How low can you go? The value of sparse data in retail databases

Received (in revised form): 1st November, 2001

Dr Tony Schellinck

is the F.C. Manning Chair in Economics and Business at the Dalhousie School of Business Administration. He is presently an Informatics Chair Associate and as such conducts research into the strategic use of information within organisations. This includes the application of data mining techniques in the development of customer relationship management systems. He is also president of Focal Research Consultants Ltd., a full service marketing research firm and teaches Marketing Informatics at Dalhousie University.

Kent Groves

is completing an interdisciplinary PhD at Dalhousie University researching Marketing Informatics, with a specific emphasis in the area of pharmacoepidemiology. He is an Informatics Chair Associate, and teaches Marketing Strategy, Direct Marketing and Marketing Management. Kent started Maritime Trading Company, a fully integrated direct to consumer catalogue, retail and Internet marketer of regional products from Atlantic Canada.

Abstract In this paper, the authors examine the effect of sparse data in retail databases on the ability of logistic regression, CHAID and neural networks to produce predictive models. The density of a large retail sales data set was systematically increased through the elimination of low-density cases and the resulting impact on prediction examined. The results suggest that market analysts should consider dropping low density cases when dealing with retail sales data and other data sets composed partly of sparse data.

INTRODUCTION

Many businesses and organisations capture transactional information at the point of sale, or point of contact, through a variety of methods. Financial institutions, health care facilities, on-line retailers, libraries, governments and retailers, to name a few, utilise these systems. The objective is generally the same; to collect purchase and/or transactional information unique to that individual's exchange with the organisation.

For each of these organisations, there is a distinct benefit to collecting this information. Within a traditional retail environment however, there are a number of factors that are making the capture and analysis of this information more critical than ever. Competition,

both virtual and traditional, is expanding the universe of customer choice. Data mining as a strategic tool is being used more than ever. A common application of data mining is the analysis of retail customer segments based on their retail shopping behaviour captured at the point of purchase, either through a loyalty card, or a retail credit card.¹

Retail shopping behaviour is usually recorded at the stock keeping unit (SKU) level, but because there can be thousands of SKUs, sales are typically aggregated to a higher level; often referred to as strategic business units (SBUs). For example, the SBU bread and rolls contains sales from over 20 SKUs including the numerous sizes and brands of breads and rolls available on the grocery shelves in the stores. In the

Tony Schellinck
Dalhousie School of
Business, 6152 Coburg Rd.,
Halifax, Nova Scotia,
Canada B3H 3J5.

Tel: +1 (902) 494 1775;
Fax: +1 (902) 494 1107;
e-mail: TONY.SCHELLINCK
@DAL.CO

authors' experience, SBUs will number somewhere between 50 and 200 for each retail operation. The matrix of data will be formed from rows of customers and columns of SBUs with the cells containing retail sales of that SBU to that customer within a specific time period.

The degree of customer purchase activity will vary significantly in the customer row, with some purchasing products or services from most of the SBUs, others purchasing from some, and still others showing no purchase activity at all during the time period. Thus, each matrix row may vary substantially in its contribution to the total variation of sale within the matrix. The analyst using these matrices will first eliminate those customers that show no activity from the analysis. Frequently this is where the process stops, and cases with activity in only a few SBUs are included in the matrix during the analysis.

The most common use of these data is to use them for predicting behaviours or category membership of customers. Such questions as, will they respond to a coupon offering, are they likely to be purchasers of the in-store bakery and what level of spending should be expected from this customer, are commonly addressed using data mining techniques. In this study, the authors assess the role of sparse data in contributing to these predictions and address the issue of whether they should be eliminated from the matrix in order to improve the power and validity of the analysis.

Another common use of the data is to segment the customers based on retail expenditure. For example, there might be three segments based on low, medium and high expenditure. These segments are profiled and the purchase behaviour within the segment analysed. The issue of a sparse data matrix becomes critical when developing predictive models with the low spending segment where density

could be quite low. There is not a one-to-one relationship between case density and case expenditure. It is possible for a customer to purchase a large ticket item in one SBU category, thus moving that case into a higher spending segment. Most of the low-density cases will, however, be found in the low expenditure segment.

The question is, is there enough information in these sparse data to allow for prediction? All analytical techniques will derive models that tend to use a subset of the SBUs or SKUs for prediction. If only five or fewer of the 133 variables have non-zero values in a low-density case, then chances are that almost all of the heavily weighted variables will have zero for their value and will therefore make no contribution to the prediction of segment membership. Effectively it would be the same as keeping those cases with no activity in the data set.

There are two primary methods of assessing the reliability and value of models. The most common measures used in statistics are the goodness of fit (GOF) measure such as the great variety of X^2 measures available and produced by most statistical packages. For example, the SAS logistic function produces the Hosmer-Lemeshow (HL) test.² There has been some research conducted on the impact of sparse data on GOFs by Farrington³ and Kuss.⁴ The main purpose of these studies was to compare the relative power of alternative GOFs. Kuss in particular concludes that where the data are very sparse there is low power for detecting lack-of-fit. It would seem then that caution must be exercised in interpreting statistics produced by very sparse data.

The second measure used to assess the reliability of models is prediction. In the case of data mining it is the prediction that is most often used as the measure of

a model's reliability. The most commonly produced statistic is the coefficient of variation. A common form of this is the Cox Snell R^2 reported by the SAS software.⁵

For the data miner, the real value of a model is in its ability to predict behaviours by correctly assigning high probability scores to those in the segment most likely to purchase an item or respond to a marketing stimulus. Large databases allow the original sample to be split into three samples: the training set (the sample used to develop the model), the testing set (the sample used to test the model for over fitting) and the validation set (the hold out sample used to determine the predictive power of the model). Using predefined criteria, the analyst then examines the lift achieved by the model in order to determine its predictive capability and thus its value to the data miner/marketer. In this paper the authors focus on the predictive value, specifically the lift value achieved at the top decile (another decile could have been used but the choice is arbitrary and the top decile represents a measure of the model's best performance).

There were several reasons to choose lift as the criterion. First, it is a measure that can be compared across all three analyses used in this study (logistic regression, CHAID and neural networks). Secondly, prediction of segment membership achieved is felt to be the critical measure of a model's success as it is applied in most marketing data mining applications. Thirdly, in particular, lift achieved on the validation set is used as the most practical method of measuring prediction.

METHOD

The researchers had available to them a large database of 50,000 customers, of

which 42,400 had recorded shopping activity. The data were based on the shopping activity of loyalty club members at several stores of a large Canadian grocery retailer over a four-month period.

The data were first aggregated from SKUs into SBUs, and the 43 SBUs showing little purchase activity (less than 1,000 customers purchasing from those SBUs in the four months) were deleted from the matrix. These variables were eliminated as they were expected to contribute little to the prediction of segment membership and analysis and data manipulation was facilitated by the smaller data set. This left 133 SBUs available for the analyses. The density of the cases was defined as the number of SBUs in which a customer had purchase activity. To calculate this value a matrix of ones and zeros, with the ones indicating some purchase activity in the SBU, was created. Summing the ones in each row then created a density variable. This variable's value ranged from one to 119 of the 133 SBUs over the measurement period. (See the density distribution in Figure 1, close to 6,000 cardholders had a density measure ranging from one to six.). The issue was whether the low-density cases contribute to or detract from the ability of the data to profile and predict segment characteristics.

The target segment was defined as those customers who spend a higher than average portion of their grocery dollars, more than 2 per cent, on the SBU snacks and chips and more than 1 per cent on the SBU breads and rolls. The 2 per cent and 1 per cent cut-off points were selected as these created segments of between 20 per cent and 30 per cent of the population (commonly referred to as the density of the dependent or target variable). These levels are considered to be desirable for this type of analysis.

Figure 1 Distribution of independent variable density

Considerable care had to be taken in selecting the dependent variable in order to avoid confounding the effects in the analyses. In particular, the variable had to be independent of the case density measure, such that the density of the dependent variable density was relatively constant no matter how sparse the data matrix. The dependent variable chosen has a density ranging from 27.3 per cent to a maximum of 35.8 per cent over all the matrices entered into the analyses for snacks and chips, and a range of 15.6 per cent to 21.3 per cent for bread and rolls.

Two separate sets of analyses were conducted for each dependent variable. The first set of analyses started with the full matrix of 42,400 customers with densities ranging from one to 119 as independent variables in CHAID, neural networks and logistic regression analyses. Tables 1 and 2 present the lift achieved in the first decile for each analysis.

The density of the matrix as a whole was then increased by eliminating all customers with densities of 30 or less,

leaving a density range of 31 to 119 in the matrix, and the analyses repeated. The process was repeated three more times, eliminating those customers with densities less than 40, 50 and 60 in each subsequent analysis. The range of density used was determined to some extent by the results found. The authors started with the full set of cases, and then selected cases with densities ranging from 31–119. When no appreciable change was found in prediction power by eliminating the bottom density cases, they continued to eliminate cases based on increments of ten in increased density until the sample size declined to a point where the predictive power of the models might be significantly reduced due to the size of the validation sample set. Thus the sample size of those with a density of 61–119 was 7,387.

A second series of analyses were conducted using the sparse matrix comprised of those customers who had a density ranging from one to 20. These analyses were repeated three more times

Table 1: Recorded lifts for increasing independent variable density matrices: Snacks and chips*

Sample size	42,400	21,800	16,624	11,720	7,387
Independent density range	1–119	31–119	41–119	51–119	61–119
Dependent density	30%	33.7%	33.8%	34.3%	35.8%
Lift — logistic regression	1.90	1.95	1.97	1.98	1.95
Lift — CHAID	1.67	1.75	1.69	1.57	1.55
Lift — neural network	1.87	1.93	1.86	1.95	1.73

*Results were obtained using Enterprise Miner 3.0, and were extracted at the first decile

Table 2: Recorded lifts for increasing independent variable density matrices: Bread and rolls*

Sample size	42,400	21,800	16,624	11,720	7,387
Independent density range	1–119	31–119	41–119	51–119	61–119
Dependent density	21.3%	20.2%	18.5%	17.1%	15.6%
Lift — logistic regression	1.91	2.12	2.19	2.37	1.15
Lift — CHAID	1.79	NS**	NS	2.02	NS
Lift — neural network	1.85	2.02	2.24	2.36	1.11

*Results were obtained using Enterprise Miner 3.0, and were extracted at the first decile

** No solution produced by procedure

for very low-density matrices, with densities ranging from 1–15, 1–10, and 1–5. The goal was to see how low the matrix density could be before the analyses lost all ability to predict segment membership. The authors focused on the density range 1–20, as the analysis of the higher density data had shown no change in predictive power of the data when the cases with density 1–30 were dropped from the analysis. This suggested that if data scarcity was going to affect prediction it would probably be somewhere below these levels of density.

RESULTS

The original hypothesis had been that the denser the matrix, the greater the ability of the data to predict the dependent value, as measured by a higher lift value. This is based on the premise that with higher density cases there are more likely to be non-zero values for the variables found to be predictive of segment membership. The authors were, however, very surprised to

find that, for snacks and chips, the lift values remained virtually unchanged as the low-density cases were eliminated (see Table 1). For example, the lift generated by the logistic regression ranged from 1.90 to 1.98, and lift generated by the neural network ranged from 1.73 to 1.95. The CHAID generally performed slightly worse than the other two techniques, ranging in lift from 1.50 to 1.75.

The results of the analysis on the bread and rolls buyers (Table 2) were closer to the hypothesised pattern, with the lift increasing from 1.91 to 2.37 for the logistic regression models and from 1.85 to 2.36 for the neural networks as the data increased in average density. That is, until the analysis was run on the highest density sample where the lift for both the logistic regression and neural network models declined to 1.15 and 1.11 respectively and the CHAID analysis could not produce a model. This suggests there may be problems modelling at the high end of the spectrum with respect to independent

Table 3: Recorded lifts for decreasing independent variable density matrices using sparse density matrices: Snacks and chips*

Sample size	11,978	9,613	5,179	1,919
Independent density range	1–20	1–15	1–10	1–5
Dependent density	30.1%	27.3%	30.0%	30.0%
Lift — logistic regression	1.78	1.56	1.26	1.44
Lift — CHAID	1.90	2.05	1.77	NS**
Lift — neural network	1.73	0.93	1.08	0.64

*Results were obtained using Enterprise Miner 3.0, and were extracted at the first decile

**No solution produced by procedure

Table 4: Recorded lifts for decreasing independent variable density matrices using sparse density matrices: Bread and rolls*

Sample size	11,978	9,613	5,179	1,919
Independent density range	1–20	1–15	1–10	1–5
Dependent density	20.7%	20.0%	20.0%	20.0%
Lift — logistic regression	1.41	1.71	1.54	1.32
Lift — CHAID	NS**	NS	1.20	NS
Lift — neural network	1.06	1.06	0.76	NS

*Results were obtained using Enterprise Miner 3.0, and were extracted at the first decile

**No solution produced by procedure

variable density (something to be examined in future research).

Given these results, the authors decided to see how sparse the dependent variable matrix could be before the analyses would lose their ability to predict. The four sets of analyses, with dependent variable density ranges of 1–20, 1–15, 1–10 and 1–5 are presented in Tables 3 and 4.

The analyses performed on the snacks and chips segment continued to perform well using the 1–20 density matrix, but started to decline as the matrices became sparser. The lifts for the logistic regression declined from 1.78 to 1.26 and the neural network lifts declined from 1.73 to 0.64. Interestingly, when compared to the lift achieved in the high-density matrices, the CHAID analyses performed even better than the first three low-density matrices, giving lifts of 1.90, 2.05 and 1.77. It was not until the extremely low-density matrix was used that the CHAID lost its ability to predict the target.

The low-density models developed in the bread and rolls analysis proved to be relatively poor, with both CHAID and neural networks unable to derive models and the lifts for the logistic regression models ranging from 1.32 to 1.71 (Table 4). This suggests that the low-density matrices are relatively poor predictors of segment membership. This might also explain why the exclusion of the low-density cases improved the model performance in Table 2. In contrast, the low-density models predicting chips and snacks segment membership based on the 1–20 density matrix still performed well. Thus, eliminating these cases would do little to improve the prediction of snacks and chips segment membership.

CONCLUSIONS

The purpose of this paper was to explore the possible influence of the sparse section of data matrices on the predictive ability of models commonly used in data mining. If there is found to be a

problem because of sparse data then the analyst should consider eliminating those cases prior to analysis in order to improve the performance of the models. There are also implications for analysis of low expenditure segments whose data may be composed primarily of cases with sparse data.

The results indicate that in some instances the low-density data in the matrix do not contribute to the predictive power of the models and that they may in fact reduce the predictive power of the total data. For snacks and chips, dropping low-density cases did not improve the predictive power of the models, and the matrix containing case with densities of 1–20 was still able to predict segment membership reasonably well. Matrices containing cases with densities 15 and lower were not always able to produce predictive models, however, suggesting that these cases could be dropped. For breads and rolls, dropping the low-density cases did improve the predictive power of the data, and the sparse matrices were usually not able to produce predictive models. The results using this target variable clearly suggest that dropping the low-density cases helped in the analysis.

How many to drop is not clear from the results of the analysis. In the chips and snacks example the model stopped predicting when the independent variable density was 15 or lower; for breads and rolls the models did not perform well with densities of 20 or less. The exploratory nature of this analysis does not provide enough information to make recommendations for standard cut-off points when eliminating cases with sparse data. Instead, analysts could start analysis on the low-density section of the matrix and determine at what level the data can provide reasonable predictions and drop

those cases with densities below this level.

If the analyst is conducting segmentation analysis, the low spending segment could be comprised largely of low-density cases. If the data are partitioned into three expenditure segments the lowest expenditure segment might be approximately 33 per cent of the sample. In this dataset, those cases with independent variable densities of 20 or less made up 28 per cent of the sample. The value of these cases thus becomes a real concern for the analyst.

The field of data mining is new, and processes for conducting analyses are still evolving. One aspect rarely considered by analysts (based on reactions by analysts at conferences) is the possible impact of cases with sparse data. This study has highlighted the need to consider the impact of these cases on the ability of models to predict. Further research utilising Monte Carlo simulations will be helpful in developing standard criteria for the analyst to consider when faced with such data sets.

References

- 1 Drew, J. H., Mani, D. R., Betz, A. L. and Datta, P. (2001) 'Targeting customers with statistical and data-mining techniques,' *Journal of Service Research*, Vol. 3, Iss. 3, Thousand Oaks, February, pp. 205–219.
- 2 Hosmer, D. W. and Lemeshow, S. (1989) 'Applied logistic regression', John Wiley and Sons, New York.
- 3 Farrington, C. P. (1996) 'Assessing goodness of fit of generalized linear models to sparse data,' *Journal of the Royal Statistical Society, Series B (Methodology)*, Vol. 58, Iss. 2, pp. 349–360.
- 4 Kuss, O. A. (2001) 'SAS/IML macro for goodness-of-fit testing in logistic regression models with sparse data', Proceedings of the Twenty-Sixth Annual SAS (r) Users Group International Conference, SAS Institute Inc., Cary, NC, Paper 2–65–26.
- 5 Cox, D. R. and Snell, E. J. (1989) 'Analysis of binary data', Second Edition, Chapman & Hall, London.