Preliminary Investigations into Knowledge Discovery for Quick Market Intelligence

William P. Alexander
Management Sciences and Statistics Program
Piero P. Bonissone
Lisa F. Rau
Artificial Intelligence Laboratory
GE Research and Development
Schenectady, NY 12301

1 Introduction

Quick Market Intelligence (QMI) is the process of identifying salient aspects of a business market to target sales more effectively. Retail businesses have begun to apply knowledge discovery techniques to their sales databases to quickly determine the optimal quantity of a certain product to stock, how well a sales promotion is doing, its location within a store, and its price, all given historical data obtained from previous sales. For example, a supermarket chain may stock certain food items only in certain stores, depending on the ethnicity or age of the local population, or may restock a new item just-in-time given greater-than-anticipated sales volume.

This paper reports on the results of applying existing software knowledge discovery methods (in particular, C4.5 and logistic regression) in support of a particular business application. The results, based on feedback from the GE business, indicate that this type of analysis is useful with very little special-purpose software development. We are currently using the results to better understand the tradeoffs between different discovery techniques, where extensions and research effort should best be spent, and how the results can be transitioned to solve real business problems and improve on existing business processes.

1.1 Application

This application involves a database of information used by a division of GE Capital Services (GE's Commercial Equipment Financing Division (CEF)). The database consists of over 400 kinds of information, used by the internal MIS organization for customer account management, as well as the marketing and sales force. The relevant portions of the database, identified through consultation with the database experts, consist of specific features of the client companies, specific features of the equipment that is being financed for them, and specific features of the financial instrument applied to the deal.

One task is to predict, given past sales, those types of companies and/or industries to target for special marketing efforts. This choice is guided by indications of growing or under-exploited market areas. We also want to detect correlations between certain types of financing arrangements and certain types of equipment, as well as relationships between financing arrangements and specifics of the companies, such as size. This type of information will allow the marketing and sales force to sell appropriate products to their customers. This general type of market analysis problem is applicable across a wide variety of businesses, and is not particular to the business of equipment sales and financing by any means. It draws heavily from related work in discovering rules from data (for example [3]).

In addition to this specific problem, we hope to better refine the information utility of the data through various analyses. For example, pre-determined geographic regional divisions may not accurately capture the data as well as divisions generated based on the data, i.e., New York may behave more like New England than the North Eastern states. Finally, exposing individuals with marketing and sales responsibilities to the types of analyses possible with state-of-the-art statistical and machine learning software facilitates the identification of the most useful data analysis that can be performed.

This paper describes our research program, the results of our first round of analyses in support of market research for this particular business application within GE Commercial Equipment Financing, and the next steps to be taken.

2 Research

Given that the data consisted of attribute-value descriptions, had pre-defined and discrete classes, and was of a relatively large size (over 50,000 cases), certain existing software tools seemed potentially applicable to the task. In particular, to approach the problem, we identified three candidate methods of data analysis: (1) the publically available package C4.5, (2) statistical methods and (3) a clustering program called Fuzzy Isodata. Each of these methods is suitable to the characteristics of the data just described. This paper reports on some of our results in applying C4.5 and statistical methods to the data - the analysis with the Fuzzy Isodata software has only just begun.

With this initial set of variables, we applied the C4.5 package and the statistical programs to the following tasks:

1. Given the magnitude of the deal in terms of dollars and the type of the equipment being financed, extract decision rules to predict the type of financial product to use (C4.5)

2. Given the entire subset of variables above minus MRKTPLN (marketing plan), extract decision rules to best predict the type of marketing plan to apply (C4.5)

3. Examine the data on a univariate and bivariate basis to assess data integrity, distributions and gross relationships between the variables (Statistics)

4. Given the magnitude of the deal and type of equipment (collateral), estimate the probabilities of the different products being used (Statistics)

The next section gives an overview of these analyses and results.

2.1 C4.5

The C4.5 package [4] is an implementation of a concept learning systems (CLS) that generates decision trees. The software was applied to a subset of the data. Although we have only begun to analyze the rules that have been generated through feedback with the domain experts, the software does find meaningful relationships between variables.

The domain experts initially identified a subset of variables, that describe three facets of the business: (1) characteristics of the financial "product" or business deal, (2) characteristics of the customer, including the geographical region of the customer and type of business of the customer, and (3) characteristics of the kind of commercial equipment being financed, "collateral".

Given that our goal in using C4.5 was that of summarizing the database, we were willing to accept an elevated percentage of misclassification for the sake of coverage. *Market Plan* is a variable used to classify the "go to market" strategy, for example, deals that were derived from referrals is a specific class of market plan. Another example of a market plan is the formulation of "private label programs" where GE CEF acts as the financier to a business so a customer can obtain financing at the place where the equipment is being purchased or loaned. The hypothesis is that there should be specific correlations between particular market plans and particular types of financial products. That is, when GE CEF is acting as a financier, they will typically be offering a certain type of financial arrangement (product). This hypothesis was born out.

In particular, in our prediction of *Market Plan* from a subset of 15 independent variables, including collateral, business code, region, and a variety of indicators, we obtained 12.5% errors in the training set and 13.3% errors in the test data. The percentage of errors almost doubled when we tried to predict *Product* from subsets of the same set of 15 variables: 21.8% in the training set and 25.4% in the test data. This percentage of errors was caused by the effort in generalizing rules to increase the data coverage and decrease the total number of rules. That is, we came across the inevitable tradeoff between the amount of pruning required for generating concise, summarized data and the inherent accuracy of the generated rules.

```
Rule    Used       Wrong         Advantage       Product
----    ----     --------        ---------       -------
   6      41      0 (0.0%)         41 (41|0)      DEALER
 550     122      0 (0.0%)        122 (122|0)     MEQMUN
 517     141     10 (7.1%)        128 (131|3)     NIMEQL
 604     304      4 (1.3%)        299 (300|1)     ELTOOL
 161      32      0 (0.0%)         32 (32|0)      ELTOOL
 602      68     18 (26.5%)        32 (50|18)     OPERLS
   1     857     11 (1.3%)        845 (846|1)     OFFICE
 640    2701      0 (0.0%)       2701 (2701|0)    TTIADB
 624     400     29 (7.2%)        366 (371|5)     TTIADB
 555      75     29 (38.7%)        41 (46|5)      FPFRRG
 636     298      5 (1.7%)          0 (0|0)       MENQSI
 471     154     10 (6.5%)          8 (17|9)      MENQSI
 552     201     13 (6.5%)        175 (180|5)     MENQSI
 634     285     94 (33.0%)        -1 (0|1)       MENQSI
 601      60     16 (26.7%)         0 (0|0)       MENQSI
```

655	6581	142 (2.2%)	6378 (6397|19)	MEREG
618	120	11 (9.2%)	101 (109|8)	MEREG
267	936	194 (20.7%)	684 (742|58)	MEREG
600	1888	620 (32.8%)	1088 (1268|180	MEREG
25	1925	37 (1.9%)	2 (3|1)	SGLINV
643	462	24 (5.2%)	63 (64|1)	SGLINV
616	1237	43 (3.5%)	1087 (1111|24)	SGLINV
480	3915	291 (7.4%)	22 (31|9)	SGLINV
467	268	79 (29.5%)	8 (10|2)	SGLINV
635	105	50 (47.6%)	17 (55|38)	SGLINV
557	18610	8449 (45.4%)	7720 (10161|24)	SGLINV

```
Tested 48052, Errors 12182 (25.4%)   <<
```

TABLE I: Evaluation on test data (48052 items) using all rules and the default class.

Clearly, several tradeoffs are available. We could produce a large number of more specific rules, we could allow disjunctions as values for the predicted variable, or we could use only those rules exhibiting higher accuracy in the training set. In the latter case, we would only use rules whose actual percentage of error in the training set was less than 10%. Under these conditions, we would not use the default rule to classify the records not covered by this reduced rule set. We would rather postpone the classification, by noting that less reliable rules were covering the case and suggesting additional verification (using for instances the statistical techniques described above).

In our example, we decide to drop the six least reliable rules used to predict *product*, without using the default classification of MENQSI. MENQSI is an indicator of how the leasee puts the lease on the books - either as an operating or a capital lease. How this is done effects who is allowed to take the depreciation benefit. Moreover this type of product is considered a tax lease from an IRS-perspective. It is a straight promissary note and hence is a "quasi" loan. The other financial products contain equally technical distinctions and will not be further defined; their "semantics" is highly specific to the business and difficult to understand without some background in finance.

As a result we dropped from 25.4% errors to 4.07% errors, at the expense of dropping our coverage to only 41.8% of the data.

Rule	Used	Wrong	Advantage	Product
6	41	0 (0.0%)	41 (41|0)	DEALER
550	122	0 (0.0%)	122 (122|0)	MEQMUN
517	141	10 (7.1%)	128 (131|3)	NIMEQL
604	304	4 (1.3%)	299 (300|1)	ELTOOL
161	32	0 (0.0%)	32 (32|0)	ELTOOL
1	857	11 (1.3%)	845 (846|1)	OFFICE
640	2701	0 (0.0%)	2701 (2701|0)	TTIADB
624	400	29 (7.2%)	366 (371|5)	TTIADB
636	298	5 (1.7%)	0 (0|0)	MENQSI
471	154	10 (6.5%)	8 (17|9)	MENQSI

552	201	13 (6.5%)	175 (180|5)	MENQSI
634	285	94 (33.0%)	-1 (0|1)	MENQSI
601	60	16 (26.7%)	0 (0|0)	MENQSI
655	6581	142 (2.2%)	6378 (6397|19	MEREG
618	120	11 (9.2%)	101 (109|8)	MEREG
25	1925	37 (1.9%)	2 (3|1)	SGLINV
643	462	24 (5.2%)	63 (64|1)	SGLINV
616	1237	43 (3.5%)	1087 (1111|24	SGLINV
480	3915	291 (7.4%)	22 (31|9)	SGLINV
467	268	79 (29.5%)	8 (10|2)	SGLINV

```
Tested 48052,  Covered:        20,104  (41.83%)     Errors:  819  (4.07%)
               Not Covered:    27,948
```

TABLE II: Evaluation on test data (48052 items) after eliminating six rules (number 602, 555, 267, 600, 635, 557) and the default classification.

The following are examples of the kind of relationships that were discovered through the application of the program:

```
Rule 640:
COLLCDE in {48000000, ..., 33000000} ; Note: Value set contains 303 elements
     BUSCODE in {C800}
     REG in {JCBREG, YALERG}
->  class TTIADB  [99.5%]

Rule 655:
     COLLCDE in {31000000, ..., 7000000} ; Note: Value set contains 15 elements
->  class MEREG  [99.6%]

Rule 616:
     GP03NUM <= 3787
     ASFMVIND in {3, 4}
     REG in {JCBREG, YALERG}
->  class SGLINV  [90.0%]
```

From Table I and II let's repeat, for the reader's convenience, the rows corresponding to rules 640, 655, and 616.

Rule	Used	Wrong	Advantage	Product
640	2701	0 (0.0%)	2701 (2701|0)	TTIADB
655	6581	142 (2.2%)	6378 (6397|19)	MEREG
616	1237	43 (3.5%)	1087 (1111|24)	SGLINV

Rule 640 above tests 3 of the 15 variables. It indicates that if the Collateral is one of 303 values, the business code is C800, and the region is either JCB or YALE, then the Product class is TTIADB. This rule was originally a path in a decision tree. However, part of the path was pruned to decrease the number of variables to be tested (in this case from 15 to 3). This generalization introduces an expected error [4] (pages 40-41). In the case of rule 640, such expected error is 0.5%. However, when applied to the test data, rule 640 correctly classified 2701 records, without any misclassification.

Rule 655 shows that collateral code is enough to classify the product: if the COLLCDE is one of 15 possible values, then the Product class is MEREG (with 0.4% generalization error). This rule was successfully applied to 6,439 records, out of 6581, leaving 142 misclassified cases. Without this rule, 6,397 records out of 6581 would not have been correctly classified (showing the unique discriminant power of this rule) and only 19 records would have been correctly classified without this rule. The difference between these two numbers is captured by the column Advantage.

We also identified problems with accurate rule generation based on small training samples, and are investigating potential biases introduced by a non-representative training sample (although we did generate the training sample through randomization). Higher accuracy rules can be obtained either by increasing the size of the training set of by increasing the number of variables. We want to point out that our training set was limited to about 10% (5166 records) of the data, while the remaining 90% of the data (47124 records) was used as test data. Retrospectively, we should have enlarged the size of our training set (and used a better sampling technique) to decrease the possibility of training bias. Of course if the training set is too large, accuracy will decrease due to overtraining.

In addition, we determined that rules with negative "advantages" should be removed from the resulting trees, and this served to increase the utility of the output.

The next section reviews the analyses we performed with the same data, but using a different statistical technique.

2.2 Polytomous Logistic Regression Analysis

In addition to the analysis described above, we applied a statistical method known as polytomous logistic regression to the data [2] (pp. 216ff). Polytomous means multiple-valued-outcome, as opposed to binary outcome where one uses dichotomous regression. There are, in fact, any number of statistical techniques that are suited to the analysis of the CEF data. These include regression, discriminant analysis, and classification and regression trees (CART).

The choice of technique(s) is driven, in part, by the distributional nature of the variable to be explained. Regression and CART are appropriate for continuous response variables. Logistic regression, discriminant analysis, and CART are appropriate for categorical response variables. Beyond this breakdown, techniques have varied strengths and weaknesses in identifying specific types of structure. Often, there is important additional insight gained by contrasting the analyses from two or more methods. We believe that the output of CART would be similar to C4.5 and hence choose to focus on the complementary technique described here, polytomous logistic regression.

As a starting point in any analysis, univariate and bivariate plots and statistics are calculated. Since businesses often summarize data by a single statistic (e.g. mean, median), presenting the entire distribution can be revealing. These insights are also important for later stages of analysis, such as model building. A number of univariate (e.g. histograms) and bivariate (e.g. boxplots of deal size by product) were run. These revealed skewed distributions for deal size and term.

At a more sophisticated level, the polytomous logistic regression was run using the software system Splus [1]. As in C4.5, we choose the dependent variable to be PRODUCT (type of financial deal) which has seven classes. The independent variables were deal size (continuous) and collateral (fourteen classes). A polytomous logistic regression is a parametric model which, for a given set of independent variables, yields the probability distribution of the dependent variable's classes. For example, for collateral 21000000 and a deal size of $300,000, the model estimates the following probabilities for each type of financial instrument (the semantics of these types of instruments are not important):

TTIREG	0.33
MEREG	0.25
SGLINV	0.21
MENQSI	0.20
Other	0.01

This ability to estimate the probability distribution of a variable for a given set of predictor variables is a technology beyond the means of standard reporting systems which present views of the database based on segmenting.

The curves of the model probabilities can be presented graphically for a given collateral code as in Figure 1. Each plot gives the probability (y-axis) of that product being selected for a given deal size (x-axis). One can see that four of the products (ELTOOL, MEREG, TTIREG, OPERLS) have a negligible concentration for this collateral. The relative importance of the other three varies with deal size. At a certain range of deal size (about $20,000), the three are used with approximately equal frequency. One would expect classification to be difficult in this range.

3 Integration of Methods

Through these experiments with common data, we have isolated a number of areas where these two techniques are complementary. In one case, a graphical display of the type illustrated in Figure 1 clearly indicated that one of the products was used almost exclusively for a particular collateral (type of equipment). This allowed us to look back at the rules that generated that particular type of financing instrument to identify the larger *class* of collateral codes for which this instrument holds; in this particular case, there were 15 out of the total 664. In this case, the use of the logistic regression pointed us towards a high-accuracy rule which in turn allowed us to pinpoint a larger set of variables for inclusion in the regression analysis.

Secondly, the results of C4.5 provide information useful for guiding the model specification to be used in the logistic analysis. Determining models of the data improves the productivity of the statistical model building step.

By imposing stronger parametric and distributional conditions on the relationships of the data, regression analysis can generate useful output with small quantities of data, whereas

C4.5 has difficulty in this case. The careful practioner will, of course, wish to validate the modeling assumptions as well as possible with limited data.

Finally, we verified that the default classes generated by the rules in C4.5 did in fact manifest themselves as the highest probability instances in the comparable regression analysis. In this case, the two methods come to the same conclusions but in different ways.

4 Future Directions

With these initial analyses in hand, the next step involves the further specification, working closely with the end users, of the key problems that can be addressed, given their understanding of the capabilities of the software systems. This further specification may entail the identification of additional supporting data that will increase the accuracy of the results, as well as specific kinds of data relationships to be focussed on.

We hope to obtain additional data at a different point in time to perform projections and trend analysis. This will address one of the problems we set out to solve; that of targeting emerging markets. The current analyses are well suited to identifying problem areas (low sales volume, etc) as well as for improving the accuracy of the mapping from a given customer to both what they finance, and how they finance it. However the static nature of the data prevents us from identifying time-dependent relationships.

In addition, we are just starting to use the Fuzzy Isodata program against the same isolated subset of variables to enable more in-depth comparison of the various methods.

5 Conclusions

This paper reported on the application of two knowledge discovery methods (machine learning and statistical) on real data in support of a common business area, quick market intelligence. The initial results have demonstrated that with guidance from the prospective end-users of the technology, and a certain amount of expertise in the correct application of the software programs, suggestive results can be obtained. These interim results are necessary to continue refining the problem definition and the program output. Our hope is that these two will meet and provide some clear competitive advantage in more effectively targeting new customers and focussing the marketing and sales efforts.

In addition, performing similar analysis with the same data using different techniques gives us insight into both how the techniques can be used to complement each other, and how they can be combined to produce better results than either one alone. The next stages of this project will entail comparing the results of another type of knowledge discovery software (Fuzzy Isodata) with the results already obtained, in parallel with generating highly targeted decision rules from C4.5, and specific analyses with the statistical methods to be used directly in the operational environment of GE CEF.

References

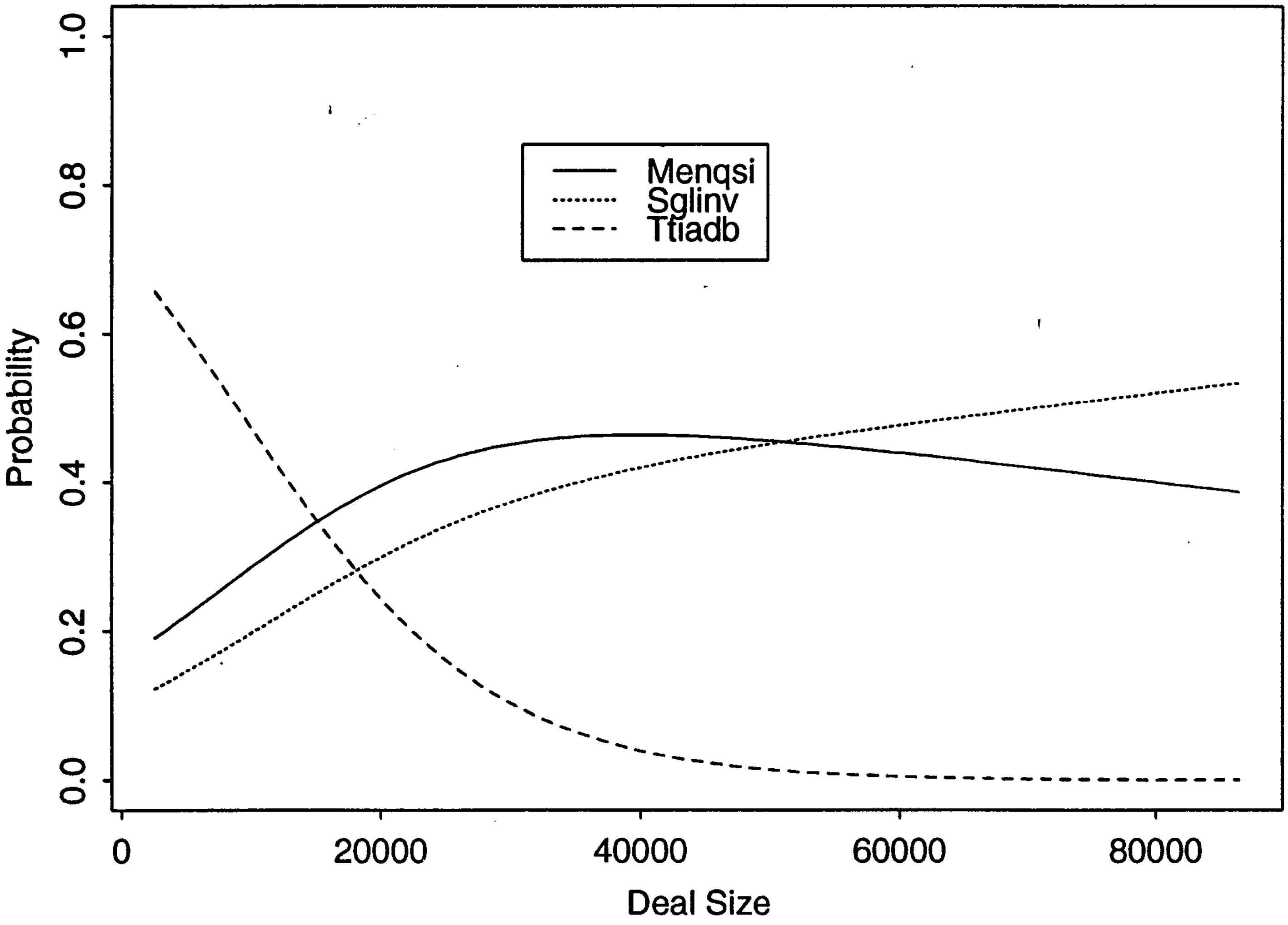
[1] *Splus Reference Manual.* Statistical Sciences, Inc., Seattle Washington, 1991.

[2] D. Hosmer and S. Lemeshow. *Applied Logistic Regression.* John Wiley & Sons, New York, NY, 1989.

[3] G. Piatetsky-Shapiro and W. J. Frawley. *Knowledge Discovery in Databases.* MIT Press, Cambridge, MA, 1991.

[4] J. Ross Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, San Mateo, CA, 1993.

Figure 1. Use of Product by Deal Size for CC 48000000


Menqsi
Sglinv
Ttiadb
Probability
Deal Size