Papers

Marketing data analysis and data quality management

Received (in revised form): 10th March, 2003

Richard J. Courtheoux

is President and founder of Marketing Analysis Applications, a management consulting firm. He has worked in direct marketing since 1978 and was formerly Senior Vice President of Database and Consulting Services for Experian, President of Precision Marketing and a consultant at Kestnbaum and Company. He has an MBA from the University of Chicago, an MS in Computer Science from the Weizmann Institute of Science and a BS from Yale.

Abstract Successful marketing data analysis depends critically on the quality of the data that are available for analytical work. This paper draws on general quality management concepts as well as recent developments in the field of data quality management to address issues related to marketing data analysis. Marketing databases have a number of particular characteristics that can be described and should be addressed in a quality management programme. Particular recommendations are made for detecting data quality problems that pertain to marketing databases that should be part of a continuous quality management programme. The common application of predictive modelling to drive target marketing programmes should include attention to data quality management issues at both the model building and customer scoring phases. A modelling professional can make significant contributions to ongoing data quality management and should be alert to data quality issues since they are a significant success factor.

INTRODUCTION

Successful customer modelling is about 80 per cent dependent on having the right data for analysis, but it seems that more than 80 per cent of the written material on customer modelling deals with analytical techniques. Virtually all of the other 20 per cent is devoted to sampling or variable construction issues. Somehow, there has been very little written for data analysts and database marketers addressing the basic subject of data quality management and its relevance. Perhaps it is thought that the need for high data quality is so self-evident that it is not worthy of comment. However, even moderately

self-evident points are often stated to remind people of their importance. Perhaps for some authors and experts the techniques are simply more interesting than data quality issues.

Whatever the explanation, without systematic attention to data quality management, a gap exists in the data analysis discussion. This paper attempts to address that gap and is built on these premises:

- data quality issues are a common and important source of suboptimal analysis and modelling effectiveness. Therefore, a data analysis professional working in the database marketing

Richard J. Courtheoux
Marketing Analysis
Applications, 566 Woodlawn
Avenue, Glencoe, Illinois
60022, USA.

Tel: +1 847 242 0318;
Fax: +1 847 631 0865;
e-mail: rcourtheoux@
maa-online.com

world needs to invest in understanding data quality management issues in order to deliver the best possible results;

- data quality has been systematically addressed by computer science and information technology researchers. Marketing database designers, operators and users now have available a much more developed set of intellectual tools than existed previously for managing data quality;
- marketing databases have particular characteristics that shape the data quality issues and relevant procedures.

Exploring these ideas yields practical recommendations for improving marketing database procedures and in turn leads to better analytical data sets.

WHAT IS DATA QUALITY?

The word 'quality' came to have a specific technical meaning for management as a result of the now famous work by W. Edwards Deming.¹ Rather than viewing quality as a general synonym for 'goodness' or 'luxury', Deming and his disciples insisted that quality could be defined more precisely and quantitatively measured. Their operative definition of quality became 'conformance to requirements'.²

That simple definition has a profound implication: someone must rigorously define requirements. Jumping to the database marketing realm, this presents a serious challenge for most companies. How often do companies have a set of requirements against which their databases can be checked for conformance? Among companies that have any requirements at all, how many have requirements that are comprehensive and complete? More common is the situation where someone reviews reports to see if they 'look OK'. In addition, the person reviewing those

reports is often not someone who is very experienced, highly numerate and motivated by a strong belief that this type of work is critical for career advancement.

The most notable applications of Deming's work were in the management of production processes for manufactured material goods. Database management processes have many similarities to traditional manufacturing, but the intangible nature of information makes the database situation somewhat different:

- a manufactured product will often either work or not work; it is either working or broken. A database will commonly be functioning, but storing some data incorrectly or supplying partially incorrect information extracts;
- a manufacturing assembly line typically pushes out and is done with a completed product. Database content may remain active and part of a database for an indefinite period of time.

The greater ambiguity of the database situation argues for a standard more nuanced than requirements. A useful operative definition of data quality provided by Loshin is 'fitness for use'.³

This definition carries with it an actionable implication: someone must anticipate and understand how the data are going to be used. With this perspective, data quality becomes a messy economic matter, not a matter of beliefs and assumptions. Some related implications are:

- not all data will necessarily need the same degree of accuracy to be of high quality;
- different types of data problems will have significantly different impacts on data quality;
- users of data may need to be advised

as to the suitability of various data elements for particular applications.

Although ideally one would like to have perfectly accurate and complete analytical data sets, some statistical models used to drive decision making may be very effective and financially valuable even though they use imperfect data.

The point of introducing these respective definitions of 'quality' and 'data quality' is that they suggest productive actions for those who work with marketing databases. Several general points from the quality management literature are very directly applicable to managing marketing databases.

- *Quality is free.* Sweeping away the definition of quality as 'goodness' or 'luxury' (for production purposes) leads to a breakthrough realisation. The costs of wasted expense and effort from not insisting on quality far exceed the costs of comprehensively pursuing quality. In the database marketing world the costs associated with an erroneous selection of names for a targeted marketing campaign will tend to exceed the costs of a quality management programme.
- *Quality problems should be caught and fixed as soon as possible.* At each step in a production process additional expense and effort is invested, so the damage from a defect keeps growing as it moves along. Similarly, the economic costs of errors in a marketing database will tend to escalate the longer they go undetected and unfixed. From the point in time when inaccurate data enter the marketing database, all subsequent analyses, reports, predictive models and selections will be corrupted. Correcting an error will tend to get more costly and less feasible the longer it has been present in the database.

- *Quality is a shared responsibility.*

Production systems sometimes rely on an inspection step at the end of the process to catch all quality problems. Not only is this more expensive, as mentioned above, but inevitable imperfections of the inspection step make the approach relatively ineffective. Technology professionals may have system responsibility for implementing various forms of quality reporting and checking. Marketers, however, may be the only likely source of information on what values or frequencies of occurrence are to be expected; marketers may also have some ability to discriminate between high value and low value pieces of data. Analysts or modelling professionals may be the best source of expertise on statistical tests to be applied for detecting quality problems.

The subject of data quality has received increased attention throughout the information technology world as data warehouses become more common and enterprise knowledge management is more systematically addressed. Some of the topics dealt with by this specialty area are data quality requirements, business rules, measurement, economics and statistical process control. The insights developed on these topics are not only applicable at the warehouse and enterprise level, but also at the level of data marts and marketing databases. There are a number of excellent sources that review data quality issues and present strategies for addressing them.⁴⁻⁷

MARKETING DATABASE CHARACTERISTICS

While it would be of minimal incremental value simply to review the data quality literature here, there is benefit in describing more specifically

what is particularly characteristic of marketing databases and their data quality issues. The following are some of those key characteristics.

Marketing database temporality

Marketing databases are temporal databases. That is, in order to fulfil its fundamental mission of recording all customer data, a marketing database must contain extensive historical data associated with either a point or an interval in time.⁸ To support basic customer modelling, the marketing database must be able to provide point-in-time views of available customer information and subsequent time interval behaviour. Systems, operational processes and underlying business characteristics evolve, making it complex to summarise or compare information collected over time. Data quality evaluations depend to some degree on the age (recency) of the data under consideration.

The temporal nature of marketing databases is probably the single biggest issue that generalist database professionals fail to understand adequately when they come to working with marketing databases. Most databases are primarily snapshot databases that contain current data and are designed to update fields as new data become available. Temporal database issues are not typically part of database training, except at very sophisticated levels, and the available commercial tools are not yet very powerful. With temporal databases, common information technology issues such as assignment of identifiers as keys or data archiving procedures can have subtle, but nasty, implications with respect to data usability.

Data collection

Marketing databases collect data from other systems. They generally import

data from operations (eg order taking, websites) and do relatively little direct generating or recording of new data. For each feeder system the marketing database needs to have some mechanism for checking data before they are loaded; wherever possible checking should be both of individual records and statistical summaries. As with production process quality management, it is much cheaper to deal with quality issues at an early stage of the process. For some quality problems, such as an erroneous extraction of data, there may be an option to fix the data before loading. For much of the data being imported, however, there is no option to fix suspect data, since the primary operational transaction has been completed. Rejecting all suspect data is not an attractive option since that would diminish the completeness of the database information. The long-term data quality management will largely depend on the marketing database system's ability to identify quality problems and push for fixes in the feeder systems.

Extraction of names

Extraction of names for targeted marketing is the major operational use. For this application, marketing database quality issues can be assessed in terms of whether a financially desirable communication does not get to an available name or, alternately, a communication is wasted on a financially undesirable name. The first situation can occur if a name has been mistakenly purged from the database or contact information is erroneous or incomplete. In addition, poor quality customer descriptive data can lead to both ineffective predictive models as well as mistaken evaluations of names. Where, however, selection decisions and communications processes are fairly robust, a certain level of data inaccuracy

or incompleteness may be financially unimportant. Highly personalised communications, on the other hand, will be highly sensitive to inaccuracies or incompleteness in certain data fields.

Statistical summaries

Statistical summaries are used for decision support. Marketing databases are used to build statistical descriptions of customer behaviour or characteristics. In general, this application has been driving quality requirements higher because more decisions are data driven and because time frames for decision making are growing shorter. Inaccuracy in the data only becomes important if it is so prevalent and so systematic that it changes the conclusions being drawn from the statistical summaries. It is important to emphasise that there will be a major difference in data fitness for use depending on whether they are systematically biased or randomly inaccurate. Lack of data coverage (completeness) becomes an issue when business opportunities are missed or when poor intuitive decisions are not stopped by convincing facts.

Name and address processing

Customer name and address processing is a core database building block. Marketing databases depend on name and address processing to link together information from multiple sources and time periods. A whole subdiscipline has developed over several decades that deals with standardising formats, fixing elements where feasible, evaluating name and address pairs as matches and grouping together records from an entity (eg a household or a business). The matching process is often characterised by relative rates of 'overkill' (two records falsely regarded as a match) and

'underkill' (two records falsely regarded as being different). There are no universal standards for appropriate levels of overkill and underkill as the intended application determines what is economically optimal. Further, the linkage of records to a single household or business may have major implications for all intended uses of the data.

Several customer data integration (CDI) systems have been developed that significantly improve the effectiveness of consumer name and address processing. These systems use the subdiscipline's processing capabilities, but add to its effectiveness by building vast reference databases linking varying versions of names and addresses. Deploying this technology requires that a marketing database be deeply tied to one vendor's system and its reference database.⁹⁻¹¹

Speed of marketing operations

Marketing operations and decision making are getting faster. Because the operations (eg a selection of names for a marketing campaign) may be specified and executed in a very short time, there is very little time buffer for checking data, fixing detected problems and then redoing the processing. Similarly, data-driven decisions will often be made on the basis of executing a single series of queries; the decision time-frames often will not allow for rerunning the queries. To be fit for use the marketing database and common types of queries run against it need to have a very high likelihood of being correct without fixing and reruns.

Attention to these characteristics can help database marketers and modelling professionals to develop data quality requirements and evaluate data fitness for use. In addition, they can provide a foundation for data quality measurements that allow for current status assessment and calibration of improvements. The

more active users there are for marketing database models and statistical summaries, the more sponsors and supporters there will be for data quality management-related expenditures.

MARKETING DATA QUALITY MANAGEMENT ISSUES AND TECHNIQUES

Marketing database quality management should include a number of practices that reflect both the particular character of the database and its intended set of applications.

Consistency

Consistency between points in time or versions should be checked. Starting with the idea that quality is 'conformance to requirements' raises the issue of developing applicable requirements. For many types of data an effective and very automated way to develop requirements is to bootstrap them over time. When a marketing database is first created the user community may be asked to inspect reports and sign off on the reasonableness of the distribution of values. Once this has been done, at subsequent updates or points in time the system should compare prior and current value distributions. In effect, the prior distribution becomes the requirement and the new distribution's conformance to it can be measured and tested for statistical significance; a chi-square test of significance can be used. For most data variables no change in value distribution should be detected, so this procedure functions mainly as a way of demonstrating the probable acceptable quality of the new values. The process will detect changes in value distributions which should receive expert attention; some changes may quite reasonably reflect changes in the business situation. An

example comparison of this type is illustrated in Table 1; in this example, a manager familiar with recent results would look at the percent change column to see if it reflects known performance. A very effective implementation of this practice is incorporated in Experian's Audit OnlineTM technology.¹²

Automation

Quality checking should be automated. Modern marketing databases will often have hundreds or thousands of different fields pertaining to each customer; the database may contain records on many millions of customers. To be effective, data quality management procedures need to be highly automated so they are quick and cheap. Both rule-based and statistical checks should be triggered each time there is a marketing database update or, alternately, at periodic intervals. The processing for such checks is usually not a particularly large burden, but manual review of the results may be far too time-consuming to be economical and feasible. Tables of the form of Table 1 should be produced automatically.

Besides automating the production of quality checking, the system should automate the identification and reporting of potential problems. A report or query interface should highlight identified problems so that people working with the database can use their time most efficiently to deal with the quality issues. A three column report of the type illustrated in Table 2 can give a person a quick overview of what potential quality issues exist.

Input, output and intermediate steps

Quality checks should be applied on input, output and key intermediate steps. The type of approach described in the preceding two sections should be applied

Table 1: Value distribution consistency

Customer type	30/11/02 count	31/12/02 count	Count change	% change
Buyer — Catalogue only	628,536	638,593	10,057	1.60
Buyer — Internet only	216,657	223,677	7,020	3.24
Buyer — Retail only	1,043,323	1,052,504	9,181	0.88
Buyer — Multichannel	270,318	275,319	5,001	1.85
Gift recipient	69,055	69,842	787	1.14
Requestor	239,254	239,254	2,321	0.97

Table 2: Database quality issues summary

Quality issue type	Variable	Value
Distribution inconsistent	Customer type	Buyer — Internet only
Distribution inconsistent	Source code	Multiple
Distribution inconsistent	Gender	Unknown
Undocumented code	Source code	B02%*>#
Undocumented code	Source code	B0004%*

not only on output (eg after the update is complete), but also to input sources as data are being collected. Major intermediate steps should be identified and automated data quality checking should be applied there as well.

Domains for each variable

Domains for each variable should be identified. Many fields, especially coding schemes, have clear sets of defined or permitted values which drive edit checks when data are entered into operational systems. Marketing database quality management systems should be constantly looking for undefined values or values that have not appeared previously; they should similarly be highlighting values that have previously appeared with some frequency and have abruptly disappeared. Sometimes it is advisable to build tables of common corrections that should be applied when certain erroneous codes are encountered (eg mistyped state or province codes). Proactive management of domain definitions can be very helpful in quickly identifying when a data source is having a quality problem.

Transitions in customer status

The temporal nature of marketing databases makes two other types of data quality management checks particularly important. First, there are only certain transitions in customer status that should occur if the data are being maintained correctly. For example, the total number of purchases or total dollars spent should only go up; the date of the most recent transaction should only get later. A well-developed set of rule-based checks and associated reports provides a hugely important defence against undetected mishandling of customer data. Customer transition errors are particularly serious because they almost always result from a logical error in the processing system.

An example data quality management report that looks at customer transitions is shown in Table 3. Customers are classified into Recency/Frequency/Monetary cells at two points in time. During the interval between the two points in time customers in any cell can either stay in that cell or migrate to another cell; customers from other segments can migrate into a cell as well. The business may also be acquiring new customers and dropping existing

Table 3: Customer update audit report

DB customer segments			Database 30/11/2002	Update customer changes 30/11-31/12/02					Database 3/12/2002	Database changes	
Recency	Freq.	Avg. order \$	Count	Stayed	Moved in	New	Moved out	Dropped	Count	Change	% Change
0-3 Mo.	6+	>500	36,399	32,468	2,268	1,843	3,640	291	36,579	180	0.49
0-3 Mo.	6+	251-500	100,128	89,314	6,184	4,245	10,013	801	99,743	(385)	-0.38
0-3 Mo.	6+	101-250	133,668	119,232	8,281	6,044	13,367	1,069	133,557	(111)	-0.08
0-3 Mo.	6+	1-100	199,124	177,619	12,297	8,429	19,912	1,593	198,345	(779)	-0.38
0-3 Mo.	6+	Total	469,319	418,633	29,030	20,561	46,392	3,754	468,244	(1,095)	-0.23
0-3 Mo.	2-5	>500	14,293	12,750	888	680	1,429	114	14,318	25	0.17
0-3 Mo.	2-5	251-500	33,487	29,870	2,067	1,408	3,349	268	33,345	(142)	-0.42
0-3 Mo.	2-5	101-250	49,162	43,853	3,056	2,380	4,916	393	49,289	127	0.26
0-3 Mo.	2-5	1-100	65,070	58,042	4,031	2,944	6,507	521	65,017	(53)	-0.08
0-3 Mo.	2-5	Total	162,012	144,515	10,042	7,412	16,201	1,296	161,969	(43)	-0.03
0-3 Mo.	1	>500	25,233	22,508	-	2,793	2,523	202	25,301	68	0.27
0-3 Mo.	1	251-500	66,040	58,908	-	6,970	6,604	528	65,878	(162)	-0.25
0-3 Mo.	1	101-250	111,628	99,572	-	11,863	11,163	893	111,435	(193)	-0.17
0-3 Mo.	1	1-100	131,996	117,740	-	14,214	13,200	1,056	131,954	(42)	-0.03
0-3 Mo.	1	Total	334,897	298,728	-	35,840	33,490	2,679	334,568	(329)	-0.10
0-3 Mo.	Total		966,228	861,876	39,072	63,813	96,623	7,729	964,761	(1,467)	-0.15
4-6 Mo.	6+	>500	23,582	21,035	2,652	-	2,358	189	23,687	105	0.45
4-6 Mo.	6+	251-500	53,525	47,744	5,514	-	5,353	428	53,258	(267)	-0.50
4-6 Mo.	6+	101-250	105,367	93,987	11,668	-	10,537	843	105,655	288	0.27
4-6 Mo.	6+	1-100	129,073	115,133	13,963	-	12,907	1,033	129,096	23	0.02
4-6 Mo.	6+	Total	311,547	277,899	33,797	-	31,155	2,493	311,696	149	0.05
4-6 Mo.	2-5	>500	11,141	9,938	1,148	-	1,114	89	11,086	(55)	-0.49
4-6 Mo.	2-5	251-500	25,932	23,132	2,759	-	2,593	207	25,891	(41)	-0.16
4-6 Mo.	2-5	101-250	43,960	39,212	4,689	-	4,396	352	43,901	(59)	-0.13
4-6 Mo.	2-5	1-100	52,323	46,672	5,864	-	5,232	419	52,536	213	0.41
4-6 Mo.	2-5	Total	133,356	118,954	14,460	-	13,335	1,067	133,414	58	0.04
4-6 Mo.	1	>500	19,809	17,670	2,072	-	1,981	158	19,742	(67)	-0.34
4-6 Mo.	1	251-500	51,794	46,201	5,576	-	5,179	414	51,777	(17)	-0.03
4-6 Mo.	1	101-250	86,822	77,445	9,628	-	8,682	695	87,073	251	0.29
4-6 Mo.	1	1-100	119,168	106,298	13,076	-	11,917	953	119,374	206	0.17
4-6 Mo.	1	Total	277,593	247,614	30,352	-	27,759	2,220	277,966	373	0.13
4-6 Mo.	Total		722,496	644,467	78,609	-	72,249	5,780	723,076	580	0.08

customers during the time interval. The overall audit process must ensure that the counts by cell at the two time periods can be balanced. However, certain transitions are not valid and should have a count of zero: previous buyers cannot buy again (becoming 0–3 month recency) and move into a frequency equal to one cell. Similarly, a customer cannot be new to file and have a 4–6 month recency.

A second set of rule-based checks should relate to reported relative levels of performance among different types of customers. For example, customers who had made a purchase in the first half of a year should (usually) be much better than average customers in terms of their productivity during the second half of the year. Developing these types of rule-based checks as a method for detecting data quality issues is tricky since they may be detecting fundamental business issues, rather than data problems. If these checks identify a likely problem, it may be necessary to reference other business reports to decide whether the problem is with the business or the data. Table 4 illustrates this type of report showing expected patterns of performance by Recency/Frequency/Monetary cells.

Rule discovery

It is sometimes recommended when designing databases that data mining or knowledge discovery in databases (KDD) technology be deployed to discover dependencies that exist in the data. These dependencies are then built into the data definitions and may lead to the selection of keys for tables. Besides the application of this technology when the marketing database is being defined, there is good reason to deploy it to monitor the quality of a database that is in production use. While data analysts and modellers are dependent on the

quality of the data they have for a particular task, they also have an important role to play in looking for suspicious data relationships. When data mining or KDD technology is available, it should be used to help update the set of rules or relationships that exist in the marketing database and should be checked with each update; discovered relationships that should not exist can be investigated and corrected.

A successful implementation of these ideas requires a detailed analysis of the processing steps and data sources. An excellent review of practical issues to address and the questions to ask is given in an article by Gendele¹³.

DATA QUALITY ISSUES AFFECTING CUSTOMER MODELLING

Customer modelling professionals are generally individuals who have received very good training in statistical concepts and analytical techniques. As part of their statistical training virtually all analysts have studied Exploratory Data Analysis (EDA) approaches and techniques. EDA deals with what a data analyst should do when given a new set of data. It is intended to catch errors in an analytical data set as well as identify data relationships that may be of interest. The issues dealt with by EDA and the techniques it deploys should continue to be part of the data investigation for analytical projects.

By the point in the process where a modelling data set has been created, however, many data quality issues may no longer be addressable or detectable. EDA may, for example, successfully detect problems in the way an analytical sample has been extracted and thereby prevent waste of effort on a deeply flawed model build. It would also successfully identify fields that were

Table 4: Customer performance report

DB customer segments				Period 1/7-31/12/02 per customer activity					
Recency	Freq.	Avg. order \$	Count	Mktg \$	Orders	Buyers %	Net sales \$	Profit \$	
0-3 Mo.	6+	>500	34,821	10.87	0.566	47.0	106.57	28.08	
0-3 Mo.	6+	251-500	95,336	10.50	0.534	44.0	87.09	20.87	
0-3 Mo.	6+	101-250	128,408	10.42	0.518	43.0	70.56	14.44	
0-3 Mo.	6+	1-100	195,828	10.33	0.517	43.0	58.15	9.57	
0-3 Mo.	6+	Total	454,393	10.44	0.525	44.0	71.44	14.72	
0-3 Mo.	2-5	>500	13,735	10.27	0.508	42.0	87.07	21.26	
0-3 Mo.	2-5	251-500	31,868	9.53	0.421	35.0	66.93	14.51	
0-3 Mo.	2-5	101-250	47,945	9.05	0.413	35.0	53.54	9.68	
0-3 Mo.	2-5	1-100	61,998	8.67	0.379	32.0	42.59	5.90	
0-3 Mo.	2-5	Total	155,546	9.09	0.409	34.0	54.88	10.20	
0-3 Mo.	1	>500	24,353	9.67	0.381	32.0	70.39	16.01	
0-3 Mo.	1	251-500	65,231	9.30	0.359	30.0	56.05	10.79	
0-3 Mo.	1	101-250	107,944	8.96	0.329	28.0	44.12	6.55	
0-3 Mo.	1	1-100	128,247	8.41	0.345	29.0	35.92	3.72	
0-3 Mo.	1	Total	325,775	8.89	0.345	29.0	45.24	6.96	
0-3 Mo.	Total		935,714	9.66	0.443	37.0	59.57	11.29	
4-6 Mo.	6+	>500	22,891	9.35	0.445	37.0	83.48	21.15	
4-6 Mo.	6+	251-500	52,316	8.68	0.425	36.0	66.37	15.11	
4-6 Mo.	6+	101-250	103,319	8.56	0.412	35.0	54.87	10.71	
4-6 Mo.	6+	1-100	126,488	8.42	0.408	34.0	45.43	7.10	
4-6 Mo.	6+	Total	305,014	8.60	0.415	35.0	55.07	10.73	
4-6 Mo.	2-5	>500	10,907	9.04	0.408	34.0	69.49	16.10	
4-6 Mo.	2-5	251-500	25,493	8.39	0.345	29.0	52.50	10.37	
4-6 Mo.	2-5	101-250	43,825	7.98	0.347	29.0	43.67	7.23	
4-6 Mo.	2-5	1-100	51,399	7.93	0.321	27.0	35.20	4.06	
4-6 Mo.	2-5	Total	131,624	8.12	0.342	29.0	44.21	7.34	
4-6 Mo.	1	>500	19,786	8.42	0.315	27.0	58.61	12.98	
4-6 Mo.	1	251-500	51,519	8.12	0.317	27.0	48.61	9.26	
4-6 Mo.	1	101-250	84,913	7.82	0.307	26.0	39.37	5.93	
4-6 Mo.	1	1-100	117,931	7.48	0.315	27.0	32.44	3.45	
4-6 Mo.	1	Total	274,149	7.79	0.313	27.0	39.51	5.98	
4-6 Mo.	Total		710,787	8.18	0.362	31.0	47.06	8.29	

unpopulated or had numerous undocumented values, but the modeller would have little chance of fixing the problem if it resulted from longstanding data quality problems in the marketing database itself. Data quality problems such as incorrectly recorded original source codes or erroneously attributed historical purchases would probably not get detected at all by normal EDA. The value of customer modelling to a database marketing business depends on the quality of the marketing database itself; techniques invoked in a modelling project are not a substitute for ongoing data quality management. It is also true that a company with severe data quality problems in its marketing database should address those issues before starting a modelling initiative.

Many of the issues related to the marketing database's fitness for use in modelling relate to its design and ability to produce derived variables. The issue is that while the content of the database may be correct, it may be incomplete or inadequately detailed to support the modelling application. This can result from design inadequacies with respect to either data definitions or processing capabilities.

Time and data quality

A set of data quality issues related to time impacts on customer modelling effectiveness. The starting point for the modelling effort is a snapshot of the customer data as they existed when names were selected for the marketing effort to be modelled. Ideally, a copy of the entire customer database would be created with all the data permanently frozen as of the specified point in time. Although it may be economically and technologically feasible to create the copy (in fact a usable database backup copy will often exist), few companies can

afford to restore and process a full copy of an old version of the database. To adapt to that reality companies typically choose among compromise strategies.

- *Some historical detailed data may be summarised.* This process may be very prone to programming errors, especially if it is done in a customised or *ad hoc* manner.
- *The current marketing database may be used with transactional data recorded after the select date filtered out.* This will work for certain information, such as buying history, but will not work for many important types of data. This approach will not work well for demographic data (were they available on the select date?), title codes (did they change after the select?) or available credit.
- *Only a random sample of customers' data is saved.* The main problem with this strategy is that many database marketing campaigns have low response rates and the reliability of a statistical model is driven largely by the number of respondents in the sample. Of course, it is impossible to know ahead of time who will respond so sampling after selects will just reduce the number of respondents available in the analytical data set. One could store the entire database and sample after the campaign is complete, but this would involve reloading the entire database prior to sampling.
- *Only data for customers contacted in the marketing campaign are saved.* This is unacceptable if behaviour of contacted customers must be compared to behaviour of non-contacted customers. It also forecloses the opportunity to project how much economic benefit might be derived from contacting additional customers using the newly constructed model.

These data quality management issues affect the development of independent (predictor) variables and sample composition. Attention to time also affects dependent variable quality in at least the following ways:

- it provides one form of filter on customer behavioural data that should be considered when building a dependent variable;
- it may provide important perspective on whether marketing activities add to the total amount of business or just shift it in time.

Quality management of time-related information will tend to be easiest and most accurate when the information, such as buying behaviour, is derived from time stamped transactional data.

Matching records

Marketing databases use name and address matching technology in order to pull together all records that should be associated with a customer. However, the matching standards (especially overkill and underkill calibration) that are specified for marketing database operations are typically not the most appropriate standards to apply to analytical data set creation. For database operations it is usually desirable to risk some underkill in order to have very minimal overkill. The reasoning is that once records are associated with each other it is very hard to ever disassociate them; in some situations (especially with credit-related data) there is a large asymmetry in the economic consequences of falsely connecting together two different customers' data versus missing a connection that should be made. When preparing an analytical data set, by contrast, minimising overkill can easily create a bias towards certain

customers (ie those with easily matched names and addresses). Modelling projects built on biased datasets produce biased, suboptimal results that reflect the data quality management decisions in addition to profit generating customer behaviour.

When tight matching logic produces very high match rates, the modeller can be relatively confident that matching issues will not degrade results. So, if the database is being maintained with tight matching logic and the standard association of orders to customers is producing high match rates to the pool of marketed names, then the analytical data set is likely to be satisfactory. If this is not the case, modellers may want to specify multiple versions of the analytical data set with varying match logic applied. Alternate models can then be built from each analytical data set and compared for plausibility and efficacy.

Classification schemes

Marketing databases should retain detailed, permanent, original identifier data such as product SKU (stock keeping unit) numbers or source codes. To make use of these detailed data for analytical or modelling purposes usually requires a classification system which groups items together. Establishing and maintaining data quality with respect to a classification scheme one confronts several quality issues.

- *Sustaining consistency over time and across business units.* Operational business exigencies often lead to changes in classification schemes; individuals in different business units may make different classification decisions. To maximise the usefulness of data for subsequent analyses every effort should be made to create mappings among classification systems when they change or while they are

coexistent; trying to build those sorts of mappings years later is very complex and error prone.

- *Ensuring that all items are classified.* When significant numbers of items are unclassified, analytical results can be misleading and there is a large potential for systematically misleading biases. Delays in loading classification data can degrade operational decision processes.
- *Maintaining different levels and dimensions of classification.* It is often the case that different levels or dimensions of classification are needed to support different applications. Classification schemes that aggregate to too small a number of categories will miss important, potentially actionable, particular situations; too many categories may obscure general, strategic patterns from being discovered.
- *Dealing with multiple classifications (ie one item can be classified to multiple categories).* These situations are difficult because representing them may require a more complex data model and querying will require more complex logic.

Classification data quality management demands that there be discipline and schedules with respect to loading data as well as ongoing verification. Classification tables should be tested by using them with detailed transactional data and creating the points in time type comparison illustrated in Table 1.

Enhancement or enrichment

Modelling success levels are usually improved when more relevant data are available, so both marketers and modellers should be constantly thinking about adding new sources of information. When outside sources are providing data to load

to the marketing database (eg demographics) there are several quality related vulnerabilities created for modelling activities. Different sources of data often have different and contradictory information (eg income) for the same customer; different sources will also code data differently. Data supplied by the same vendor may also change significantly over time (eg the vendor's method of estimating income changes). The schedule for loading enhancement data may also seriously affect the use of data variables. For example, if a relatively long time has passed since enhancement data were acquired, then the absence of such information will correlate highly with being a new customer. In all of these situations, the utility of these data to the modeller is seriously diminished if there is no effective documentation, calibration and standardisation. The points in time type comparison should be made with special attention paid to the quantities of missing values.

Archiving

In the realm of operational information processing, archiving older data is done to save costs related to online storage of large data volumes. The archiving rules can be developed based on the likelihood that a piece of information will ever be needed again. Marketing databases, however, need to be able to describe a customer's behaviour over time which makes archiving a potentially serious impediment to accurate analyses. A number of types of problems can occur when data have been archived.

- If archiving is done only for currently inactive customers, then all analyses of performance by cohort groups of new customers become either biased or much more difficult to execute. Those sorts of studies, which include

- customer lifetime value analyses, depend on being able to track subsequent activity for all customers who were acquired at a point in time.
- If archiving is done for all transactions that occurred prior to a certain date, then aggregate measures of activity (eg total sales, number of orders) become either inaccurate or more complicated to calculate. Trends in customer activity may become impossible to calculate.
 - Archiving schedules or changes in archiving rules become part of the interpretation of variables that are created for analysis or scoring. An undisciplined process of archiving (eg archiving when the system is running out of disk space) can create inconsistencies in variables over time.

With the economics of data storage that exist today, marketing database administrators should strongly resist archiving as a strategy and wherever possible should use either less expensive data storage media or implement data compression technology. If archiving must be deployed, it should be documented and data analysts should be made aware of the practices that affect their data sets.

Scoring data sets

The commercial value of a customer model for marketing is realised when the model is applied to a current version of customer data. At that point in the process the potential model value can dissipate partially or completely if there are data quality issues. The normal marketing progression from model building to customer scoring is based on the assumption that variables used have the same essence and impact at both phases in the process. Unfortunately, it is not infrequent that this assumption is violated, often due to issues with

classification, enhancement or archiving. When production systems are set up to carry out scoring processing, it is often the case that as long as the input file layout meets specifications, there is no check on erroneous scoring. In a production environment this can very quickly lead to erroneous selections and significant financial consequences.

The key check that should be in place is a comparison of model building and scoring data sets in terms of value distributions by variable (ie Table 1 format, again). If statistically significant differences are found for variables that have an impact on the model, there should be an inspection of the differences by a modelling professional. Sometimes the differences, though statistically significant, may be benign if they are explainable as characteristics of the business or if they have minimal impact on customer scores. In many cases, however, in order to avoid a financial fiasco it is necessary to either fix an error and regenerate the scoring data set or find an alternative to use of the model.

Commercial world problems with the application of statistical modelling of customer data often have their origins in one or more of the issues discussed above. Marketing database design and maintenance responsibilities are often organisationally very separate from the modelling function; the training and background of professionals in the respective areas are generally very different. Consequently, organisational and expertise gaps with respect to data quality are created and go undetected with significant detrimental effects.

SUMMARY

The area of marketing database quality and its impact on customer data analysis has not been systematically addressed by database marketing professionals. Several key

functional constituencies — information technology, data analysis (modelling) and marketing — will all be able to achieve higher levels of performance by understanding and implementing fundamental principles of quality management. There is also a developed set of knowledge and approaches to the subject of data quality management that can be applied to marketing databases. The specific characteristics of marketing databases and their applications suggest particular quality management approaches.

References

- 1 Dobyns, L. and Crawford-Mason, C. (1991) 'Quality or else', Houghton Mifflin, New York.
- 2 Crosby, P. B. (1980) 'Quality is free: The art of making quality certain', Penguin Group, New York.
- 3 Loshin, D. (2001) 'Enterprise knowledge management: The data quality approach', Morgan Kaufmann, San Diego.
- 4 *Ibid.*
- 5 English, L. P. (1999) 'Improving data warehouse and business information quality: Methods for reducing costs and increasing profits', John Wiley & Sons, New York.
- 6 Redman, T. C. (1997) 'Data quality for the information age', Artech House, Norwood, MA.
- 7 Wang, R. Y., Ziad, M. and Lee, Y. W. (2001) 'Data quality', Kluwer Academic Publishers, Boston.
- 8 Date, C. J. (2000) 'An introduction to database systems', Addison Wesley, Reading, MA.
- 9 See www.acxiom.com/abilitec for information on Abilitec.
- 10 See www.experian.com/truvue/index_us.html for information on TruVue.
- 11 See www.fairisaacs.com/pdf/smartlink for information on Smartlink.
- 12 See www.auditonline.experian.com for an online demonstration.
- 13 Gendelev, B. (1989) 'Making your marketing data usable', *Direct Marketing*, October. Also available as an article at www.auditonline.experian.com.