A Review of Software Packages for Data Mining

Dominique HAUGHTON, Joel DEICHMANN, Abdolreza ESHGHI,
Selin SAYEK, Nicholas TEEBAGY, and Heikki TOPI

We present to the statistical community an overview of five data mining packages with the intent of leaving the reader with a sense of the different capabilities, the ease or difficulty of use, and the user interface of each package. We are not attempting to perform a controlled comparison of the algorithms in each package to decide which has the strongest predictive power, but instead hope to give an idea of the approach to predictive modeling used in each of them. The packages are compared in the areas of descriptive statistics and graphics, predictive models, and association (market basket) analysis.

As expected, the packages affiliated with the most popular statistical software packages (SAS and SPSS) provide the broadest range of features with remarkably similar modeling and interface approaches, whereas the other packages all have their special sets of features and specific target audiences whom we believe each of the packages will serve well. It is essential that an organization considering the purchase of a data mining package carefully evaluate the available options and choose the one that provides the best fit with its particular needs.

KEY WORDS: Clementine; Ghostminer; Quadstone; SAS Enterprise Miner; XLMiner.

1. INTRODUCTION

The term “data mining” has come to refer to a set of techniques that originated in statistics, computer science, and related areas that are typically used in the context of large datasets. The purpose of data mining is to reveal previously hidden associations between variables that are potentially relevant for managerial decision making. The exploratory and modeling techniques used in data mining are familiar to many statisticians and include exploratory tools such as histograms, scatterplots, boxplots, and analytic tools such as regression, neural nets, and decision trees.

This article’s objective is to present to the statistical community an overview of five data mining packages, and to leave the reader with a sense of the different capabilities, the ease or difficulty of use, and the user interface of each package. We are not attempting to perform a controlled comparison of the algorithms in each package to decide which has the strongest predictive power, but instead aim to give an idea of the approach to predictive modeling used in each of them.

The article is structured as follows: we first outline the methodology we used to evaluate the packages and give a summary of key characteristics of each package. We continue by focusing on descriptive statistics and exploratory graphs. The section that follows is devoted to predictive modeling, covering model building and assessment. A section on association (market basket) analysis is then provided, followed by a conclusion.

2. METHODOLOGY

The list of packages we have selected for this review is by no means exhaustive. We have chosen to cover the data mining packages associated with the two leading statistical packages, SAS and SPSS. We also decided to review two “stand-alone” packages, GhostMiner and Quadstone, and an Excel add-on, XLMiner.

We compare the packages in the areas of descriptive statistics and graphics, predictive models, and association (market basket) analysis. Predictive modeling is one of the main applications of data mining, and exploratory descriptive analyses always precede modeling efforts. Association analysis, in which “baskets” of goods purchased together are identified, is also very commonly used.

For the descriptive and modeling analysis, we used the Direct Marketing Educational Foundation dataset 2, merged with Census geo-demographic variables from dataset 6 (www.the-dma.org/dmef). The dataset contains 19,185 observations and concerns a business with multiple divisions, each mailing different catalogs to a unified customer database. The target variable, BUY10, equals unity if a customer made a purchase from the January 1996 division D catalog, zero if not. Data available (through June 1995) as potential predictors, for the whole business and each division, include: life-to-date orders, dollars, and number of items; orders and dollars in the most recent 6, 12, 24, and 36 months; recency of first and latest purchases; payment method and minimal demographics. Census geo-demographic variables give race, population, age profiles, as well as information on property values at the zip-plus-four level. The number of candidate predictor variables is nearly 200, representing a realistic situation in database marketing situations.

For our association analysis, we chose to use the Direct Marketing Educational Foundation’s Bookbinders Club Case dataset including data from 1,580 customers.

A typical hardware environment used in our tests was an 800MHz IBM A22m with 256 MB RAM (except for Quadstone, which required 512MB of RAM) and a 30 GB hard drive.

3. SUMMARY OF KEY CHARACTERISTICS

Table 1 presents a brief summary of the main characteristics of the packages reviewed here. Later sections will discuss many

Dominique Haughton, Joel Deichmann, Abdolreza Eshghi, Selin Sayek, Nicholas Teebagy, and Heikki Topi are members of the Data Analytics Research Team, Bentley College, 175 Forest Street, Waltham, MA 02452 (E-mail: dhaughton@bentley.edu). Selin Sayek is also affiliated with Bilkent University, Turkey. The authors thank each of the vendors of the reviewed packages for their assistance in dealing with installation concerns and for their very prompt replies to all of the authors’ questions. Our sincere thanks also go to Section Editor Joe Hilbe and his editorial staff for their careful reading of the manuscript and their support for the project, as well as for their many useful comments. We also thank Carter Rakovski and other members of the Academic Technology Center at Bentley College for all their help.

DOI: 10.1198/0003130032486

Table 1. Summary of Key Characteristics of the Packages

Software	*SPSS Clementine*	*XLMiner*	*Quadstone*	*GhostMiner*	*SAS Enterprise Miner*
Version	Client 7.1 (2002)	1.1.7.3 (5 June 2003)	4.0 b	Developer 1.0	4.1
Operating System	WinME/98/XP/2000/ NT 4.0	Win98/2000 XP	Win NT Server 4.0, Win 2000	Win NT/ 2000/XP	Win NT/2000/XP (client); UNIX, Linux, MVS
Hard Drive Space	320 MB	1GB	100 MB	35 MB	Not specified
Processor	Not specified	133 MHz	Varies (#users)	Pentium	Not specified
Other requirements	High resolution (1024 x 768 recommended)	Microsoft Excel 2000/XP	Java JRE 1.3.1 (installed with Quadstone)	None	Base SAS and SAS/STAT required
Easy to install	✓	✓	—	✓	Part of the SAS installation process
Input Data Format					
–Excel	—	✓	✓ ODBC	✓ ODBC	✓ Import Wizard
–CSV	✓	✓	✓	—	✓ Import Wizard
–SPSS	✓	✓ Wizard	✓ ODBC	✓ Using .xls or .dbf	✓ Using .xls or .dbf
–SAS	✓	✓ Wizard	✓ ODBC	✓ Using .xls or .dbf	✓
–.dbf	—	✓	✓ ODBC	✓ ODBC	✓ Import Wizard
View data table	✓ (executable)	✓ (default)	✓ (executable)	✓ (default)	✓ (default)
Sort data	✓	✓	?	Only one sorting criterion	✓
Simple descriptives*	✓	✓	✓	✓	✓
Variable binning	✓	—	✓	—	✓
Graphing:					
–Scatterplot	✓	✓	—	✓	✓
–Distribution	✓	✓	✓	—	✓
–Histogram	✓	✓	✓	—	✓
–Multiplot	✓		—	—	—
–Boxplot	—	✓	—	✓	✓
–Other graphic features?	Web chart, collection	Several (in Excel)	3D, map, grab, manipulate, drill down	3D	3D rotating plot Interpolation/Contour lines
Easy exports to SPSS, SAS, Excel, flat files	✓	✓ indirect	✓	✓ Through a flat file format	✓
Grade for help menu	A	B+	C	B–	A
Demos/Tutorials	✓			✓	✓

NOTE: * includes mean, standard deviation, minimum, maximum, count.

of the key features and provide more details.

We were able to obtain pricing information for most of the packages. An academic server license for Enterprise Miner is available for $40,000–100,000, and a mainframe license for $47,000–222,000. Commercial licenses cost $119,000–281,000 (mainframe $140,000–629,000). GhostMiner costs $2,500–30,000 plus undisclosed annual maintenance fees; or $10,000–75,000 for perpetual licenses, with the exact prices based upon the type of users-academic, government, or commercial. Nearly all Quadstone users are commercial, and licenses cost between $200,000 and approximately $1,000,000, plus annual maintenance fees, depending upon the number of users and number of customers to be analyzed. XLMiner is available in an educational version at $1,499 per site license or $49 each per student in class. The standard version two-year site license is available after July 2003 for $199 (academic) or $899 (nonacademic). The makers of Clementine did not disclose the cost of their software.

4. DESCRIPTIVE STATISTICS AND GRAPHICS

This section reviews the descriptive statistics and graphics capability that are commonly used to gain a better understanding of the data prior to more complex modeling procedures. We will discuss the packages in alphabetical order: Clementine (SPSS), Enterprise Miner (SAS), GhostMiner, Quadstone, and XLMiner.

4.1 Clementine (SPSS)

Clementine provides a useful set of descriptive tools enhanced by excellent graphics. Users accustomed to data streaming based upon icons will find this software very easy to work with provided they are aware of the need to connect icons by right-clicking the mouse. Those unaccustomed to using such icons will find Clementine fairly easy to learn through its several demonstrations; upon using the actual software, clarification can be obtained through an excellent help menu or by hovering the

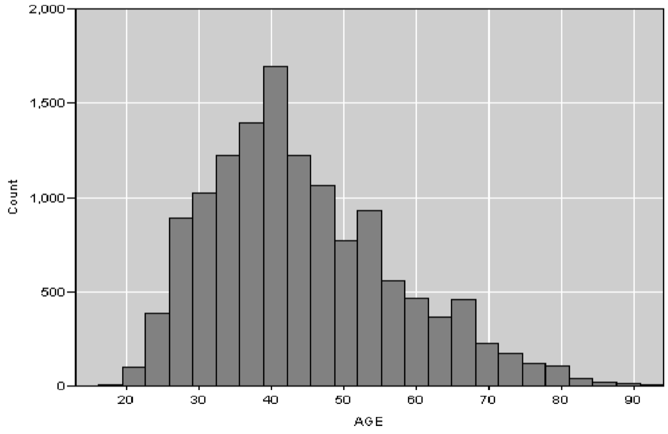


Figure 1. Histogram in Clementine.

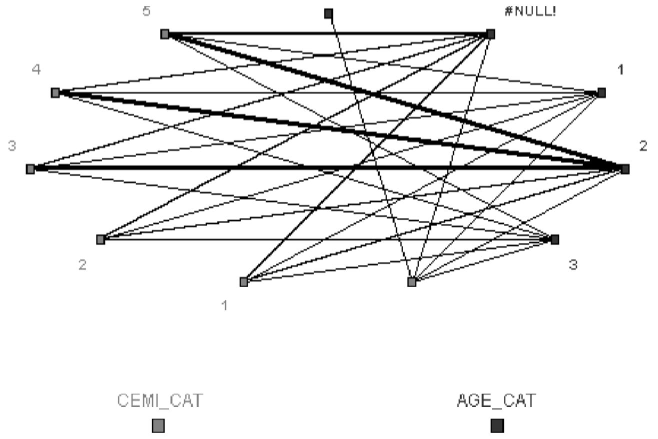


Figure 3. Web graph in Clementine.

mouse over the commands to invoke descriptions of the icons. Nearly all functions symbolized by icons are duplicated in the drop-down menus, an advantage to Clementine neophytes and more experienced users who can save time by recognizing and clicking on these simple icons.

Because Clementine is an SPSS package, importing SPSS format .sav data is straightforward, as is data in SAS or CSV format. However, data in Excel or .dbf format must first be converted. No explicit limit exists to the size of the file, a clear advantage over XLMiner, for example. The main menu bar at the bottom of the screen includes common operations under the “Favorites” tab; these operations are duplicated under subsequent logical menu tabs entitled “Sources”, “Field Ops”, “Record Ops”, “Graphs”, “Modeling”, and “Output”. The help menu is extremely useful in exploring options using logical keywords such as “import”, “export”, and “graph”. Missing values are identified by obtaining a data quality report in “Quality” node, and subsequent treatments can be executed easily in the “Type” node. To view the data, users must execute a table, found both in the Favorites and Outputs tabs. There, users can view 37 rows and several columns of data at a time on one screen. In all, the user interface is easy to understand and is self-explanatory. Clementine also features easy exports to SPSS, SAS, Excel, and flat file formats via nodes.

Several types of graphs are easy to create in Clementine, including scatterplots (simply called “plot”), distributions, histograms, collection, multiplots, and web plots, the latter illustrating coincidences of categorical values through the thickness of lines. The single conspicuously missing standard type of graph is the box plot. Clementine’s graphics are very good, and it is easy to produce a histogram (see Figure 1) and export it to one of several graphics formats (.jpg, .bmp, .png) as done here. Moreover, unlike in XLMiner, the process of specifying data classes (based upon percentiles, quartiles, quintiles, etc.) is straightforward in Clementine’s “Evaluation” feature, so this or any other graph could easily be altered.

The scatterplot in Figure 2 of two continuous variables provides another good indication of the strong graphics capabilities of Clementine. By virtue of the many output options, this or another graph could easily be imported into a Word, Powerpoint, or .html document for viewing without using excessive storage space.

An example of a web graph is shown in Figure 3. Among the five packages, this type of graph is unique to Clementine. The web graph represents a sort of graphical cross-tabulation in which thicker lines indicate relatively larger cell counts, and thinner lines the opposite. We found this to be a simple and useful technique for visualizing data.

Overall, Clementine is easy to learn and a pleasure to use for descriptive statistics and graphics. Among the software’s greatest advantages is its ability to bin data into user-specified groups (percentiles or any level of quantiles). Clementine’s ability to produce professional grade/publication quality graphics in compact formats is to be especially commended.

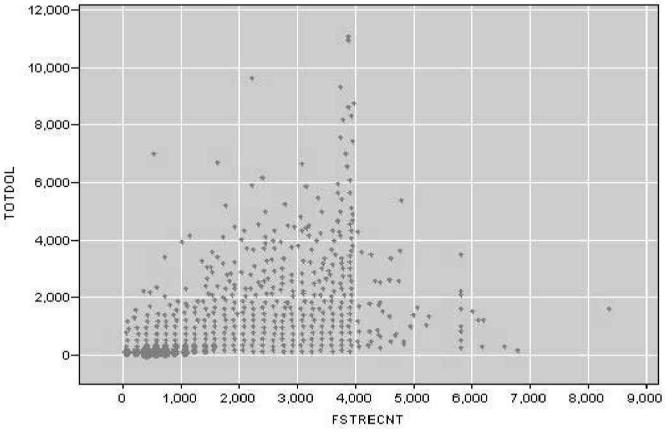


Figure 2. Scatterplot in Clementine.

4.2 Enterprise Miner (SAS)

The Enterprise Miner’s main tool for data visualization and descriptive statistics is the Insight module, which is a SAS system component. Insight offers a rich selection of tools for initial data analysis before the more complex modeling activities. Insight can be easily added to the analysis as one of the nodes in the Enterprise Miner network and, thus, it can accept not only raw data as input, but also data from data transformation modules (e.g., Replacement and Transform Variables).

Insight can be run either on the entire dataset or a random sample. In its tutorial, SAS recommends against loading very large datasets into Insight and suggests that a sample of 2,000 observations is sufficient for most purposes. In our tests, Insight was occasionally somewhat slow with our dataset of 19,185

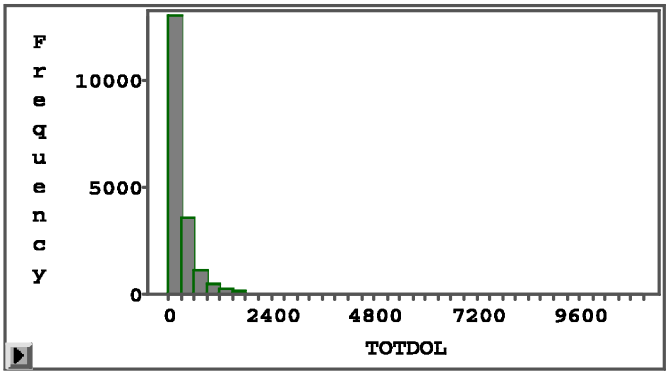


Figure 4. SAS Insight histogram.

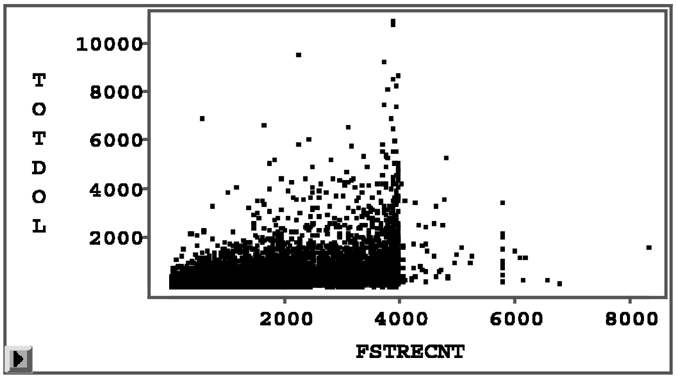


Figure 6. SAS Insight scatterplot.

observations and 296 variables, but otherwise performed flawlessly.

Initially, Insight presents the data to the user in a table format. The observations can easily be sorted by an unlimited number of variables either in ascending or descending order. The order of the variables can be changed easily, too. The tool provides a very comprehensive set of features for finding and evaluating specific observations. For example, it is possible to use the search tool to select a set of observations that satisfy a complex set of criteria and move them to a desired location in the table.

The main function of the Insight module is, however, to generate descriptive statistics and data visualization. As such, Insight provides a wide range of options that allows the user to explore the data from a number of perspectives. It can even quickly perform complex analyses prior to high-end modeling procedures. The available tools include: Histogram/Bar Chart (Figure 4), Box Plot/Mosaic Plot (see Figure 5), Line Plot, Scatter Plot (see Figure 6), Contour Plot, and Rotating Plot. While the first four tools offer a rich variety of options to modify the characteristics of the graph, the remaining two graph types allow three-dimensional representation of data. Contour Plots enable the user to visualize three-dimensional data in a two-dimensional space using contour lines (or equipotential curves) and associated color densities that represent a constant value of the dependent value. This technique is frequently used in weather maps. Another three-dimensional tool is the Rotating Plot, which allows the user to freely choose the perspective from which to analyze the patterns in three-dimensional data. Both three-dimensional tools are highly versatile and give the user a lot of freedom.

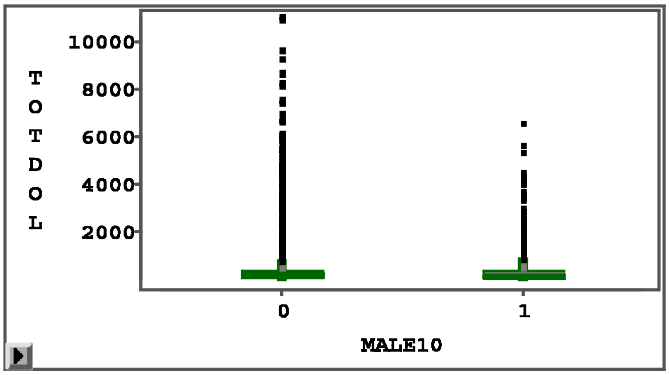


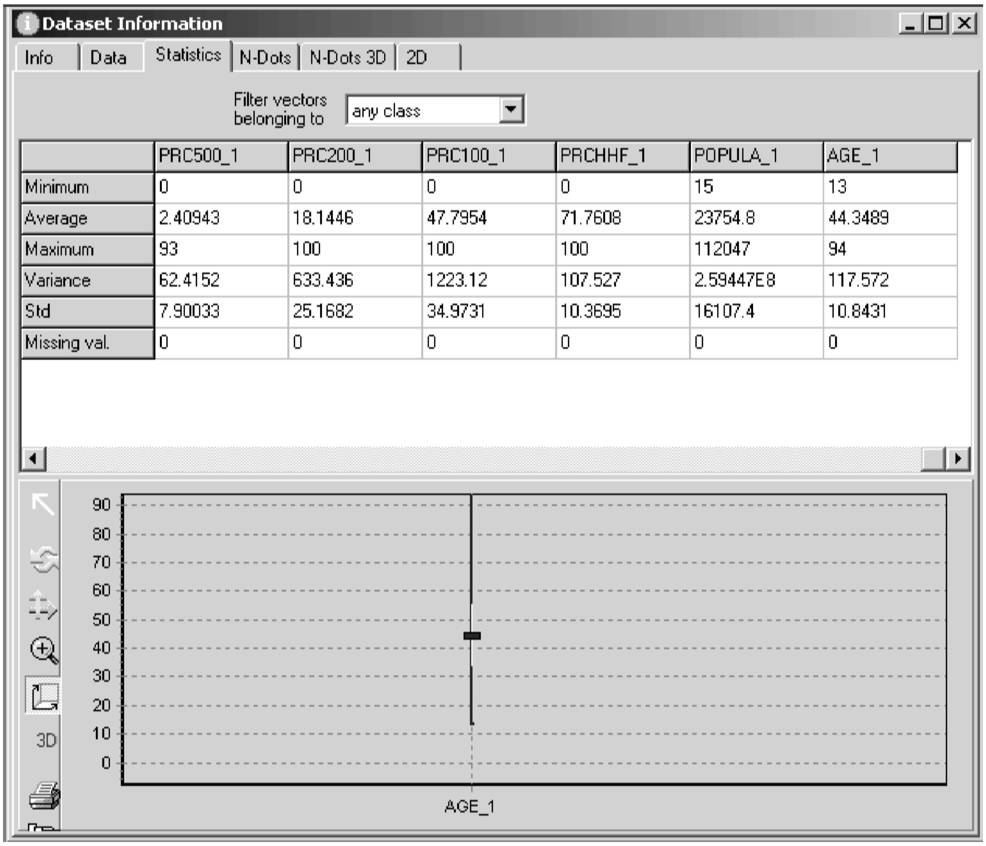
Figure 5. SAS Insight box plot.

Although the graphical tools are highly versatile, they suffer from user interface problems that are somewhat surprising in a high-end package such as Enterprise Miner. It seems that much of the code that implements the graphical user interface is still based on an interface development environment from the era when interfaces were mostly command-based and graphical elements were add-ons. There are relatively few opportunities for direct manipulation of the graphical elements and those that exist are awkward. The images are not visually attractive compared, for example, to those generated with some other packages in this review, or the graphical tools available in a tool such as Excel. The data visualization capabilities of Enterprise Miner are impressive, but they could still be significantly better if the tools to manipulate the characteristics of the charts and plots were easier to use. Two additional problems are nonstandard dialog boxes that determine the characteristics of the graphical elements and the cumbersome switching between the various options available through the Tools menu (regular pointing, moving a graphical object, and zoom).

In addition to the graphical tools, there are three excellent tools for quick analyses, including Distribution, Fit, and Multivariate. They all give the user a rich array of options for data analysis, including additional tools for graphical analyses. For example, one can select a variable from a table and then select the Distribution option from the Analyze menu. A user can create a Box Plot or a Histogram and then analyze basic descriptive statistics and the most important quantiles of the data. In addition, the number of observations available for evaluating and formally testing the distributional characteristics of the data is very impressive and definitely sufficient for most purposes associated with preliminary data analysis prior to data mining modeling.

Space constraints prevent us from including a comprehensive review of all the characteristics of the Fit tool, which offers a variety of both parametric and nonparametric methods for fitting curves and as such is an excellent tool for identifying trends and relationships in two-dimensional data. Moreover, the same tool can be used to run a wide range of analyses (including multiple regression, ANOVA, and ANCOVA) which rely on the least-squares method. In addition, Fit can be used for Logistic and Poisson regressions. A rich variety of graphical tools exist for analyzing residual and surface plots and fit curves.

Finally, the Multivariate tool is available for examining correlations, covariances, principal components, canonical discrimi-



Filter vectors belonging to: any class

	PRC500_1	PRC200_1	PRC100_1	PRCHHF_1	POPULA_1	AGE_1
Minimum	0	0	0	0	15	13
Average	2.40943	18.1446	47.7954	71.7608	23754.8	44.3489
Maximum	93	100	100	100	112047	94
Variance	62.4152	633.436	1223.12	107.527	2.59447E8	117.572
Std	7.90033	25.1682	34.9731	10.3695	16107.4	10.8431
Missing val.	0	0	0	0	0	0

Figure 7. GhostMiner Statistics view.

nant analyses, and evaluating relationships between two sets of variables with canonical correlation analysis and maximum redundancy analysis. In sum, Distribution, Fit, and Multivariate tools together offer a range of analytical tools that will be sufficient for exploratory analysis even for a demanding user. As with the graphical tools, the practical implementation of the user interface characteristics could be better, but the variety of analytical tools and their quality is excellent.

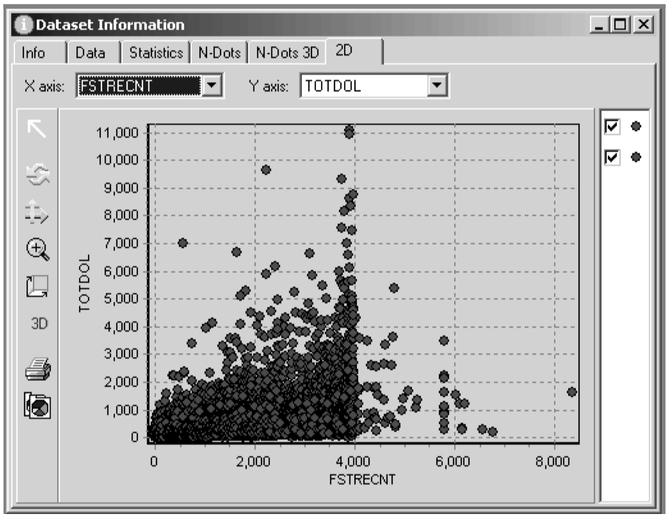
4.3 GhostMiner 1.0

GhostMiner's main tool for visualization and exploratory data analysis is the Dataset Information Window, the program's default view of the data. This window offers six different views of the data: Info, Data, Statistics, N-Dots, N-Dots 3D, and 2D.

Info provides an overall view of the data, including the numbers of cases (vectors in GhostMiner vocabulary), possible values of the dependent variable (classes), independent variables (features), and missing values. In addition, it gives the user both numeric and graphical representations of how the cases are distributed between the different classes. Data allows the user to view the data in a table format. The cases can be filtered based on the class to which they belong and sorted by any one variable (it does not seem to be possible to sort the cases based on multiple variables).

Statistics provides basic descriptive statistics (minimum, average, maximum, variance, standard deviation, and number of missing values) for all the independent variables and shows a box plot representation of the distributions of all the variables (see Figure 7). This feature works well with a small number of independent variables but, unfortunately, the mechanism breaks down when the number of independent variables increases. It is possible to zoom into a single variable or a subset of variables, but the results are neither visually attractive nor clear. This is unfortunate, because the tool is clearly useful with datasets that have only a small number of independent variables.

N-Dots, N-Dots 3D, and 2D are the primary descriptive visualization tools available in GhostMiner. N-Dots 2D/3D provides a mechanism for evaluating visually the distributions of each of the independent variables separately for each of the



X axis: FSTRECNT Y axis: TOTDOL

Figure 8. GhostMiner 2-D scatterplot.

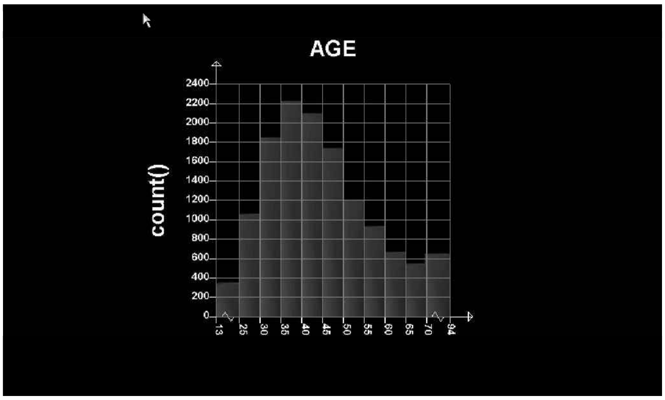


Figure 9. Histograms in Quadstone.

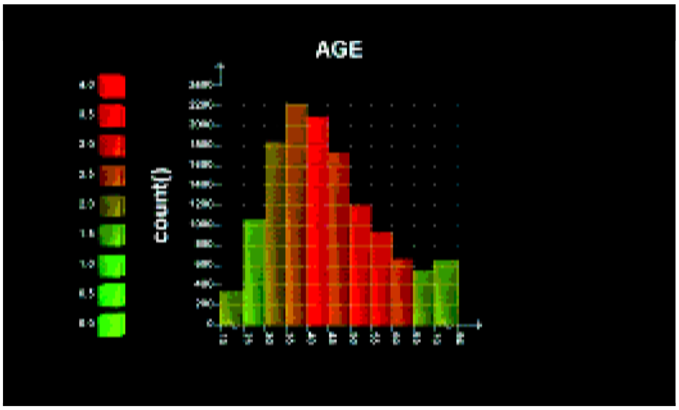


Figure 10. Color-enhanced histograms in Quadstone.

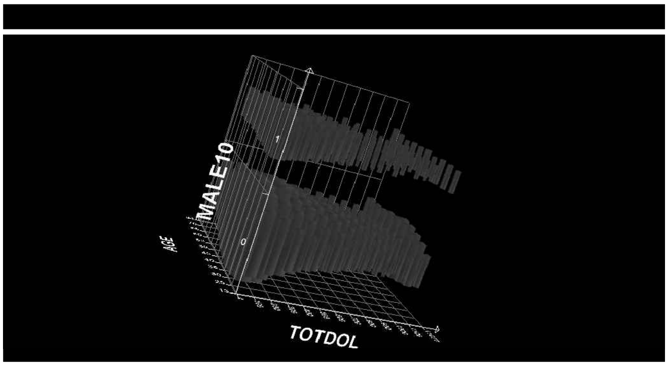


Figure 11. 3-D bar graphs in Quadstone.

classes, which in some cases can provide very useful insights just on the basis of visual inspection. We were, unfortunately, able to test N-Dots 2D/3D only with a small subset of our data because of capacity constraints and because this visualization technique does not lend itself well to datasets with a large number of predictive variables. 2D provides a fairly standard two-dimensional scatterplot in which the various classes are separated with different colors (see Figure 8).

Overall, GhostMiner provides a relatively modest set of tools for data visualization and analysis of the descriptive statistics. The usefulness of these tools is mostly limited to datasets with a small number of independent variables.

4.4 Quadstone

Quadstone's approach to exploratory graphical analysis and descriptive statistics differs from that of the applications that are built on the foundation of more traditional statistical packages. The exploratory analysis tools of Quadstone are available in the Decisionhouse module, and they include Crosstab Viewer, Crossdistribution Viewer, and a Profile Viewer. In addition, the Binning Editor can be used to view the distributions of the variables and change the methods used for categorization (binning). The Map Viewer function can be used to produce graphs that link data with various types of geographical images.

All Quadstone analysis tools are based on abstract cross-tabulations, which consist of one or more basic statistics (mean, minimum, maximum, etc.) for each of the virtual cells at the intersections of categories of the included variables. The different tools just provide different views of the same data. Crosstab Viewer shows the data in a table format, and both Profile and Crossdistribution Viewers allow the user to view the data graphically. Profile Viewer focuses on the characteristics of individual variables and Crossdistribution Viewer provides multidimensional views of the data.

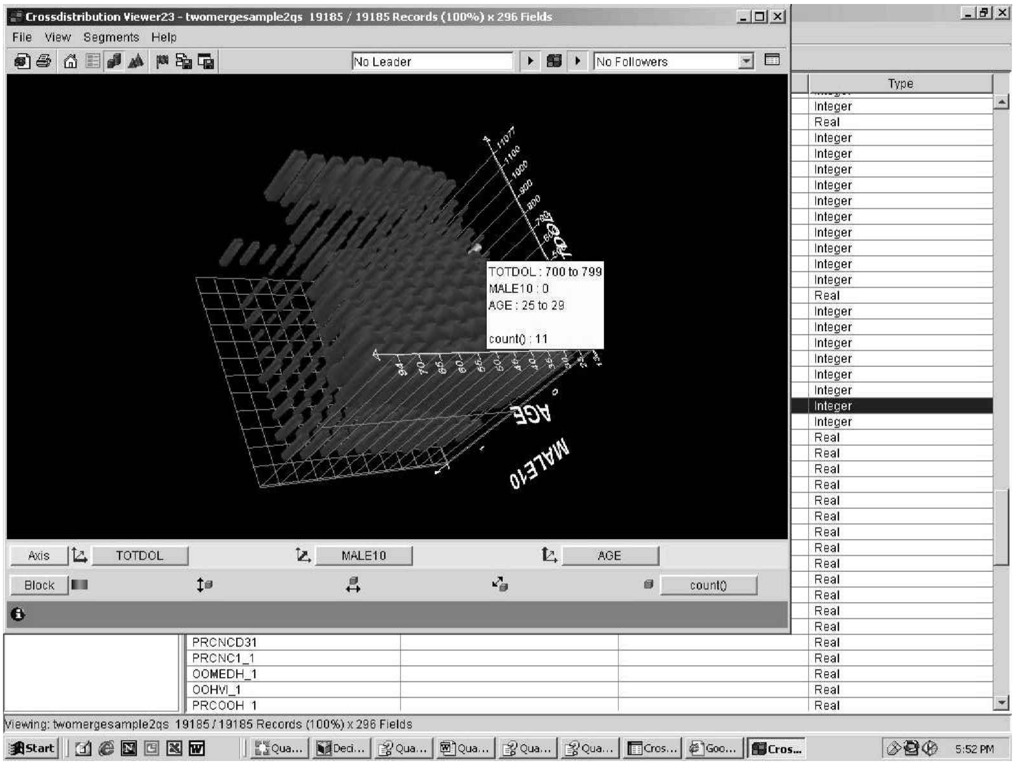
A simple example of the use of the Profile Viewer would be to analyze the age distribution within our sample (see Figure 9). This histogram was very easy to create with Profile Viewer, as was another histogram (Figure 10) which combines information about the average purchase amount (displayed with different colors) with the age distribution (displayed by the histogram). An example of the possibilities that are available in Crossdistribution Viewer is included in Figure 11; this bar graph shows the numbers of male and female observations in cells defined on the basis of age and the total number of dollars spent by the customer.

An additional powerful exploration feature in Quadstone allows the user to drill down by choosing a single cell (e.g., 25–29-year-old males whose life-to-date spending is between 700 and 799 dollars) using the graphical interface and applying any available analytical tool to that category of observations only (see the selection process in Figure 12).

Overall, Quadstone provides the user with an exploratory analysis module featuring a relatively small, but very well-designed, set of tools that make it possible to easily drill into the data.

4.5 XLMiner

XLMiner is an add-on program for Microsoft Excel that nests within an interface that users of Excel will find entirely familiar. Therefore, a license for Excel is prerequisite to installing this package. XLMiner enables the user to conduct an array of descriptive and graphical tasks. Although it is relatively inexpensive and easy to use—especially for those accustomed to other Microsoft products—it is also the most limited of the five packages in its ability to handle large datasets. Users can conduct descriptive operations on up to 200 columns and 2,000 rows at one time (partitioned), with a maximum file size of 6,000 records. This constraint required us to examine XLMiner with a subset of our database. Even after portioning the data, we found this package to drain system resources, precluding simultaneous use of other software, causing in some cases the



TOTDOL : 700 to 799
MALE10 : 0
AGE : 25 to 29

count() : 11

	Type
	Integer
	Real
	Integer
	Integer
	Integer
	Integer
	Integer
	Integer
	Integer
	Integer
	Integer
	Integer
	Real
	Integer
	Integer
	Integer
	Integer
	Integer
	Integer
	Integer
	Integer
	Real
	Real
	Real
	Real
	Real
	Real
	Real
	Real
	Real
	Real
	Real
	Real
	Real
PRCNCD31	Real
PRCNC1_1	Real
OOMEDH_1	Real
OOHVI_1	Real
PRCOOH 1	Real

Viewing: twomergesample2qs 19185 / 19185 Records (100%) x 296 Fields

Figure 12. Quadstone user interface.

user to reboot the computer.

Importing data into XLMiner is done with ease—an advantage of affiliation with its mainstream parent package Excel. We had no difficulty importing Excel, CSV, SPSS, SAS, and .dbf files either directly or by using Import Wizard. When finished importing the data, viewing, evaluating, and rearranging the data was easy to accomplish. Simple descriptives can be accessed in parent Excel (most conveniently using the Data Analysis toolkit). The single redundancy we found in XLMiner (when added to Excel) is the availability of Histograms both in the Excel Tools menu (under the Data Analysis toolkit) and under XLMiner. Data can be exported to several standard formats, but intermediate formats are needed to convert data to SAS and SPSS.

XLMiner's graphic capabilities are limited to histograms (see Figure 13), box plots (see Figure 14), and matrix plots. For all other types of graphs (including scatterplots), users can employ the Chart option within Excel, which offers several additional styles of graphs including scatter charts, pie charts, bar charts, and radar charts. This association with Excel is an advantage for XLMiner in comparison to Ghost Miner and Quadstone, which offer a much more limited array of charting options. A major disadvantage of the charting feature in Excel is that variables to be charted must be in preselected adjacent columns, requiring the user to cut-and-paste.

Before creating a simple box plot like the one shown in Figure 13, the user is required to treat all missing values, a procedure that repeatedly caused our computer to crash while work-

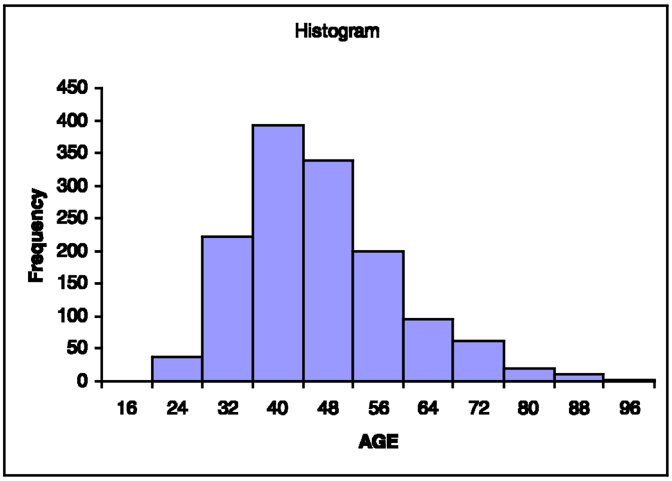


Figure 13. Histogram in XLMiner.

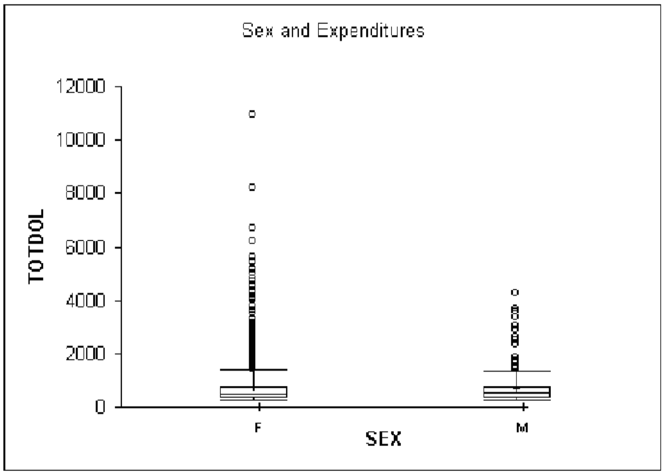


Figure 14. Boxplot in XLMiner.

ing on our database of 5,696 cases. Moreover, as a result of a handful of missing entries (1.6% of the dataset) the binary variable MALE01 could not be charted as desired until treatment was carried out. We found this treatment to be slow and rigid while using our dataset. It was not possible to treat only one column, but rather the entire dataset was treated over a period of about eight minutes, and records with *any* missing values in any columns were deleted. Moreover, the variable names did not appear in the window, but rather nondescript variable numbers based upon the columns. When attempting to change the options in the display, again the machine locked up and we were required to reboot the computer. The default treatment is "delete record", but users may also choose to replace missing values with the variable's mode or any user specified value. The absence of "mean" as an option is unfortunate, as was an apparent inability to select and treat only one column. The result of this challenge was the creation of the box plot (Figure 14) using the alternate categorical variable of "SEX" rather than "MALE10".

Scatterplots are not offered in XLMiner, but users can create them in Excel. To create a scatterplot, columns must be adjacent, and the default arrangement is that the first column is the X-axis and the second column becomes the Y-axis. This can be altered by changing the selected columns manually, or working specifying different columns in the chart wizard. Charts are easy to create in Excel provided the user is satisfied with the automatic binning procedures. Alternately, users can specify their own bins in a separate sheet in Excel. An advantage to the wide variety of graphs in Excel is that changes can be made by clicking on any part of the graph or its background. The corresponding disadvantage is that these graphics are large and require substantial space on the clipboard or in the resulting file-they can be pasted as .jpg or .bmp files, but then cannot easily be edited.

Overall, XLMiner would be most useful for users of databases of modest size. It is relatively easy to use but its capacity is limited. XLMiner is an excellent, inexpensive add-on that greatly expands the capabilities of Excel.

5. PREDICTIVE MODELS

This section describes our attempts to perform on each package an analysis that is fairly standard in database marketing-that is, building a predictive model for the response to an offer on the basis of a training (or analysis) dataset, and then evaluating it on a validation dataset. Because SAS Enterprise Miner (SAS EM) is the only package among the five considered which can automatically perform all steps of the analysis, we briefly describe the methodology using SAS EM output, and then move on to the other packages in alphabetical order.

5.1 Enterprise Miner

The main steps of the analysis can be visualized below in the SAS EM Diagram (see Figure 15). The node SASUSER.TWOMERGESAMPLE2 is the node identifying the data source, here a SAS dataset. The Data Partition node splits the dataset randomly into a training sample, a validation sample, and a test sample. The proportions of data in each sample are selected by the user, and in our case are chosen to be 40%, 30%, and 30%, respectively. The training sample is used by SAS to build a predictive model, which is refined in some cases on the basis of the validation sample. The performance of the model is then evaluated on data which did not intervene in the model building, namely the test dataset. In the Regression node, a logistic regression model (using stepwise selection) is built to predict who is more likely to purchase from the division D catalog (BUY10). In the Tree node, a decision tree which is very similar to a CART (classification and regression trees) tree, is built to the same effect. The performance of both models, in the form of lift charts, is provided in the Assessment node.

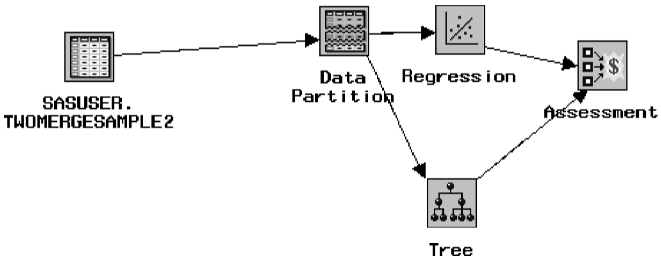


Figure 15. SAS EM diagram.

We briefly explain the decision tree process and refer the reader to, for example, Breiman, Friedman, Olshen, and Stone (1984) for a detailed exposition of the CART tree building methodology, or to the Salford-Systems Web site (www.salford-systems.com) for an overall introduction to CART and extensions of it such as multiple adaptive regression splines (MARS; Friedman 1991). See Deichmann et al. (2002) for a study of the performance of MARS in a database marketing context.

Essentially, CART splits the training dataset into two parts at each stage such that it reduces as much as possible the amount of impurity in the parts. Impurity (measured by a Gini coefficient) occurs when responders as well as nonresponders are present in a node of the tree. CART tries to split the data by considering rules of the form X ≤ C for each continuous variable X, and for each value C of that continuous variable, and for each possible arrangement into two sets of the levels or factors of a categorical variable, and by selecting that split which most reduces the impurity (of the two children nodes, compared to their parent node). Of course, trees tend to grow to the point where each observation ends up alone in a node (and where there are as many nodes as observations) if no pruning takes place. Very large trees tend to generalize very poorly to an independent validation dataset, so CART essentially prunes the tree to optimize the performance on a validation dataset.

As can be seen below (see Figure 16), SAS EM produced a simple tree after pruning, and decided to split the data according to the Total Number of Life to Date Orders from Division D. We can see that overall, 2.5% of the 7,674 people in the training sample, and 2.2% of the 5,756 people from the validation sample purchased from catalog D. Response is much higher among those people who had placed in the past at least 5 orders to date from catalog D, as one might expect (5.9% on the training file, 4.7% on the test file).

Several options are available in SAS EM for pruning the tree and selecting the final number of "terminal" nodes. We selected

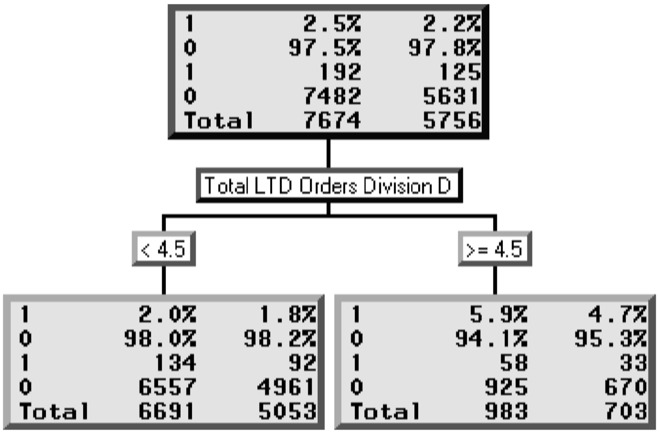


Figure 16. SAS EM tree.

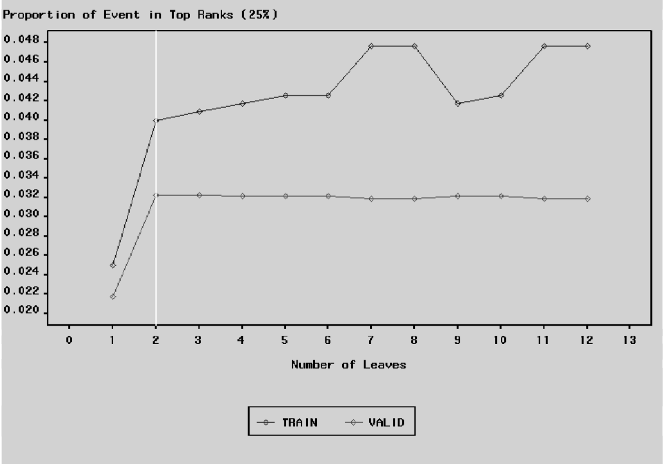


Figure 17. SAS EM tree plot.

to choose a number of nodes such that the proportion of responders in the best nodes covering up to 25% of the data was the highest. We can see below (Figure 17) that for the validation data, the proportion is as high as it will get after just two nodes. We can see a tendency for the proportion to rise up with the number of nodes in the training data, but these improvements do not hold up in the validation sample. This phenomenon is often referred to as "overfitting": a tree with seven nodes would "overfit" the data.

With the node Regression, SAS EM builds a stepwise logistic regression model for the logarithm of the odds of someone purchasing from catalog D. Among the available output, the most useful we have found is the traditional logistic regression output (familiar to statisticians). We have included in Figure 18 the latter part of the output, corresponding to the last model in the stepwise process.

The interpretation of this output is standard: for example, when other variables are held constant, one more past order from catalog D (to date) raises the estimated odds of purchase from catalog D by about 12% (because the odds ratio for the variable ORDLTDD is 1.117).

In the Assessment node, the performance of both models is evaluated. A common way of evaluating the performance of predictive response models is to sort a test file from the most likely to respond to the least likely to respond-as predicted by the model-then to divide the sorted file into, for example, deciles. If the model is performing well, one would expect the top decile to have a higher response rate than other deciles. The ratio of a response rate for a decile to the overall response rate for the whole file is commonly referred to as lift, and graphical representations of lifts (or equivalently of response rates) for all ten decile as lift charts. Cumulative lift charts are similar to lift charts, but response rates are evaluated on the top decile, then the top two deciles together, the top three deciles together, and so on.

Lift charts are provided by SAS EM on the training, validation, and test datasets. We can see (Figure 19) that the response

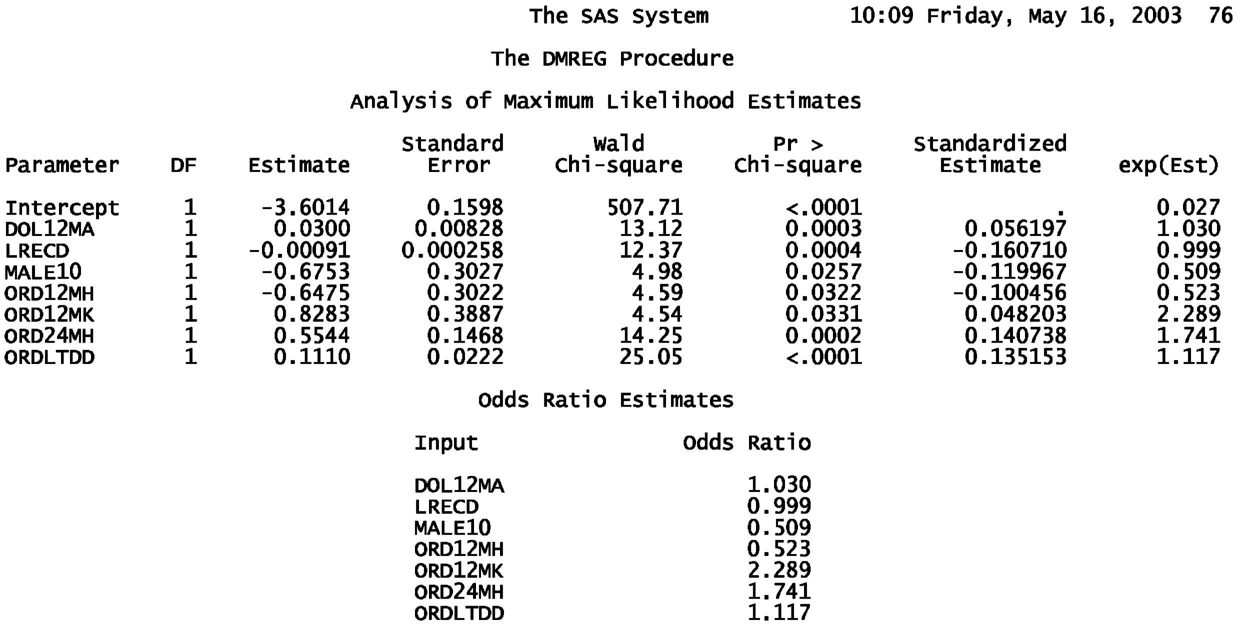
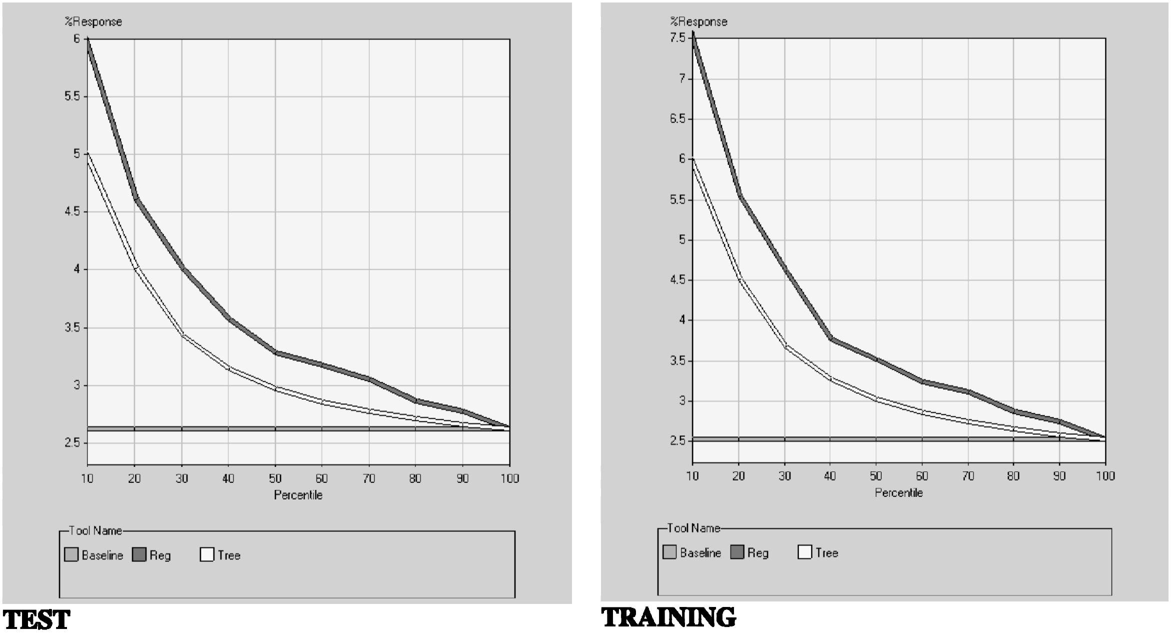

```
                                The SAS System         10:09 Friday, May 16, 2003  76

                              The DMREG Procedure

                       Analysis of Maximum Likelihood Estimates

                                  Standard       Wald          Pr >      Standardized
Parameter    DF     Estimate       Error      Chi-Square    Chi-Square     Estimate      exp(Est)

Intercept     1      -3.6014      0.1598        507.71        <.0001              .         0.027
DOL12MA       1       0.0300     0.00828         13.12        0.0003       0.056197         1.030
LRECD         1     -0.00091    0.000258         12.37        0.0004      -0.160710         0.999
MALE10        1      -0.6753      0.3027          4.98        0.0257      -0.119967         0.509
ORD12MH       1      -0.6475      0.3022          4.59        0.0322      -0.100456         0.523
ORD12MK       1       0.8283      0.3887          4.54        0.0331       0.048203         2.289
ORD24MH       1       0.5544      0.1468         14.25        0.0002       0.140738         1.741
ORDLTDD       1       0.1110      0.0222         25.05        <.0001       0.135153         1.117

                              Odds Ratio Estimates

                           Input                Odds Ratio

                           DOL12MA                   1.030
                           LRECD                     0.999
                           MALE10                    0.509
                           ORD12MH                   0.523
                           ORD12MK                   2.289
                           ORD24MH                   1.741
                           ORDLTDD                   1.117
```

Figure 18. SAS EM stepwise logistic regression output (partial).



TEST **TRAINING**

Figure 19. SAS EM cumulative lift chart for the test and training datasets.

rate in the top decile is slightly above 7.5% on the training file, and about 6% on the test file, representing a lift higher than 2 on the test file. Lifts on the test file tend to be less impressive than on the training file, but are more representative of lifts one might expect when applying the model to a new dataset, for which it is not yet known who the respondents are. One can infer from the charts that by using a logistic regression model and applying it to a dataset of prospective buyers, if one mailed a catalog to the top 20% of the file (sorted according to estimated probabilities of response), one might expect a response rate of about 4.6%. The lift charts reveal that lifts in the top deciles are higher when using the logistic regression model, as compared to the decision tree. The logistic regression model has more predictive power in the top deciles, although the decision tree is predictive in the lower deciles: one would lose fewer responders by dropping bottom deciles from a mailing using the decision tree, as compared to the regression model.

5.2 Clementine

The SPSS Clementine package has a lot of similarities with SAS EM, as can be seen on the Clementine stream in Figure 20.

The node twomergesample2.sav is the node which brings the dataset into Clementine, in the form of an SPSS dataset (.sav). The nodes Statistics and Table provide summary statistics and a view of the data, respectively. Training and validation files are created by the user, through a node which selects a subset of the data. We used a uniform random variable to select about

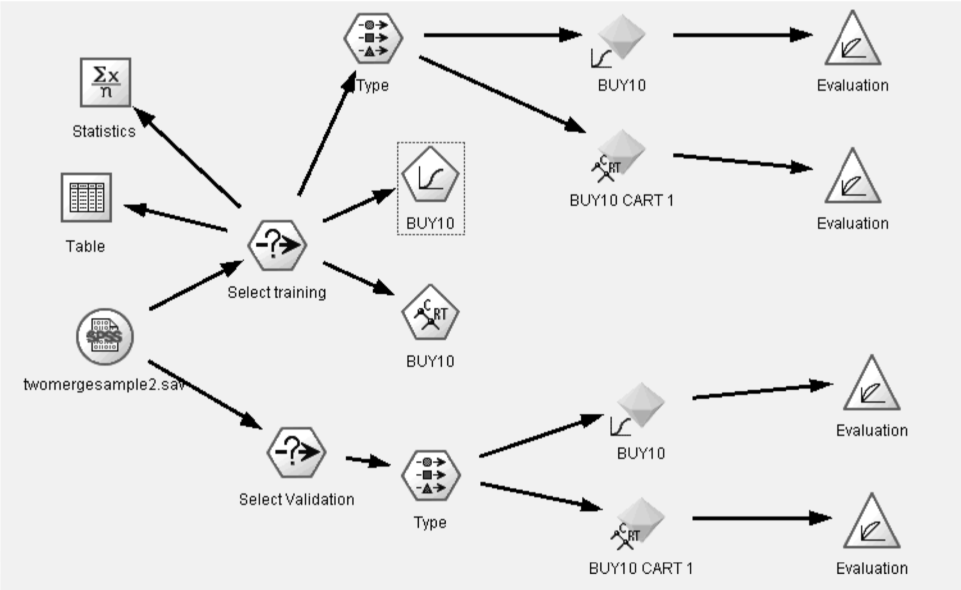
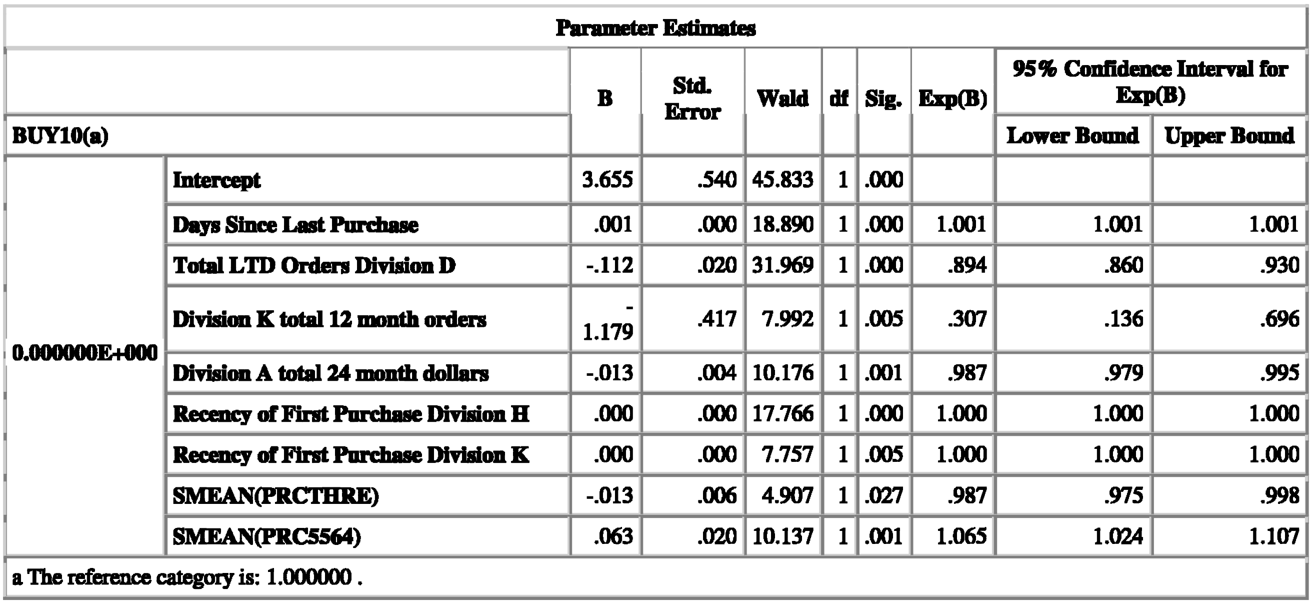


Figure 20. SPSS Clementine stream.



Parameter Estimates

		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
BUY10(a)								Lower Bound	Upper Bound
0.000000E+000	Intercept	3.655	.540	45.833	1	.000			
	Days Since Last Purchase	.001	.000	18.890	1	.000	1.001	1.001	1.001
	Total LTD Orders Division D	-.112	.020	31.969	1	.000	.894	.860	.930
	Division K total 12 month orders	-1.179	.417	7.992	1	.005	.307	.136	.696
	Division A total 24 month dollars	-.013	.004	10.176	1	.001	.987	.979	.995
	Recency of First Purchase Division H	.000	.000	17.766	1	.000	1.000	1.000	1.000
	Recency of First Purchase Division K	.000	.000	7.757	1	.005	1.000	1.000	1.000
	SMEAN(PRCTHRE)	-.013	.006	4.907	1	.027	.987	.975	.998
	SMEAN(PRC5564)	.063	.020	10.137	1	.001	1.065	1.024	1.107

a The reference category is: 1.000000 .

Figure 21. Clementine logistic regression output (partial).

60% of the file (11,450 cases) as a training file, and 40% of the file as a validation file (7,735 cases). Clementine successfully built a stepwise logistic regression model on the training file, but it took 1.5 hours with the hardware configuration we used. The logistic regression node is labeled BUY10 (for the name of the dependent variable) and marked with an icon featuring the graph of an S-shaped logistic function; the results of the model are presented in the yellow diamond with the same icon attached. The logistic regression output is given in Figure 21, and is similar to logistic regression output from standard statistical packages.

A peculiarity of this output is that the reference category was defined as category 1 (responders), so that in effect the model predicts the log of the odds of nonresponse; this is unusual, and the response rates in further evaluation nodes do indeed provide with proportion of responders, not proportions of nonresponders. This is of no serious consequence, but one needs to keep in mind that signs should be changed on all coefficients on the B column to predict response.

Clementine builds a CART decision tree quite rapidly (nodes are marked BUY10; yellow diamonds marked BUY10 CART 1 give results), producing the output presented in Figure 22.

We found no way of including in the output any more information about the nodes, such as number of observations, responders, and so on. In contrast, the output from a C4.5 tree (where an algorithm different from CART is used), which can be built with Clementine, is much more complete in general. In this particular case, however, a C4.5 option produces no tree, finding that all observations should be classified as nonresponders.

Decision trees (including CART, but excluding C4.5 trees) can be built with SPSS Answer Tree, a module of SPSS which is not part of this review, and the output from that module is much more complete, at least as far as CART trees are concerned.

Clementine provides lift charts, either in the form of response rates (cumulative or not), or of lifts. Because we found it quite awkward to modify the scale of the plots (1 – 100%, which makes for plots which occupy a small portion of the available space on the graph, and are hard to read), we present (see Figures 23 and 24) graphs with lifts from Clementine for both the training and validation file, and for both the logistic regression and decision tree models. Although it is quite straightforward to produce the lift charts for both models separately, there is, at least to our knowledge, no easy way to place the lift charts on the same graph for both models. We note in passing that the tree model gives very poor lifts; we did not dwell on this issue because our purpose is more to provide an overview of the packages and their features than to make a careful comparison of the predictive powers of the default models built by each of the packages.

5.3 GhostMiner

GhostMiner (Figure 25) provides a procedure to build a decision tree referred to as an SSV tree (separability of a split value). The algorithm here-the only one available-is different from the CART algorithm, but the splitting procedure is closely related to that of a CART tree. Algorithms differ in how the tree

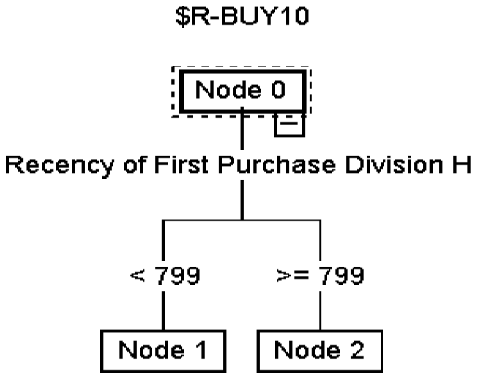
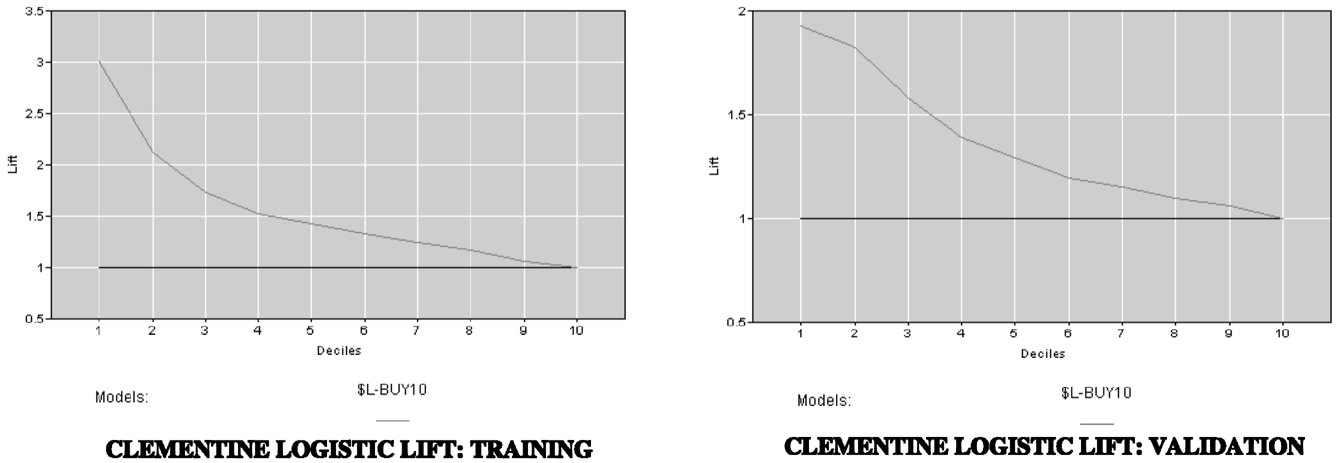


Figure 22. Clementine CART tree.



CLEMENTINE LOGISTIC LIFT: TRAINING

CLEMENTINE LOGISTIC LIFT: VALIDATION

Figure 23. Clementine lift charts for the logistic regression model.

is pruned, and how a final tree is selected. GhostMiner provides interesting algorithms, such as a neuro-fuzzy system algorithm which to the best of our knowledge is not readily available in the other reviewed packages. On the other hand, GhostMiner does not provide a logistic regression procedure.

Training and validation files must be created separately; only independent variables and the dependent variable must be included in the input dataset, because once the file is in GhostMiner, it is not possible to exclude any variable for consideration as an independent variable. We use the same training and validation files as used for Clementine.

We present GhostMiner output from an SSV tree (see Figure 26). GhostMiner splits the file according to the variable LSTRECNT, the number of days since last purchase (from any division of the catalog). More recent customers yield a higher response rate (3.7%, compared to 1.54% for less recent customers).

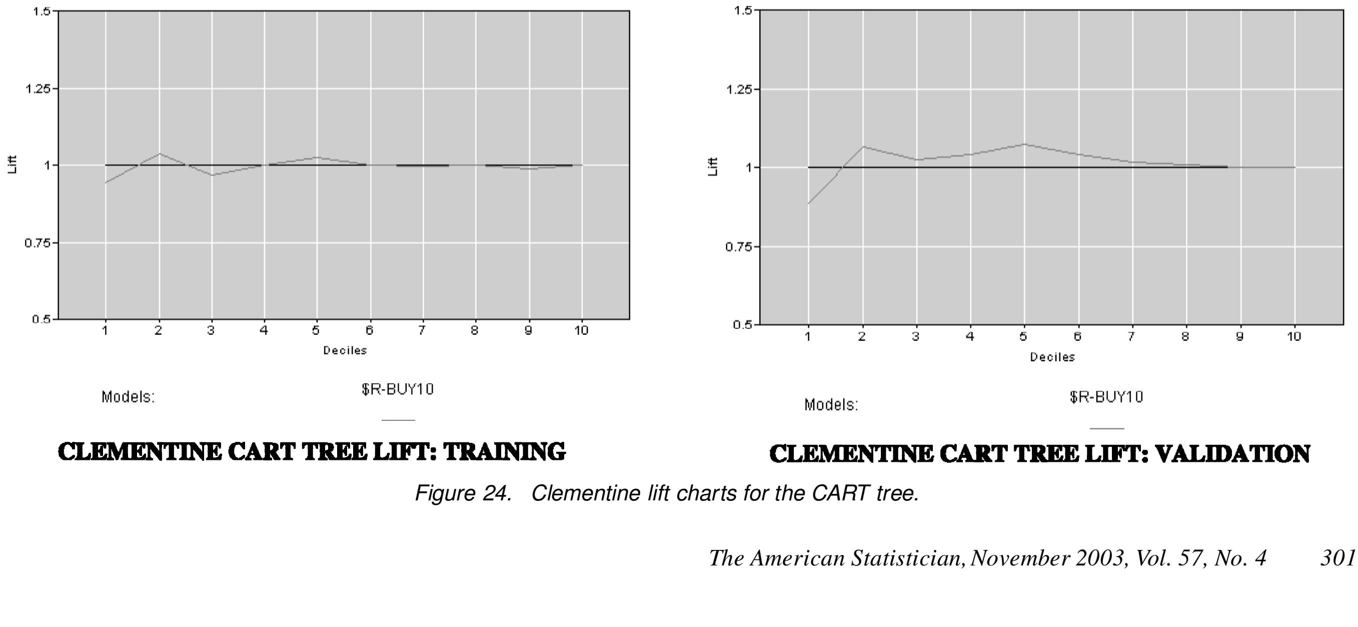
Deploying a model, here a decision tree, through the validation file is straightforward, and the results given are presented in Figure 27. Of course, for database marketing environments where response rates are often low, lift charts are perhaps more informative than confusion matrices. GhostMiner does not output lift charts or response rates on each node of the tree for the validation file.

5.4 Quadstone

In the Quadstone package, a project is referred to as a “focus”, and training and validation files must be created separately. We found it easiest to create two subfoci, with each subfocus containing the training and validation data, respectively. One inconvenience is that we found it impossible to rename the subfoci. However, it is quite easy to tell which subfocus is which, because the number of observations is given: 11,450 out of a full focus with 19,185 records. The user interface appears in Figure 28.

Quadstone provides a methodology to build a decision tree with a binary objective (yes/no response as in our case) using by default the ID3 algorithm (Quinlan 1993). The tree thus produced on the training file is presented in Figure 29.

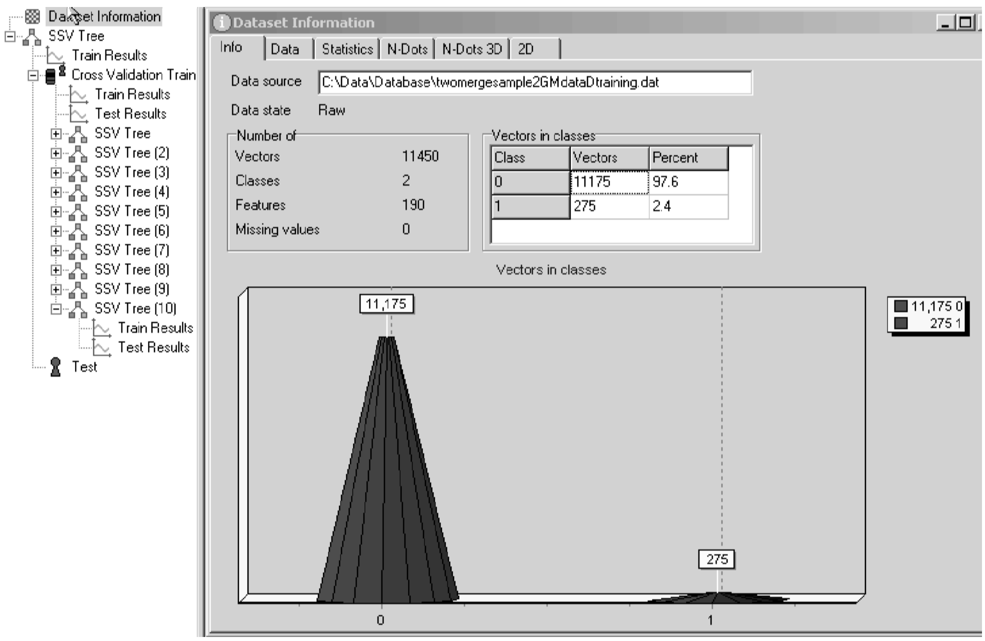
The output is easy to read, response rates are given for each node, and are color coded. The variables involved in the tree include NUMORDS (number of orders to date), LRECH (number of days since last purchase from division H), LSTRECNT (number of days since last purchase, from any division), TOTMSTRC (any use of MasterCard to date yes: 1 no: 0), INCMIN_1 (income index), PRNCDN_1 (percent households with one unit structures in the customer’s neighborhood). Quadstone provides a Gains Table, very similar to output given



CLEMENTINE CART TREE LIFT: TRAINING

CLEMENTINE CART TREE LIFT: VALIDATION

Figure 24. Clementine lift charts for the CART tree.



Data source C:\Data\Database\twomergesample2GMdataDtraining.dat

Data state Raw

Number of

Vectors	11450
Classes	2
Features	190
Missing values	0

Vectors in classes

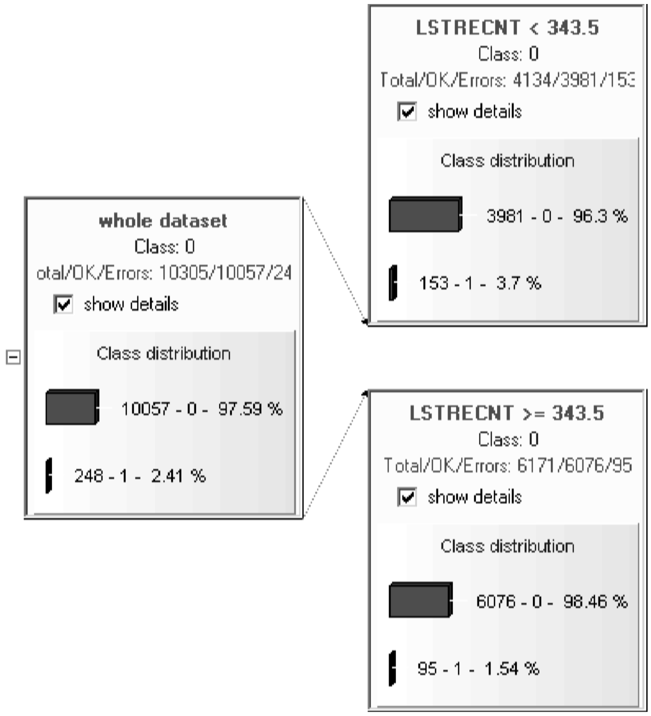
Class	Vectors	Percent
0	11175	97.6
1	275	2.4

Figure 25. GhostMiner user interface after building a decision tree on the training file.

in SPSS AnswerTree, presented in Figure 30.

The table is easily read; match rates are response rates. Note the very high (relatively) response rate in the relatively small node 12, which one might suspect not to hold up on the validation file.

It is a bit tricky, but feasible, with the Quadstone help menu to deploy the tree to the validation subfocus and to show the results of the tree on both the training and validation files, as presented in Figure 31.



whole dataset
Class: 0
otal/OK/Errors: 10305/10057/24
☑ show details
Class distribution
10057 - 0 - 97.59 %
248 - 1 - 2.41 %

LSTRECNT < 343.5
Class: 0
Total/OK/Errors: 4134/3981/153
☑ show details
Class distribution
3981 - 0 - 96.3 %
153 - 1 - 3.7 %

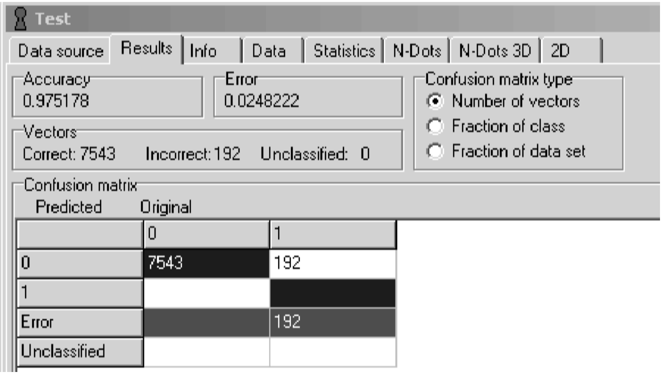
LSTRECNT >= 343.5
Class: 0
Total/OK/Errors: 6171/6076/95
☑ show details
Class distribution
6076 - 0 - 98.46 %
95 - 1 - 1.54 %

Figure 26. GhostMiner SSV tree.

It does appear on this output that the very high response rate of node 12 does not hold up on the validation file (12.61% on the training file, 5.48% on the validation file). The tree built by default by Quadstone is more complex than the trees built by either SAS EM, GhostMiner, and Clementine. There are ways to get Quadstone to prune the tree differently, but we let the tree be built as per default settings. The lift in the top decile does compare with the lift from, for example, the SAS EM tree.

The closest procedure we found to the traditional logistic regression approach is that of the scorecard model, for which we selected the "logistic" option. The output, presented in Figure 32, is very different from standard logistic regression output, and no lift charts are available, at least without further manipulation. An interesting-and powerful-feature is that all independent variables are automatically "binned", and the underlying logistic model used in the scorecard model takes this binning into account.

The Gini graph provided by the Quadstone scorecard output is constructed by sorting the cases in increasing order of model score (from the least likely to respond to the most likely to re-



Accuracy 0.975178

Error 0.0248222

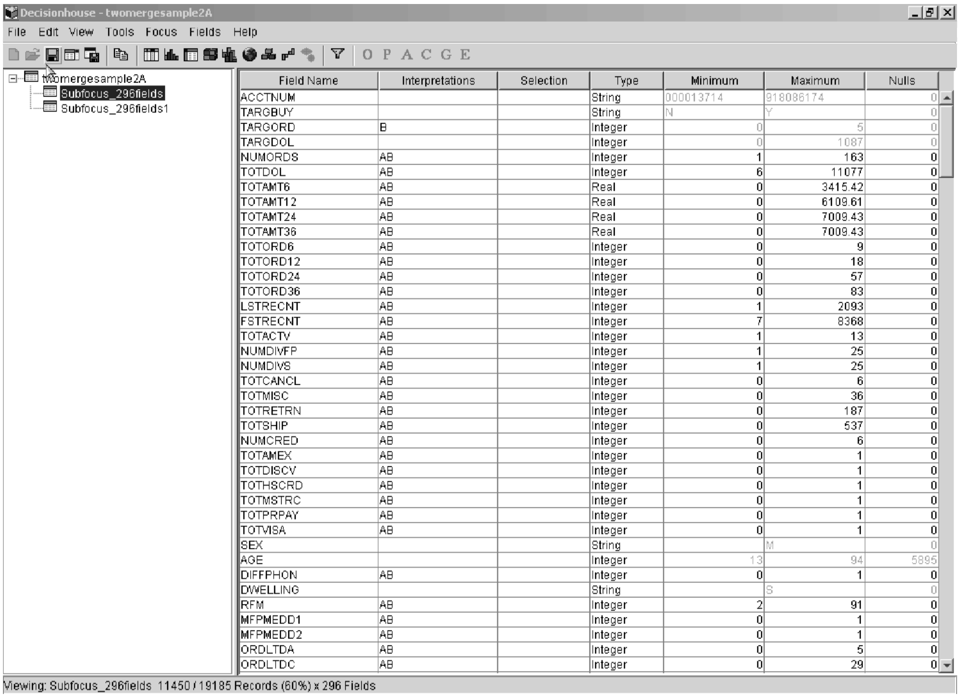
Vectors Correct: 7543 Incorrect: 192 Unclassified: 0

Confusion matrix type
● Number of vectors
○ Fraction of class
○ Fraction of data set

Confusion matrix
Predicted Original

	0	1
0	7543	192
1		
Error		192
Unclassified		

Figure 27. GhostMiner test results; decision tree deployed on the validation file.



Field Name	Interpretations	Selection	Type	Minimum	Maximum	Nulls
ACCTNUM			String	000013714	918086174	0
TARGBUY			String	N	Y	0
TARGORD	B		Integer	0	5	0
TARGDOL			Integer	0	1087	0
NUMORDS	AB		Integer	1	163	0
TOTDOL	AB		Integer	6	11077	0
TOTAMT6	AB		Real	0	3415.42	0
TOTAMT12	AB		Real	0	6109.61	0
TOTAMT24	AB		Real	0	7009.43	0
TOTAMT36	AB		Real	0	7009.43	0
TOTORD6	AB		Integer	0	9	0
TOTORD12	AB		Integer	0	18	0
TOTORD24	AB		Integer	0	57	0
TOTORD36	AB		Integer	0	83	0
LSTRECNT	AB		Integer	1	2093	0
FSTRECNT	AB		Integer	7	8368	0
TOTACTV	AB		Integer	1	13	0
NUMDIVFP	AB		Integer	1	25	0
NUMDIVS	AB		Integer	1	25	0
TOTCANCL	AB		Integer	0	6	0
TOTMISC	AB		Integer	0	36	0
TOTRETRN	AB		Integer	0	187	0
TOTSHIP	AB		Integer	0	537	0
NUMCRED	AB		Integer	0	6	0
TOTAMEX	AB		Integer	0	1	0
TOTDISCV	AB		Integer	0	1	0
TOTHSCRD	AB		Integer	0	1	0
TOTMSTRC	AB		Integer	0	1	0
TOTPRPAY	AB		Integer	0	1	0
TOTVISA	AB		Integer	0	1	0
SEX			String		M	0
AGE			Integer	13	94	5895
DIFFPHON	AB		Integer	0	1	0
DWELLING			String		S	0
RFM	AB		Integer	2	91	0
MFPMEDD1	AB		Integer	0	1	0
MFPMEDD2	AB		Integer	0	1	0
ORDLTDA	AB		Integer	0	5	0
ORDLTDC	AB		Integer	0	29	0

Viewing: Subfocus_296fields 11450 / 19185 Records (60%) x 296 Fields

Figure 28. Quadstone user interface and subfoci.

spond), then by going through the sorted cases and making one horizontal step when an actual responder appears and a vertical step when an actual nonresponder appears. In the ideal model, all responders would have lower scores than all nonresponders, so that all vertical steps would precede all horizontal steps. This would give a Gini graph in the shape of an upper triangle. The Gini value given in Quadstone output is the area between the diagonal line and the Gini graph, divided by 1/2 (the area of an upper or lower triangle). A perfect model would have a Gini value of 1. Note that this is a nonstandard definition of a Gini measure, and that the Gini graph is reminiscent of but not identical to a more standard ROC (Receiver Operating Characteristic) curve for logistic models.

The output in Figure 33 provides an idea of the predictive power of each variable separately, as measured by the Gini value from a model with that variable only. NUMORDS (number of orders) is the most predictive among the variables shown on the output, with a Gini value of 27.13%.

The output in Figure 34 provides the estimated contribution to the score of each bin for each variable: for example, 6 past orders add 15.42 to the base score of 387.38. For each variable and each bin, the number of responders and nonresponders is provided.

Even though we used the auto-include and auto-exclude options, which appear to approximate a stepwise process, it seems that many variables are involved in the calculation of the

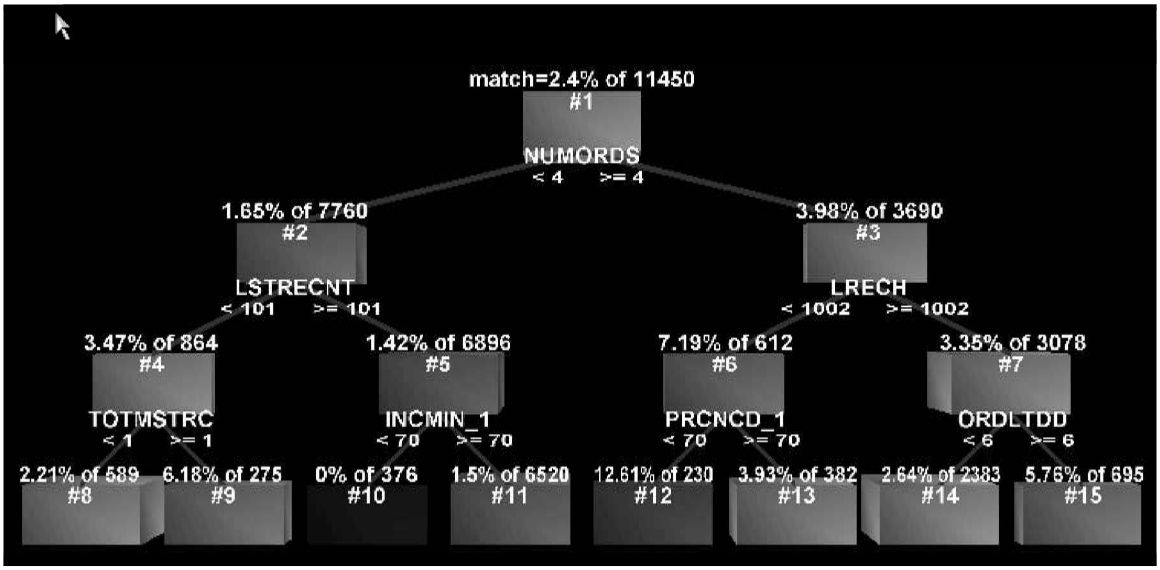


Figure 29. Quadstone decision tree (training file).

Segment	Match Rate (%)	Size	Size (%)	Cumulative Size	Accept Rate (%)	Reject Rate (%)
12	12.61	230	2.01	230	2.01	97.99
9	6.18	275	2.4	505	4.41	95.59
15	5.76	695	6.07	1200	10.48	89.52
13	3.93	382	3.34	1582	13.82	86.18
14	2.64	2383	20.81	3965	34.63	65.37
8	2.21	589	5.14	4554	39.77	60.23
11	1.5	6520	56.94	11074	96.72	3.28
10	0	376	3.28	11450	100	0

Match	Cumulative Match	Non Match	Cumulative Non Match	Cumulative Match Rate (%)	Segment Description
29	29	201	201	12.61	NUMORDS >= 4 AND LRECH < 1002 AND PRCNCD_1 < 70
17	46	258	459	9.11	NUMORDS < 4 AND LSTRECNT < 101 AND TOTMSTRC >= 1
40	86	655	1114	7.17	NUMORDS >= 4 AND LRECH >= 1002 AND ORDLTDD >= 6
15	101	367	1481	6.38	NUMORDS >= 4 AND LRECH < 1002 AND PRCNCD_1 >= 70
63	164	2320	3801	4.14	NUMORDS >= 4 AND LRECH >= 1002 AND ORDLTDD < 6
13	177	576	4377	3.89	NUMORDS < 4 AND LSTRECNT < 101 AND TOTMSTRC < 1
98	275	6422	10799	2.48	NUMORDS < 4 AND LSTRECNT >= 101 AND INCMIN_1 >= 70
0	275	376	11175	2.4	NUMORDS < 4 AND LSTRECNT >= 101 AND INCMIN_1 < 70

Figure 30. Quadstone gains table for decision tree (training file).

score, although the formula for the score is not evidently clear from the output.

5.5 XLMiner

As stated earlier, XLMiner is an add-on to MS Excel and, because it is an educational version, with the full version due to come out shortly at the time of writing, there are limitations on the number of variables to be used in the analysis (30) and the number of cases in the training and validation files (2,000). So our results for XLMiner are not immediately comparable to those of other packages, where we used full versions.

XLMiner provides an automatic way of splitting the file into training, validation, and test samples, which is convenient. Given a set of 28 independent variables, selected rather arbitrarily with some consideration as to which variables might be predictive of response, XLMiner easily built the tree presented in Figure 35.

The first split uses the variable RFM (recency frequency monetary, an index of how recently, how frequently, and how

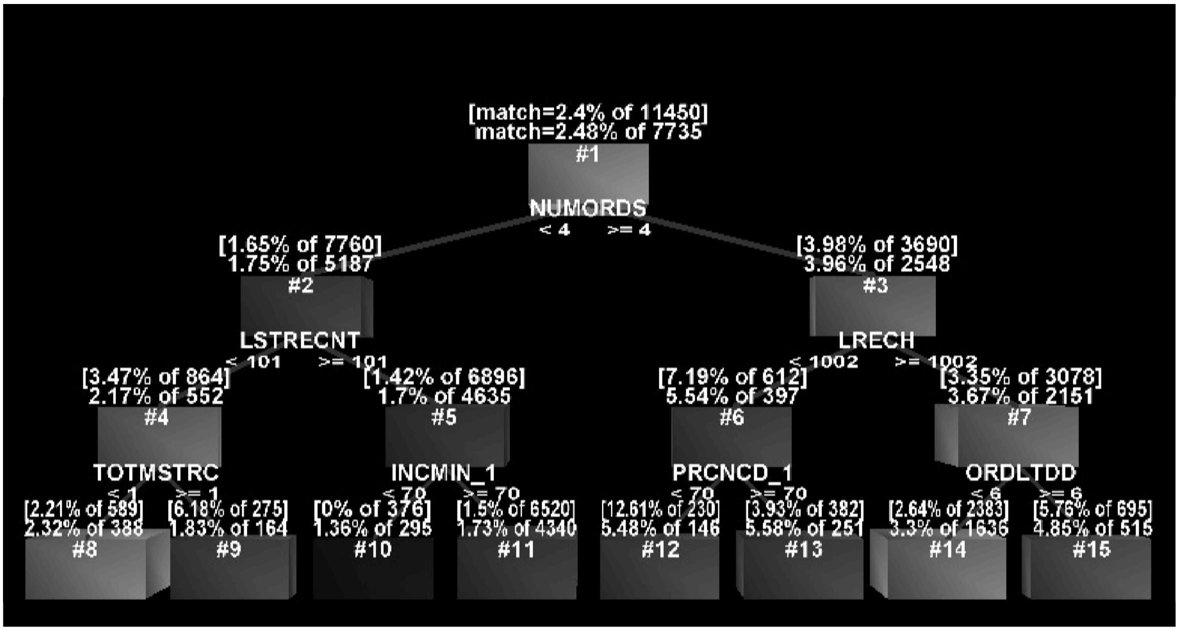


Figure 31. Quadstone tree built on training file and deployed on validation file.

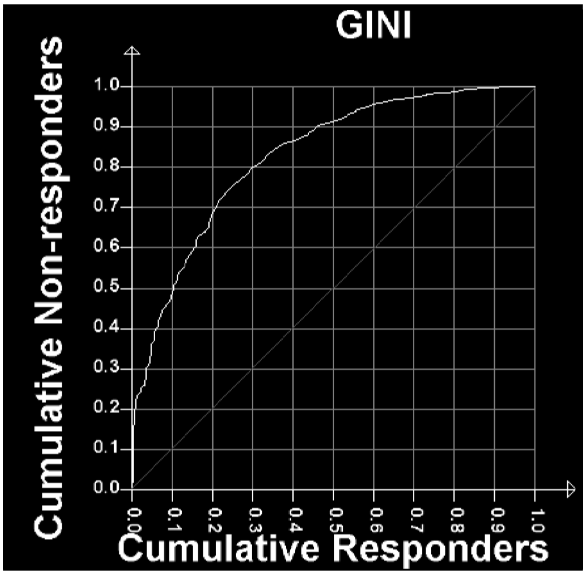


Figure 32. Quadstone scorecard model (logistic regression, auto-include, auto-exclude options); Gini graph.

much a customer has bought). Cases with RFM less than or equal to 10.5 move to the left (872 cases), and others move to the right (1,121 cases). The tree also uses variables ORDLTDD and ORDLTDH (number of orders to date from divisions D and H, respectively), as well as PRC556_1 (percent of people in customer's neighborhood aged 55 to 60). Nodes with zeros under them are terminal nodes, split no further; the percentages inside terminal nodes refer to the ratio of the number of cases in a particular node to the number of cases in the whole (training) file. The tree is more complicated than that of SAS EM, for example, and it is not clear how well the tree would hold up on the validation or test files. No lift charts are available for trees in XLMiner.

A logistic regression tool is available in XLMiner and, because we could not get the stepwise procedure to work properly with an input set of 30 variables, we moved the training file created by XLMiner to another package, ran a stepwise logistic regression on the training file and with the same input variables, and then reran the logistic regression with the variables selected by the other package, to yield the output in Figure 36.

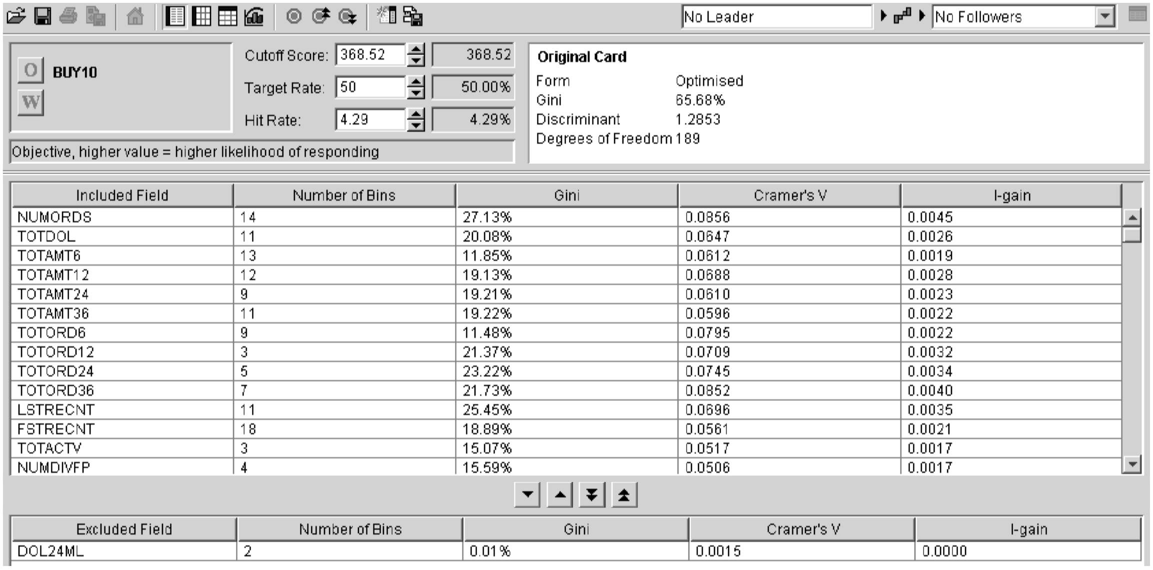
Lift charts are available for logistic regression in XLMiner, and are presented in Figure 37.

The decile-wise lift charts give the response rate for each decile of the file (training or validation) sorted in descending order of estimated probability of response. A good model would have steeply decreasing response rates as one progresses through the deciles. In the cumulative plots, the farther the graph is from the diagonal line, the more predictive the model.

6. ASSOCIATION ANALYSIS

Association analysis, or what is commonly referred to as market basket analysis, is one of the most popular techniques in database marketing and customer relationship management. In its most typical application, market basket analysis determines what products/services are purchased together by the consumers in a retail setting.

Patterns in the dataset are explained by rules. For example, the analysis may reveal that when items A and B are purchased item C is also purchased. In this case, A and B are called antecedents and C is the consequent. Any number of items can be antecedents or consequents. The output from association analysis includes three quantities that measure the degree of uncertainty associated with a given rule. Support, expressed as a percentage, is the probability that a randomly selected set of transactions from a database include items A, B, and C. Confidence, also expressed as a percentage, is the conditional probability that a randomly selected set of transactions will include C given that the transaction includes A and B. Finally, the analysis produces another measure of interest: lift. Lift is a value that mea-



BUY10

Cutoff Score: 368.52 368.52
Target Rate: 50 50.00%
Hit Rate: 4.29 4.29%

Objective, higher value = higher likelihood of responding

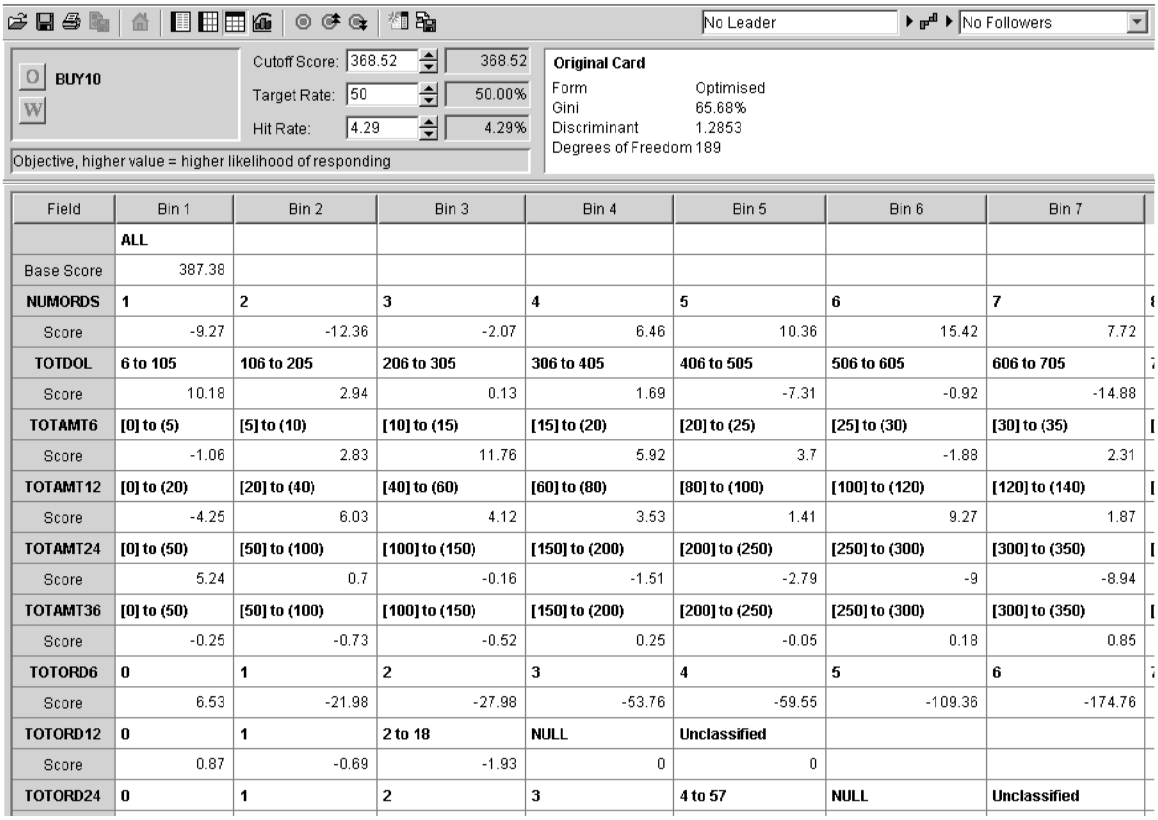
Original Card

Form	Optimised
Gini	65.68%
Discriminant	1.2853
Degrees of Freedom	189

Included Field	Number of Bins	Gini	Cramer's V	I-gain
NUMORDS	14	27.13%	0.0856	0.0045
TOTDOL	11	20.08%	0.0647	0.0026
TOTAMT6	13	11.85%	0.0612	0.0019
TOTAMT12	12	19.13%	0.0688	0.0028
TOTAMT24	9	19.21%	0.0610	0.0023
TOTAMT36	11	19.22%	0.0596	0.0022
TOTORD6	9	11.48%	0.0795	0.0022
TOTORD12	3	21.37%	0.0709	0.0032
TOTORD24	5	23.22%	0.0745	0.0034
TOTORD36	7	21.73%	0.0852	0.0040
LSTRECNT	11	25.45%	0.0696	0.0035
FSTRECNT	18	18.89%	0.0561	0.0021
TOTACTV	3	15.07%	0.0517	0.0017
NUMDIVFP	4	15.59%	0.0506	0.0017

Excluded Field	Number of Bins	Gini	Cramer's V	I-gain
DOL24ML	2	0.01%	0.0015	0.0000

Figure 33. Quadstone scorecard model (logistic regression, auto-include, auto-exclude options); Gini values for individual variables.



No Leader | No Followers

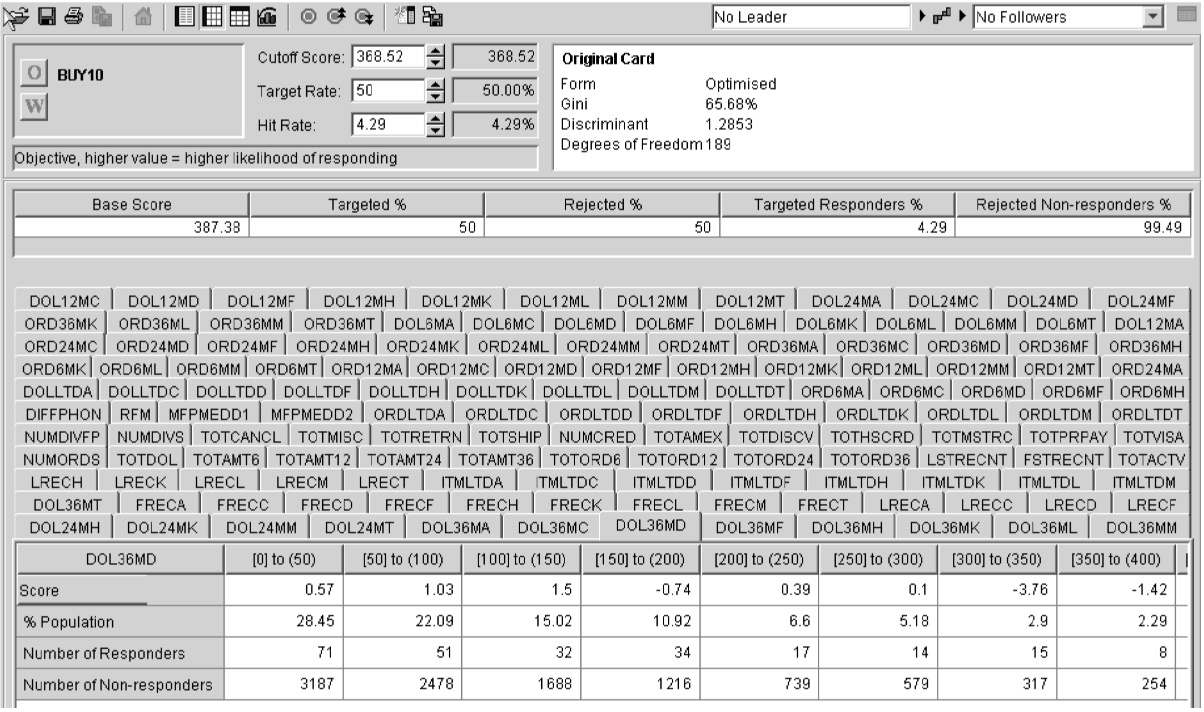
Cutoff Score: 368.52 | 368.52
Target Rate: 50 | 50.00%
Hit Rate: 4.29 | 4.29%

Objective, higher value = higher likelihood of responding

Original Card

Form	Optimised
Gini	65.68%
Discriminant	1.2853
Degrees of Freedom	189

Field	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6	Bin 7
	ALL						
Base Score	387.38						
NUMORDS	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Score	-9.27	-12.36	-2.07	6.46	10.36	15.42	7.72
TOTDOL	**6 to 105**	**106 to 205**	**206 to 305**	**306 to 405**	**406 to 505**	**506 to 605**	**606 to 705**
Score	10.18	2.94	0.13	1.69	-7.31	-0.92	-14.88
TOTAMT6	**[0] to (5)**	**[5] to (10)**	**[10] to (15)**	**[15] to (20)**	**[20] to (25)**	**[25] to (30)**	**[30] to (35)**
Score	-1.06	2.83	11.76	5.92	3.7	-1.88	2.31
TOTAMT12	**[0] to (20)**	**[20] to (40)**	**[40] to (60)**	**[60] to (80)**	**[80] to (100)**	**[100] to (120)**	**[120] to (140)**
Score	-4.25	6.03	4.12	3.53	1.41	9.27	1.87
TOTAMT24	**[0] to (50)**	**[50] to (100)**	**[100] to (150)**	**[150] to (200)**	**[200] to (250)**	**[250] to (300)**	**[300] to (350)**
Score	5.24	0.7	-0.16	-1.51	-2.79	-9	-8.94
TOTAMT36	**[0] to (50)**	**[50] to (100)**	**[100] to (150)**	**[150] to (200)**	**[200] to (250)**	**[250] to (300)**	**[300] to (350)**
Score	-0.25	-0.73	-0.52	0.25	-0.05	0.18	0.85
TOTORD6	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Score	6.53	-21.98	-27.98	-53.76	-59.55	-109.36	-174.76
TOTORD12	**0**	**1**	**2 to 18**	**NULL**	**Unclassified**		
Score	0.87	-0.69	-1.93	0	0		
TOTORD24	**0**	**1**	**2**	**3**	**4 to 57**	**NULL**	**Unclassified**



No Leader | No Followers

Cutoff Score: 368.52 | 368.52
Target Rate: 50 | 50.00%
Hit Rate: 4.29 | 4.29%

Objective, higher value = higher likelihood of responding

Original Card

Form	Optimised
Gini	65.68%
Discriminant	1.2853
Degrees of Freedom	189

Base Score	Targeted %	Rejected %	Targeted Responders %	Rejected Non-responders %
387.38	50	50	4.29	99.49

DOL36MD	[0] to (50)	[50] to (100)	[100] to (150)	[150] to (200)	[200] to (250)	[250] to (300)	[300] to (350)	[350] to (400)
Score	0.57	1.03	1.5	-0.74	0.39	0.1	-3.76	-1.42
% Population	28.45	22.09	15.02	10.92	6.6	5.18	2.9	2.29
Number of Responders	71	51	32	34	17	14	15	8
Number of Non-responders	3187	2478	1688	1216	739	579	317	254

Figure 34. Quadstone scorecard model (logistic regression, auto-include, auto-exclude options); scores and number of responders for bins of individual variables.

sures the improvement in probability of C occurring in a transaction given that the transaction includes A and B.

6.1 Applications of Association Analysis

Although association analysis is widely used in direct marketing and catalogue sales, it can be applied in other contexts. For example, it can be used to determine patterns in insurance claims submitted by patients. This will help insurance companies to not only identify medical procedures that are performed together, but also gain insights into possible fraudulent activity.

Regardless of the context, association analysis can offer several benefits to companies. First, association analysis may be

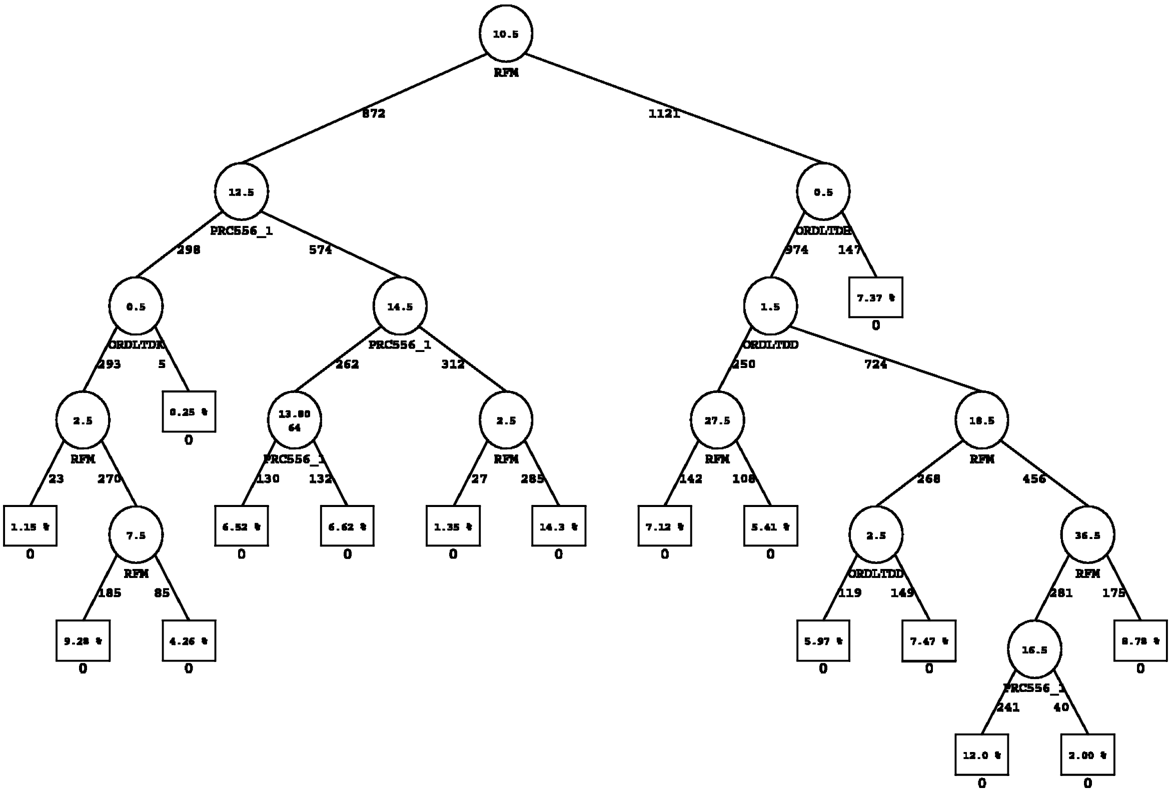


Figure 35. XLMiner Tree.

used to segment the customer base into similar “baskets.” In this application companies can monitor revenues from different basket segments and develop promotional campaigns for up-selling and cross-selling. Second, the brick-and-mortar retailers and catalogue companies can use association analysis to make decisions about product placement in the store and catalogue. Similarly, online businesses can benefit by identifying pages that are accessed together.

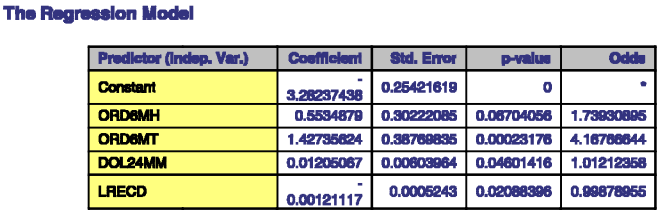
6.2 Package Review

Of the packages under review in this study, only three offered the association analysis procedure: SAS Enterprise Miner (SAS EM), SPSS Clementine, and XLMiner. We used the Bookbinder dataset from the Direct Marketing Educational Foundation to review these packages. Overall, we found the three packages to be quite similar in terms of ease of use, input data, and the output. However, we also discovered a number of differences that analysts must keep in mind.

User Friendliness. We found that all three packages are equally easy to use. XLMiner offers the familiar MS Excel format. SAS EM and Clementine offer a graphical user interface to run the procedure. All three offer the capability to set a minimum level of support and confidence. This is a huge advantage because in a typical database consisting of millions of transactions the total number of possible rules can be overwhelming and quite meaningless in many cases.

Input Data Format. XLMiner can handle two input data formats: binary matrix format (where each row of the matrix represents a customer, each column represents a product, and the matrix entries are ones and zeros indicating whether the product was purchased or not) and item list format (each row represents a transaction); SAS EM and Clementine can only handle the item list input format.

Output. All three packages calculate support and confidence parameters; SAS EM and XLMiner also calculate lift whereas Clementine’s output does not include lift. The methods of calculating these parameters are also different between the three packages. In SAS EM and Clementine, cases with no transactions are excluded from the analysis, but they are included in XLMiner. Another difference is the way the results are presented. We found the SAS EM presentation of results to be most intuitive as it lists all antecedents in a row followed by consequents. But, in XLMiner and Clementine, each antecedent is listed separately, which can make the output very long. All three packages offer the capability to sort the output measures by ascending or descending order. As an example of association analysis output, we have included the results from SAS EM (Figure 38).



The Regression Model

Predictor (Indep. Var.)	Coefficient	Std. Error	p-value	Odds
Constant	-3.28237438	0.25421619	0	*
ORD6MH	0.5534879	0.30222085	0.06704056	1.73930895
ORD6MT	1.42735624	0.38769835	0.00023176	4.16766644
DOL24MM	0.01205067	0.00603964	0.04601416	1.01212358
LRECD	-0.00121117	0.0005243	0.02088396	0.99878955

Figure 36. XLMiner logistic regression output.

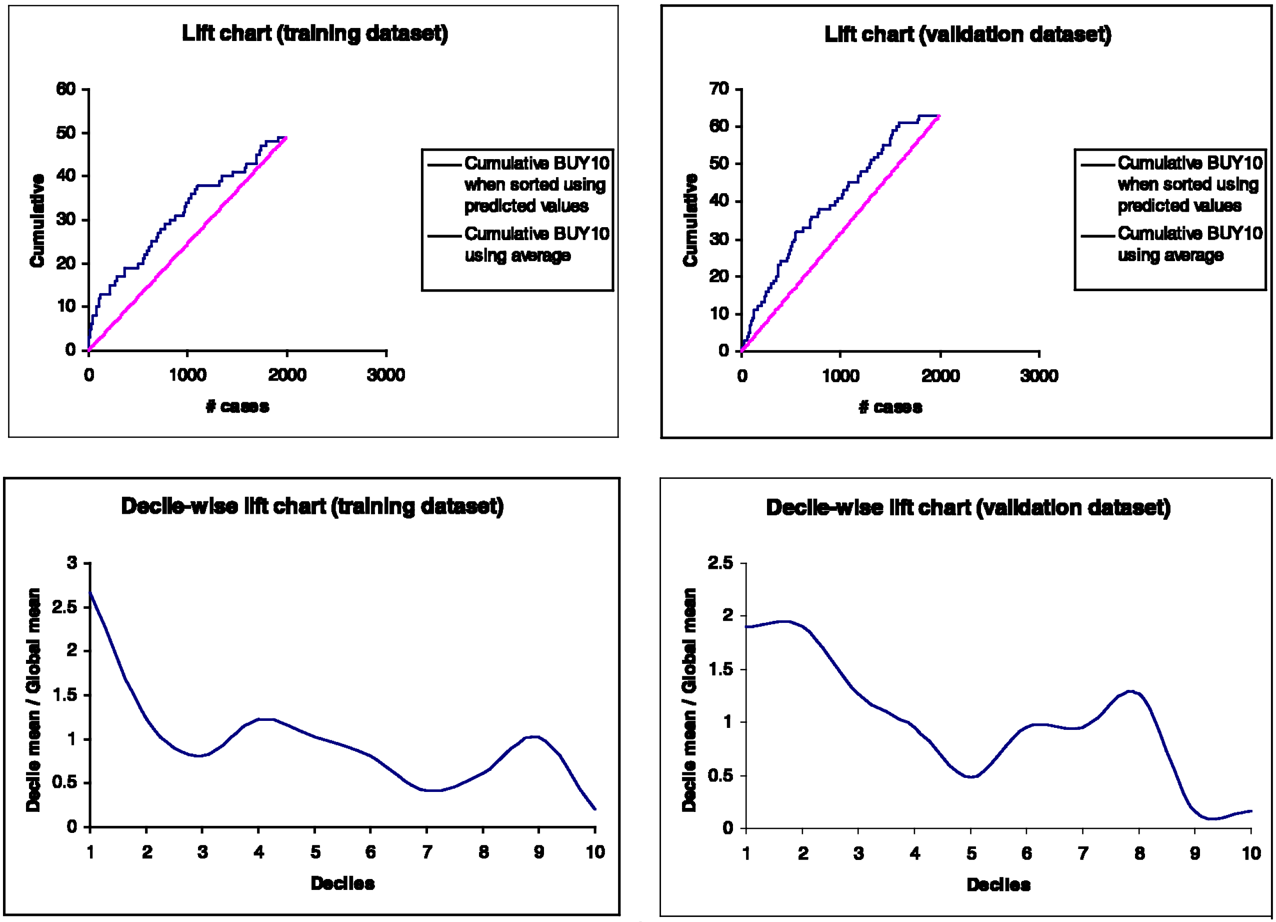
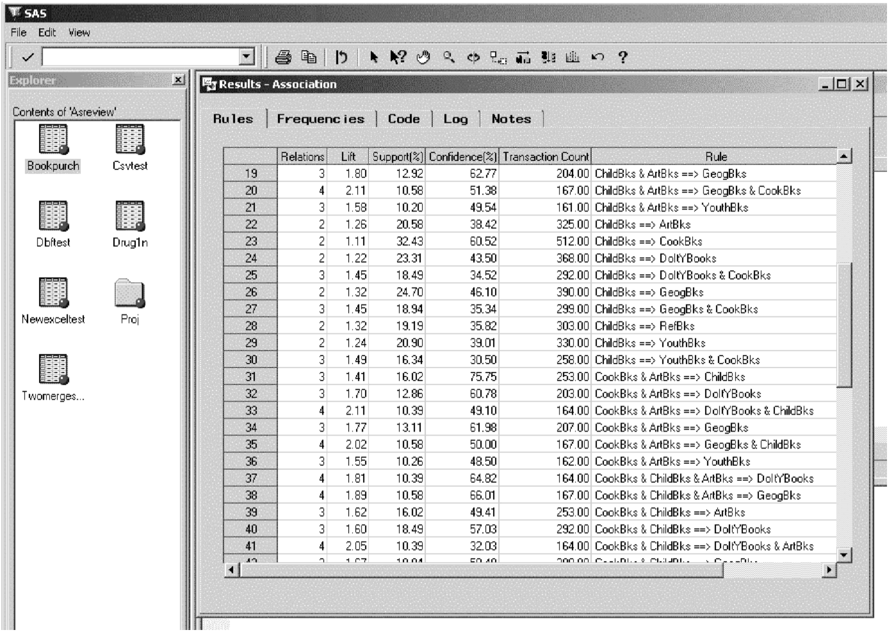


Figure 37. XLMiner lift charts for logistic regression.

7. CONCLUSION

Of all the packages, SAS EM is the most complete, although its graphics are not as attractive as those, for example, in Clementine. Quadstone has the most powerful graphics; in this package, all variables are binned and cross-tabulations can be represented in a variety of ways. XLMiner provides a respectable set of capabilities for a package with modest hardware requirements and low cost. Of course, file sizes will be limited to those fitting in Excel even in the upcoming professional version, which is a serious limitation for some applications. GhostMiner has interesting exploratory graphs, but they are



	Relations	Lift	Support(%)	Confidence(%)	Transaction Count	Rule
19	3	1.80	12.92	62.77	204.00	ChildBks & ArtBks ==> GeogBks
20	4	2.11	10.58	51.38	167.00	ChildBks & ArtBks ==> GeogBks & CookBks
21	3	1.58	10.20	49.54	161.00	ChildBks & ArtBks ==> YouthBks
22	2	1.26	20.58	38.42	325.00	ChildBks ==> ArtBks
23	2	1.11	32.43	60.52	512.00	ChildBks ==> CookBks
24	2	1.22	23.31	43.50	368.00	ChildBks ==> DoItYBooks
25	3	1.45	18.49	34.52	292.00	ChildBks ==> DoItYBooks & CookBks
26	2	1.32	24.70	46.10	390.00	ChildBks ==> GeogBks
27	3	1.45	18.94	35.34	299.00	ChildBks ==> GeogBks & CookBks
28	2	1.32	19.19	35.82	303.00	ChildBks ==> RefBks
29	2	1.24	20.90	39.01	330.00	ChildBks ==> YouthBks
30	3	1.49	16.34	30.50	258.00	ChildBks ==> YouthBks & CookBks
31	3	1.41	16.02	75.75	253.00	CookBks & ArtBks ==> ChildBks
32	3	1.70	12.86	60.78	203.00	CookBks & ArtBks ==> DoItYBooks
33	4	2.11	10.39	49.10	164.00	CookBks & ArtBks ==> DoItYBooks & ChildBks
34	3	1.77	13.11	61.98	207.00	CookBks & ArtBks ==> GeogBks
35	4	2.02	10.58	50.00	167.00	CookBks & ArtBks ==> GeogBks & ChildBks
36	3	1.55	10.26	48.50	162.00	CookBks & ArtBks ==> YouthBks
37	4	1.81	10.39	64.82	164.00	CookBks & ChildBks & ArtBks ==> DoItYBooks
38	4	1.89	10.58	66.01	167.00	CookBks & ChildBks & ArtBks ==> GeogBks
39	3	1.62	16.02	49.41	253.00	CookBks & ChildBks ==> ArtBks
40	3	1.60	18.49	57.03	292.00	CookBks & ChildBks ==> DoItYBooks
41	4	2.05	10.39	32.03	164.00	CookBks & ChildBks ==> DoItYBooks & ArtBks

Figure 38. Association analysis results from SAS Enterprise Miner.

suitable for small datasets mainly.

SAS EM was the only package where the full modeling analysis, from partitioning the data into training/validation files to building the model to drawing lift charts was possible automatically, with lift charts from various models available on the same graph. Clementine is quite similar to SAS EM, but a little more awkward to use for predictive modeling, although its interface is visually quite attractive. GhostMiner's modeling options are overall fairly limited (no logistic regression option), and its data manipulation tools are the least flexible of all the packages. GhostMiner's user interface is visually quite attractive. Quadstone has quite powerful modeling options but its output differs significantly from the standard, at least for the scorecard procedure, and some users may find this disconcerting. Quadstone has the capability of analyzing very large datasets, and its user interface is attractive.

All packages were straightforward to install except for Quadstone, which is intended to be installed by an expert. The most complete documentation, with an overall fairly clear description of the algorithms used, is found in SAS EM, with an added convenience of immediate availability from within the software. XLMiner's documentation is remarkably good, with useful examples of each tool. GhostMiner comes with a good user manual, with a good description of the algorithms. Clementine's documentation, such as available within the software, is also quite extensive, but does not usually provide details of algorithms. Quadstone comes with a good set of documentation, but this is available on the Internet separately from the software with a different username and password, although it is possible to download files from that source once to refer to them later off-line.

As expected, the packages affiliated with the most popular statistical software packages (SAS and SPSS) provide the broadest range of features with remarkably similar modeling and interface approaches, whereas the other packages all have their special sets of features and specific target audiences whom we believe each of the packages will serve well. It is essential that an organization considering the purchase of a data mining package carefully evaluate the available options and choose the one that provides the best fit with its particular needs.

[Received July 2003. Revised September 2003.]

REFERENCES

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Clementine, http://www.spss.com/SPSSBI/Clementine, accessed on June 25, 2003.

Deichmann, J., Eshghi, A., Haughton, D., Sayek, S., Teebagy, N. (2002), "Application of Multiple Adaptive Regression Splines (MARS) in Direct Response Modeling," Journal of Interactive Marketing, 16, 15—27.

GhostMiner, www.fqspl.com.pl, accessed on June 18, 2003.

Quadstone, www.quadstone.com, accessed on June 28, 2003.

Quinlan, R. (1993), *C4.5: Programs for Machine Learning*, Burlington, MA: Morgan Kaufmann.

SAS Enterprise Miner, www.sas.com/technologies/analytics/datamining/miner/, accessed on June 17, 2003.

XLMiner, http://www.resample.com/xlminer/index.shtml, accessed on June 22, 2003.