression will have problems including that input in the model. Scarcity of data becomes even more of a problem if interactions or cross-terms are included in the model. For example, if your inputs include geographical region with 50 levels, and product code with 20 levels, a regression model with a region by product interaction will create 50+20+50*20 cells in the contingency table. You would need a large amount of data in each of these 1070 cells to fit a regression model.

Training, Validation, and Testing of Models

In the health care fraud case study, only about 7 percent of the records in the data set represent fraudulent cases. In addition, many of the inputs have a high number of nominal levels. Combined, these two factors make analysis with a regression model difficult. However, for comparison a forward stepwise regression was tested, and it was determined that its selection of significant predictors agreed with the decision tree’s fit.

Figure 3 shows how SAS Enterprise Miner was used to create a process flow diagram for analyzing the health care fraud data.

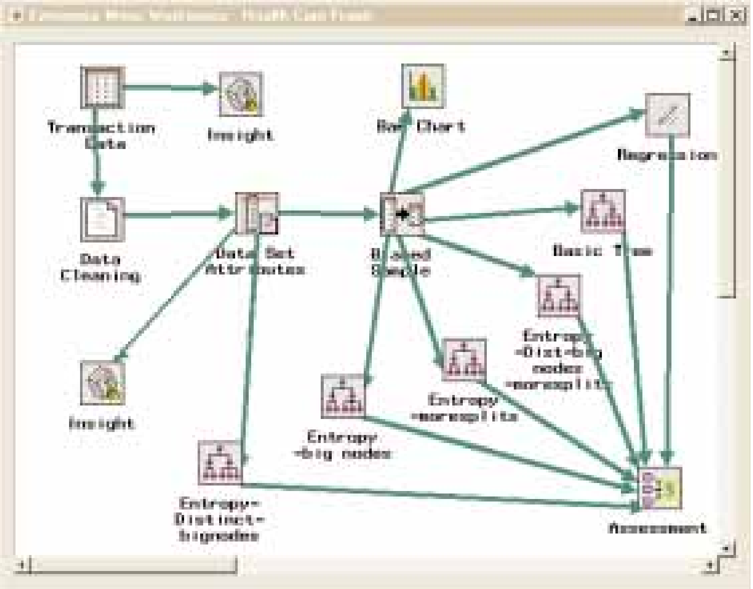


Figure 3: Analysis Path for Health Care Data

In fitting the decision tree, many of the options were adjusted to find the best tree. Unfortunately there were only 2107 records of which only 153 were considered fraudulent.

This is not enough data to create a validation data set that would ensure generalizable results. Options used include:

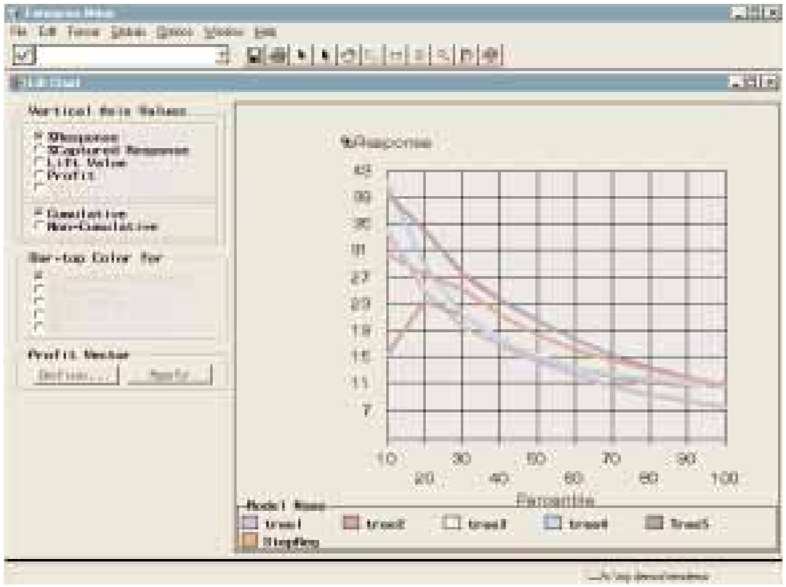
- CHAID type of Chi-square splitting criterion, which splits only when a statistically significant threshold is achieved
- entropy reduction splitting criterion, which measures the achieved node purity at each split
- evaluating the tree based on the node purity (distinct distribution in nodes)
- evaluating the tree based on the misclassification rate (general assessment)
- allowing more than the default number of splits (2) at each level in the tree
- adjusting the number of observations required in a node and required for splitting to continue.

Analyzing Results

To analyze the results of the health care study, we used lift charts and confusion matrices. A lift chart displays results from different models allowing a quick initial model comparison. A confusion matrix involves a comparison of predicted values to actual values.

Analyzing Results Using Lift Charts

Lift charts compare the percentage of fraudulent observations found by each of the decision trees and the regression model. In Figure 4 the lift chart shows the percent of positive response (or lift) on the vertical axis. The lift chart reveals that two of the trees, 4 and 5, outperformed the other models.



Vertical Axis Values
● %Response
○ %Captured Response
○ Lift Value
○ Profit

● Cumulative
○ Non-Cumulative

Figure 4: Lift Chart Comparing Multiple Models

Tree 4 allows only 2 splits at each branch, requires at least 20 observations in a node to continue splitting, and uses the entropy reduction algorithm with the best assessment option. Tree 5 allows 4 splits at each branch, requires at least 20 observations in a node to continue splitting, and uses the entropy reduction algorithm with the distinct distribution in nodes option. To decide which of these trees to use, several factors need to be considered:

- What is the desired tradeoff between false alarms and false dismissals?
- How much fraud overall is each tree able to identify?
- How have the algorithmic options that have been set affected the tree results?
- Does one tree make more sense than the other does from a business perspective?
- Is one tree simpler to understand and implement than the other?

Analyzing Results Using Confusion Matrices

Individual model performance of supervised learning methods is often assessed using a confusion matrix. The objective, typically, is to increase the number of correct predictions (sensitivity) while maintaining incorrect predictions or the false alarm rate (specificity) at an acceptable level. The two goals, getting as much of the target field correctly predicted versus keeping the false alarm rate low, tend to be inversely proportional. A simple example can illustrate this point: to catch all the fraud in a data set, one need only call health care claims fraudulent, while to avoid any false alarms one need only call all claims non-fraudulent. Reality resides between these two extremes.

The business question typically defines what false alarm rate is tolerable versus what amount of fraud (or other target) needs to be caught.

Table 3 displays the layout of a confusion matrix. The confusion matrix compares actual values of fraud (rows) versus model predictions of fraud (columns). If the model predicted fraud perfectly, all observations in the confusion matrix would reside in the two shaded cells labelled "Correct Dismissals" and "Correct Hits." Generally, the objective is to maximize correct predictions while managing the increase in false alarms.

		Model Predictions	
		Model Predicts Non-Fraud	**Model Predicts Fraud**
Actual Values of Fraud	**Actual Transaction is Non-Fraudulent**	**Correct Dismissals**	***False Alarms***
	Actual Transaction is Fraudulent	***False Dismissals***	**Correct Hits**

Table 3: Layout of a Confusion Matrix

When predicting for classification problems, each record receives a score based on the likelihood that the record represents some target value. Because the likelihood is a probability, its values range from zero to one inclusive. While most modeling packages apply a standard cutoff or threshold to this likelihood and then determine the predicted classification value, SAS Enterprise Miner enables the analyst to modify the default threshold. Changing the threshold value changes the confusion matrix.

Figures 5 and 6 display the confusion matrix of tree 5 for thresholds of 0.10 and 0.50 respectively. With a 0.10 threshold, records that have a 10 percent or higher chance of being fraudulent are predicted as fraudulent. The 10 percent threshold predicts more false alarms while the 50 percent threshold predicts more false dismissals. A correct classification diagram enables evaluation of the confusion matrix over a wide range of thresholds.

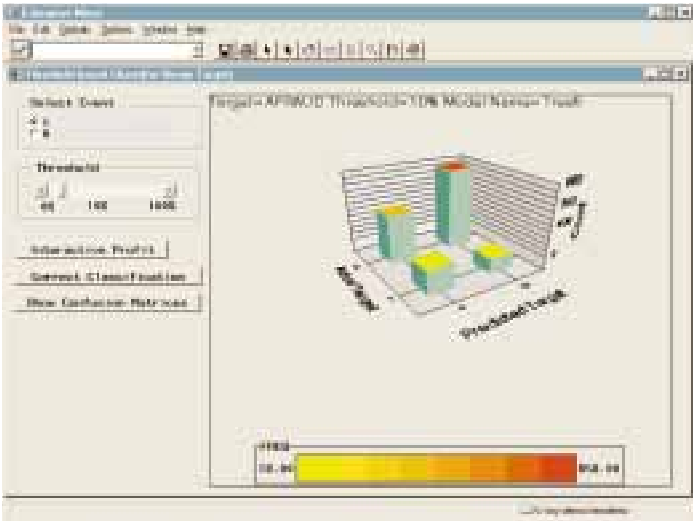


Figure 5: Tree 5 Confusion Matrix at 10% Threshold

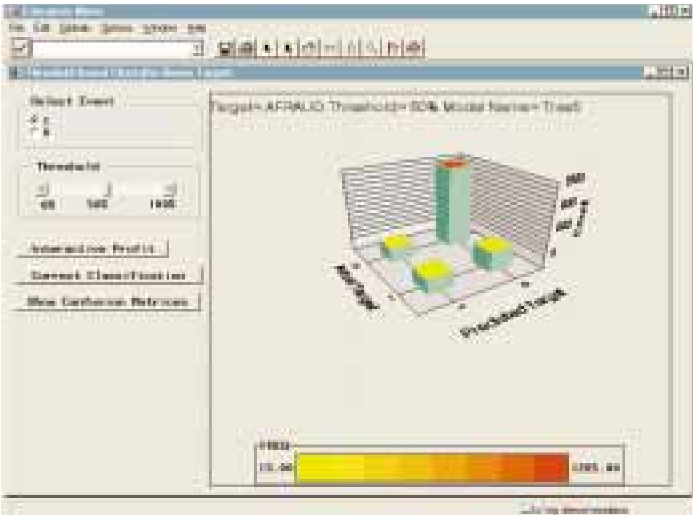


Figure 6: Tree 5 Confusion Matrix at 50% Threshold

Figure 7 displays a correct classification rate diagram for tree 5. This plot shows the tradeoff between sensitivity and specificity enabling the analyst to determine an appropriate cut-off value for the likelihood that a particular record is fraudulent. For this example, the curve for Target Level of both illustrates that the threshold can vary upwards of 20 percent without significantly affecting the results.

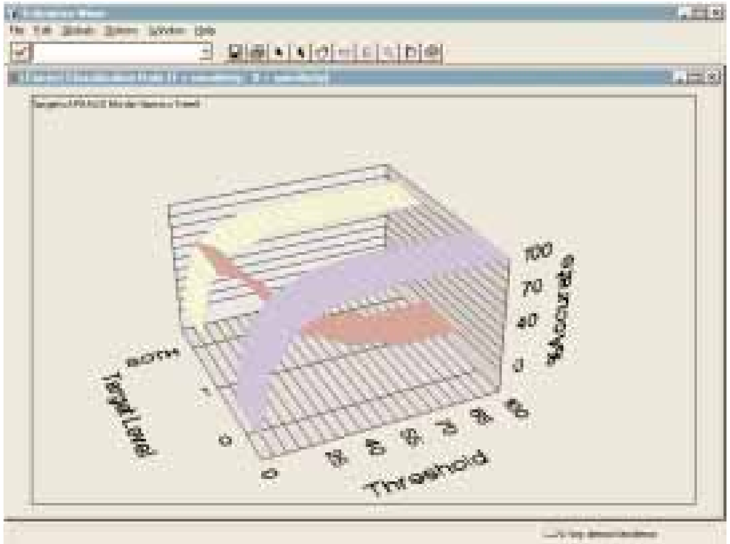


Figure 7: Tree 5 Correct Classification Diagram

Table 4 displays statistics for decision trees 4 and 5. In this case, both trees produce quite similar results, first splitting on gender then other inputs in different orders. Because of the options chosen after the initial split, tree 4 has focused on isolating fraud in the males (68 percent of all fraud) while tree 5 has found fraud for both males and females. At first glance, tree 4 with fewer splits may seem simpler, yet tree 5 with more splits has less depth and simpler rules. Tree 5 also isolates 37 percent of all fraudulent cases whereas tree 4 only isolates 25 percent of all fraud.

Time	Node	# Fraud (%)	Gender
4	61	22 (14%)	Male
4	106	7 (4.6%)	Male
4	93	8 (5.2%)	Male
Tree 4 Total		**3 (25%)**	
5	16	7 (4.6%)	Male
5	49	4 (2.8%)	Male
5	46	11 (7.2%)	Male
5	43	15 (9.8%)	Male
5	65	6 (3.9%)	Female
5	63	7 (4.6%)	Female
5	32	4 (2.6%)	Female
5	2	3 (2%)	Female
Tree 5 Total		**5 (37%)**	

Table 4: Table of Tree Statistics

The visualization techniques available in SAS Enterprise Miner also are helpful when analyzing the trees. For example, tree ring diagrams provide holistic views of decision trees. Figure 8 displays a tree ring diagram for tree 4. The center of the ring represents the top tree node including all the data. The first ring represents the first tree split, in this case on gender.

Subsequent rings correspond to subsequent levels in the tree. Colors are assigned to show the proportion of fraudulent records correctly classified in each node of the tree. Light sections correspond to less fraud, dark sections to more fraud.

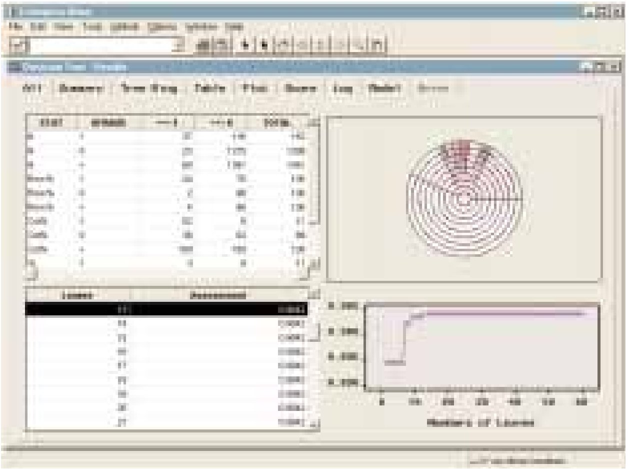


Figure 8: Tree Ring Diagram for Decision Tree 4

Using the diagnostics readily available in the software enables analysts to investigate quickly the darker portions of the diagram and to generate a subset of the tree that displays the rules required to create the associated data subset. Figure 9 displays rules for decision tree 4.

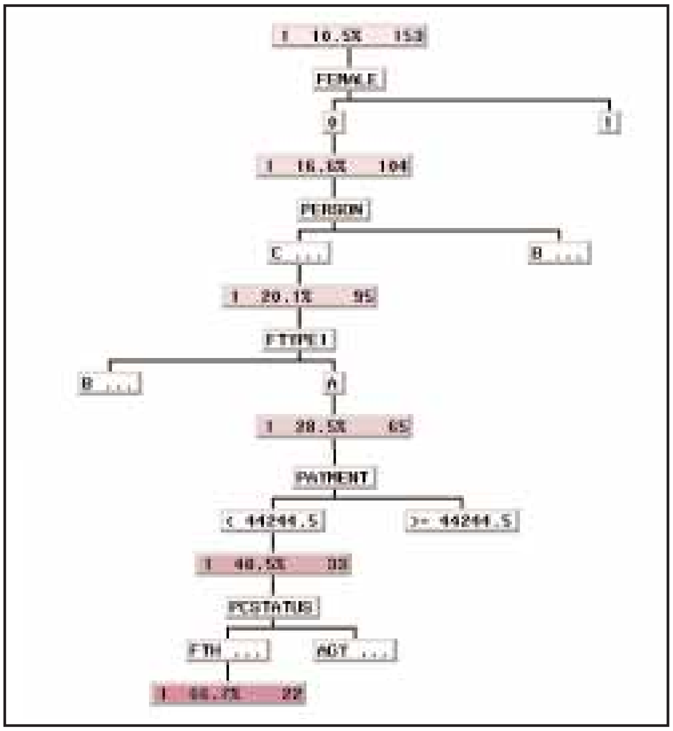


Figure 9: A Subset of the Rules for Tree 4

An example rule from Figure 9 specifies that 14 percent of the fraudulent records can be described as follows:

- male
- from four specific person categories of file type A
- received payments of between $19,567 and $44,500
- one of three ‘pc’ status values.

Deriving these rules from Figure 9 is straightforward; however, notice that the payment amount is addressed twice in the model. Model options for tree 4 specified that only one split point could be defined on an input at each level of the tree. This algorithmic setting often causes decision trees to create splits on a single input at multiple levels in the tree making rules more difficult to understand.

Following a similar set of steps for tree 5 enables a comparison of the two trees at a more granular level. Figure 10 displays the tree ring for decision tree 5 for which four split points for each input were allowed in each level of the tree. Allowing the algorithm more freedom in splitting the inputs resulted in a tree with fewer levels that addresses more of the data – in particular both males and females. A quick glance at the tree ring may suggest tree 5 is more complex than tree 4. However, each input appears at only one level in the tree, making the rules easier to understand.

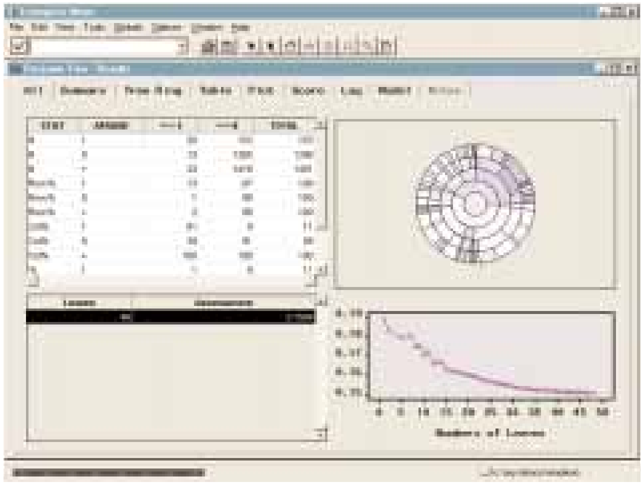


Figure 10: Tree Ring Diagram for Decision Tree 5

Decision tree 5 is displayed graphically in Figure 11 as a set of splits (decisions).

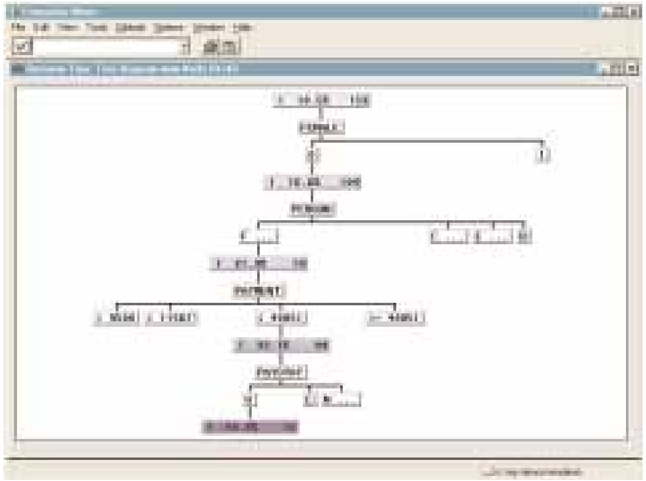


Figure 11: Rules for Tree 5

An example rule from Figure 11 specifies that 9.8 percent of the fraudulent records can be described as follows:

- male
- from two specific person categories or the ‘missing’ category
- of payment status U
- received payments of between $11,567 and $40,851.

Conclusions for Case Study 1

Based on the amount and type of data available, a decision tree with rules that are simple to follow provides the best insight into this data. Of course, the best recommendation is to obtain more data, use a validation data set, and have subject matter expertise applied to enhance this analysis.

Case Study 2: Purchase Card Fraud Detection

A federal agency has collected data on its employees’ purchase card transactions and on the 40,000 employees’ purchase card accounts. The transaction data contain information on the date purchases are made, the purchase amount, the merchant’s name, the address of the merchant, and the standard industrial classification (SIC) code of the merchant among other fields. The account data contain information about the individuals’ accounts such as information about the account holder, the single transaction limit of the account, the billing cycle purchase limit for the account, and purchase histories for each account among other fields.

Problem Definition

A government organization seeks to determine what groups of purchases exist in its purchase card program that may be indicative of a misuse of public funds. The organization has collected information on purchase characteristics that signify a misuse of government funds. This information is resident in reports about purchase card risk. Additional information resides with domain experts. The organization seeks to determine what other types of transactions group together with the existing knowledge for the sake of preventing continued misuse of funds by authorized individuals.

The organization wishes to build an effective fraud detection system using its own data as a starting point.

Data Collection and Enhancement

After defining the business problem, the next step in the data mining process is to link the disparate data sources. In this case study, data from account and transaction files are linked. Data are joined at the transaction level because the business question is focused on determining the inherent properties of transactions that signify fraudulent use of funds.

Typically, not all of the data that have been joined will be selected for model inputs. Some subset of the fields will be used to develop the data mining models.

Data transformations can be performed on the collected data. Data transformations involve converting raw inputs. For example, data transformations might group very granular categorical variables such as SIC codes into more general groups, or aggregating records. Data transformations make more efficient use of the information embedded in raw data. Data transformations can be made with the assistance of domain experts. In this case study, domain experts have indicated some SIC code purchases indicate a misuse of funds.

Typically, data mining requires the drawing of samples from the records in the joined data due to the intense resources required in the training of data mining algorithms. Samples need to be representative of the total population so that models have a chance to “see” possible combinations of fields.

Modeling Strategies

In this case study, no target field exists because the organization has never analyzed purchase card data in search of fraud. Therefore, the decision is made to use unsupervised learning methods to uncover meaningful patterns in the data. Unsupervised learning will be used to group the data into sets of similar cases.

Figure 12 displays the selection of an unsupervised learning method using SAS Enterprise Miner. A sample of approximately 13,000 accounts is created. Cluster analysis segments the sample data into sets of similar records.

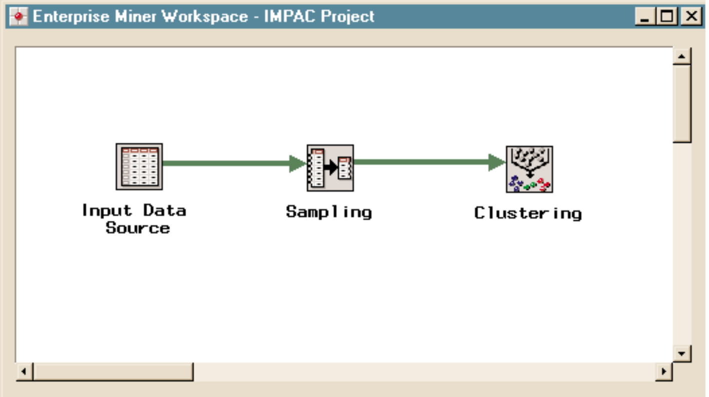


Figure 12: Selection of Clustering as an Unsupervised Learning Method

The unsupervised learning method selected in Figure 12 performs disjoint cluster analysis on the basis of Euclidean distances computed from one or more quantitative variables and seeds that are generated and updated by a clustering algorithm. Essentially the clustering method bins the data into groups in such a way as to minimize differences within groups at the same time that it maximizes differences between groups.

The cluster criterion used in this example is Ordinary Least Squares (OLS), wherein clusters are constructed so that the sum of the squared distances of observations to the cluster means is minimized.

Training, Validation, and Testing of Models

Figure 13 displays a hypothetical clustering result. The large crosses represent cluster centers. The cases are assigned to three clusters (each with an ellipse drawn about it). In the space represented, there is no better way to assign cases to clusters in order to minimize distance from each data point to the center of each cluster. Of course, this example displays a simple two-dimensional representation; cluster analysis performs its optimization routines in *m*-dimensional space, where *m* is the number of fields or variables. Therefore, if there are 20 variables in the clustering operation, the space in which clustering is performed is 20-dimensional space.

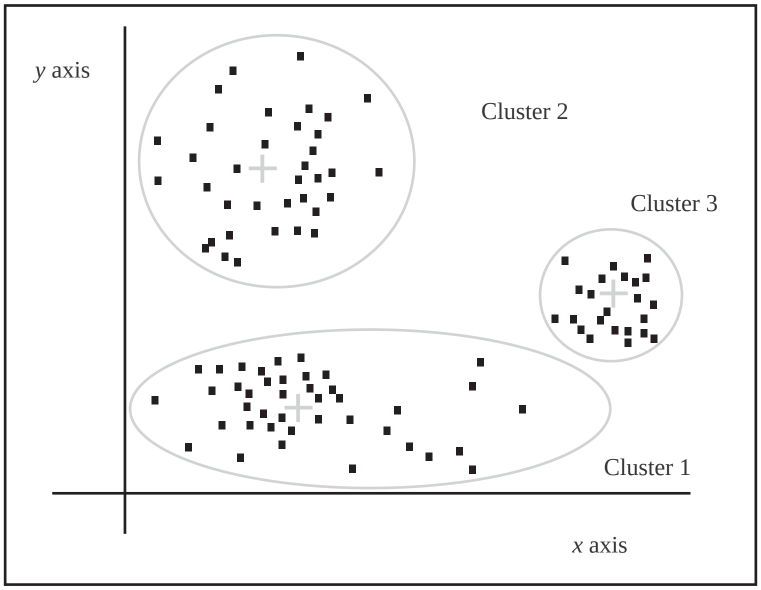
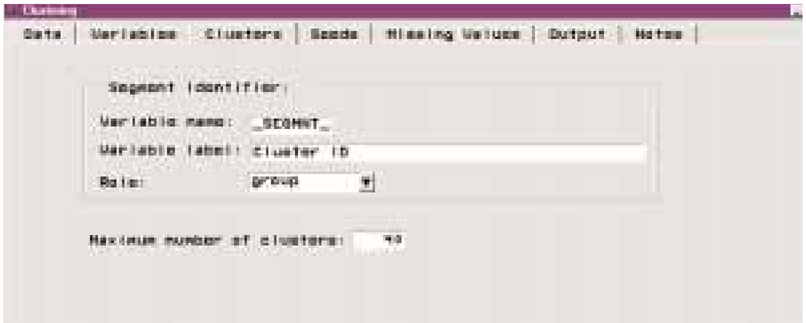


Figure 13: Cluster Analysis Efficiently Segments Data into Groups of Similar Cases

The difference between exploratory analysis and pattern discovery in clustering concerns what constitutes a result and how the results will be put to use. Exploratory analysis may be satisfied to discover some interesting cases in the data. Pattern discovery will leverage the existing clusters and the general patterns associated with those clusters to assign new cases to clusters. As a result of this more forward-looking objective, cluster analysis in pattern discovery requires cluster models to be tested prior to deployment. This testing ensures a reliable result, one that can help ensure that "discovered" clusters in the data persist in the general case.

In this case study, cluster analysis is used as a pattern detection technique; therefore, the resulting cluster model would need to be tested were it to be applied.

Part of the model training process involves selecting parameters for the cluster model. Figure 14 shows the parameter settings for the current cluster model. In this case, four cluster centers are selected.



Segment identifier:

Variable name: _SEGMNT_

Variable label: Cluster ID

Role: group

Maximum number of clusters: 40

Figure 14: Selecting Cluster Model Parameters

The model settings in Figure 14 will produce a cluster model with four centers. The algorithm will try to arrange the data around the four clusters in such a way as to minimize differences within clusters at the same time that it maximizes differences between clusters.

Analyzing Results

Figure 15 displays results for a cluster analysis using the purchase card data. The parameters for the cluster analysis were set to 40 clusters. The height and color of each pie slice represent the number of cases in the cluster. The slice width refers to the radius of the circle that covers all data points in the cluster as measured from the center of the cluster. Cluster 31 holds the largest number of cases at 6,334, while the clusters 1,11, and 19 each have in excess of 500 cases. Cluster 6 has 345 cases.

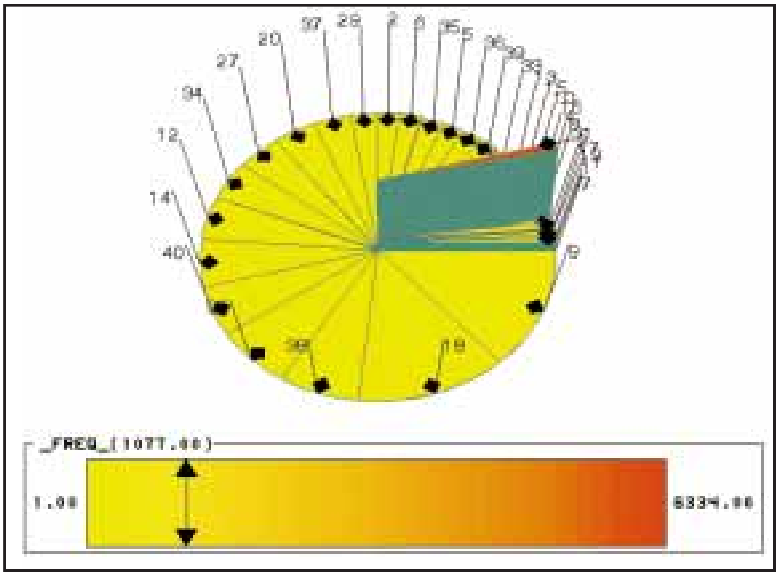


Figure 15: Results of Cluster Analysis

Figure 16 displays cluster statistics. The column titles represent standard industrial classification (SIC) codes where purchases have taken place. The number in each cell corresponds to the average frequency of purchases made by the account holders in that account. In this case, cluster 6 (with 345 cases) is highlighted as account holders in this cluster make an average of 6.51 sports and leisure purchases each.

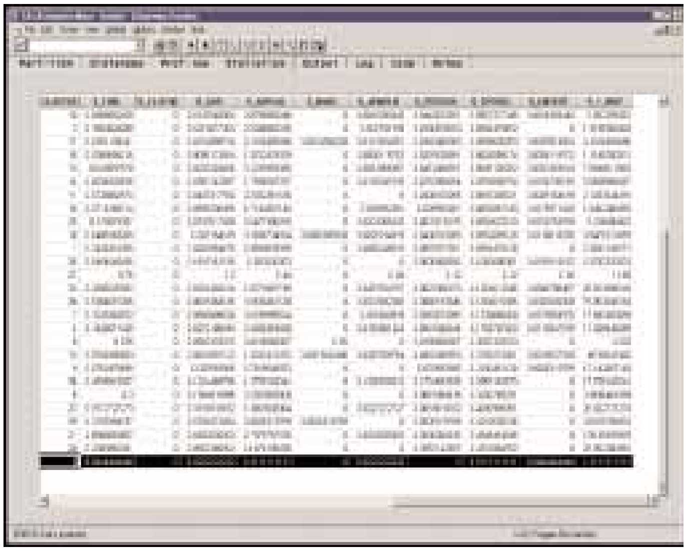


Figure 16: Cluster Statistics

Looking at the raw data for the cases in cluster 6, we find that account holders in that cluster also make a high amount of weekend and holiday purchases, restaurant purchases and hotel purchases. These accounts are problematic as the patterns exhibited by them clearly indicate improper use of purchase cards for personal and/or unwarranted expenses.

As investigation of clusters proceeds, it also is necessary to ensure that enough of a split occurs between the clusters, which will demonstrate that sufficient difference exists between clusters. Lastly, it is important to identify the relevance of the clusters, which is achieved with the aid of domain expertise. Individuals who are knowledgeable of purchase card use can help indicate what batches of data are promising given the business question.

The model would still need to be tested by using new data to ensure that the clusters developed are consistent with the current model.

Building from Unsupervised to Supervised Learning

Pattern detection provides more information on fraudulent behaviors than simply reporting exceptions and can prove valuable in the future for building a knowledge base for predicting fraud. For example, the cluster analysis in this case study yields interesting results. In fact, one of the clusters holds the promise of uncovering fraudulent transactions, which may require investigation through account audits. The ultimate findings of the investigations should be stored in a knowledge base, which can be used to validate the cluster model. Should investigation show the model's judgment to be erroneous, the cluster analysis would need to be revisited.

The tested cluster model can continue to be applied to new data, producing cases for investigation. In turn, the knowledge base will accumulate known fraud cases.

Conclusions for Case Study 2

Cluster analysis yields substantive results in the absence of a target field. Used wisely, cluster analysis can help an organization interested in fraud detection build a knowledge base of fraud. The ultimate objective would be the creation of supervised learning model such as a neural network that is focused on uncovering fraudulent transactions.

Overall Conclusions

Data mining uncovers patterns hidden in data to deliver knowledge for solving business questions. Even in the absence of target fields, data mining can guide an organization's actions toward solving its business questions and building a growing knowledge base. The powerful data mining tools found in SAS Enterprise Miner software make it easy for organizations to extract knowledge from data for use in solving core business questions.

When followed, the steps in the data mining process (problem definition; data collection and enhancement; modeling strategies; training and validating models; analyzing results; modeling iterations; and implementing results) provide powerful results to organizations.

Biographies

I. Philip Matkovsky

Federal Data Corporation
4800 Hampden Lane
Bethesda, MD 20814
301.961.7024
pmatkovsky@feddata.com

As manager of operations for Federal Data Corporation's Analytical Systems Group, Philip Matkovsky provides technical lead and guidance on data mining, quantitative analysis, and management consulting engagements for both public and private sector clients. Philip has a BA in Political Science from the University of Pennsylvania, an MA in Political Science/ Public Policy from American University and is currently completing his doctoral research in Public Policy at American University. Philip has successfully applied numerous analytical/ research approaches (including survey research, game theoretic models, quantitative modeling, and data mining) for public and private sector clients.

Kristin Rahn Nauta

SAS Institute, Inc.
SAS Campus Drive
Cary, NC 27513
919.677.8000 x4346
saskrl@wnt.sas.com

As part of the Federal Technology Center at SAS Institute Inc., Kristin Rahn Nauta is the federal program manager for data mining. Formerly SAS Institute's data mining program manager for Canada and the analytical products marketing manager for the US, Kristin has a BS in mathematics from Clemson University and a Masters of Statistics from North Carolina State University. Kristin has consulted in a variety of fields including pharmaceutical drug research and design, pharmaceutical NDA submissions, database marketing and customer relationship management.

Recommended Reading

Data Mining

Adriaans, Pieter and Dolf Zantinge. (1996) Data Mining. Harlow, England: Addison Wesley.

Berry, Michael J. A. and Gordon Linoff, (1997), Data Mining Techniques, New York: John Wiley & Sons, Inc.

SAS Institute Inc., (1997), SAS Institute White Paper, Business Intelligence Systems and Data Mining, Cary, NC: SAS Institute Inc.

SAS Institute Inc., (1998), SAS Institute White Paper, Finding the Solution to Data Mining: A Map of the Features and Components of SAS® Enterprise Miner™ Software, Cary, NC: SAS Institute Inc.

Weiss, Sholom M. and Nitin Indurkhya, (1998), Predictive Data Mining: A Practical Guide, San Francisco, California: Morgan Kaufmann Publishers, Inc.

Data Warehousing

Berson, Alex and Stephen J. Smith (Contributor). (1997) Data Warehousing, Data Mining and OLAP, New York: McGraw Hill.

Inmon, W. H., (1993), Building the Data Warehouse, New York: John Wiley & Sons, Inc.

SAS Institute Inc., (1995), SAS Institute White Paper, Building a SAS® Data Warehouse, Cary, NC: SAS Institute Inc.

SAS Institute Inc., (1996), SAS Institute White Paper, SAS Institute's Rapid Warehousing Methodology, Cary, NC: SAS Institute Inc.

Singh, Harry, (1998), Data Warehousing Concepts, Technologies, Implementations, and Management, Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Credits

Using Data Mining Techniques for Fraud Detection was a collaborative work. Contributors to the development and production of this paper included the following persons:

Consultants

SAS Institute Inc.
Padraic G. Neville, Ph.D.

Writers

SAS Institute Inc.
Kristin Rahn Nauta, M.Stat.
Federal Data Corporation
I. Philip Matkovsky

Technical Reviewers

SAS Institute Inc.
Brent L. Cohen, Ph.D.
Bernd Drewes
Anne Milley
Warren S. Sarle, Ph.D.
Federal Data Corporation
Steve Sharp
Paul Simons

Technical Editor

SAS Institute Inc.
John S. Williams

Copy Editors

SAS Institute Inc.
Rebecca Autore
Sue W. Talley

Production Specialist

SAS Institute Inc.
Kevin Cournoyer

Copyright© 1999 by Federal Data Corporation, Bethesda, MD and SAS Institute Inc., Cary, NC. All rights reserved. Credit must be given to the publisher. Otherwise, no part of this publication may be reproduced without prior permission of the publisher.

SAS and all other SAS Institute Inc. product or service names are the registered trademarks or trademarks of SAS Institute Inc. ® indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective owners.

SAS Institute
Europe, Middle East & Africa
P.O. Box 10 53 40
Neuenheimer Landstr. 28-30
D-69043 Heidelberg, Germany
Tel: (49) 6221 4160 Fax: (49) 6221 474850

SAS Institute Inc.
World Headquarters
SAS Campus Drive, Cary, NC 27513 USA
Tel: (919) 677 8000, Fax: (919) 677 4444
In Canada, Tel: (800) 363 8397
www.sas.com

Federal Data Corporation
4800 Hampden Lane
Bethesda, MD 20814
Tel: (301) 961-0500
Fax: (301) 961-0685

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 1999, SAS Institute Inc. All rights reserved.