Tradeoffs in Managing the Quality of Marketing Data

Adir Even
Ben-Gurion University of the Negev, adireven@bgu.ac.il

G. Shankaranarayanan
Boston University, gshankar@bu.edu

Paul D. Berger
Bentley University, pberger@bentley.edu

Recommended Citation
Even, Adir; Shankaranarayanan, G.; and Berger, Paul D., "Tradeoffs in Managing the Quality of Marketing Data" (2009). AMCIS 2009 Proceedings. 279.
http://aisel.aisnet.org/amcis2009/279

Follow this and additional works at: http://aisel.aisnet.org/amcis2009

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 2009 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
Tradeoffs in Managing the Quality of Marketing Data

Adir Even
Ben-Gurion University of the Negev, Israel
adireven@bgu.ac.il

G. Shankaranarayanan
Boston University, USA
gshankar@bu.edu

Paul D. Berger
Bentley University, USA
pberger@bentley.edu

ABSTRACT
A large majority of work in database marketing deals with what to do with data when it is available. This paper focuses on an aspect of data that has been infrequently examined in the database marketing literature – managing quality of data resources from a profit perspective. The notion that “more is better” often prevails in quality management decisions, with very little consideration, if any, of cost. This paper suggests that such decisions should be driven by consideration of cost-benefit tradeoffs and profit maximization. It specifically addresses data-quality decisions which are relevant in the database marketing area: the time-span covered by and targeted quality levels within datasets. These decisions are routinely made based on satisfying technical and functional requirements. We propose a model that quantifies the benefits and costs associated with these decisions and helps maximize profit. The paper describes the model development, discusses its implications for data-quality management decisions, and highlights its contributions with illustrative examples.

Keywords
Database Marketing, Interactive Marketing, Data Quality Management.

INTRODUCTION
A large majority of work in database marketing deals with what to do with data when it is available. The efficiency of and the benefits gained from information system environments that support database/interactive marketing depend largely on the data within – customer profiles, transaction history, past contact efforts, and promotions. We argue that as the use of data resources for interactive marketing grows, along with related costs, it is important to understand associated economic factors. This paper focuses on an aspect of data management visited infrequently in the database/interactive marketing – optimizing data quality management (DQM) decisions from a profit perspective. In this study, we consider the association between high quality of data and the business value it generates, the costs for achieving and sustaining high data quality, and the overall profitability. We argue that the economic aspects of DQM in the context of database marketing are understudied and deserve further examination.

Our study addresses decision regarding the subset of records in the dataset targeted for quality improvement (reflecting time-span covered), and targeted quality levels (the rate of non-defective data). In real-world marketing systems, these (DQ) decisions are made based on satisfying technical (e.g., processing and storage capacity) and functional requirements (e.g., from analysts). We suggest that DQM decisions should also be based on economic considerations. We argue that the technical and economical factors should be addressed simultaneously; i.e., maximizing the overall business benefits should inform and guide DQM decisions.

To highlight the managerial concerns this research seeks to explore, we use an illustrative case of “Great-Store,” a retail firm that manages a chain of stores and sells online. “Great-Store” has recorded its sales transactions in a database. With loyalty cards and with the introduction of online sales, transactions are associated with individual customers. Marketing analysts within the firm are using the data for studying consumer behavior, list segmentation and customer lifetime value. “Great-Store” is hence considering several quality improvements to its dataset. Some key DQM decisions are: (1) how many years of data should be audited and improved? (2) What level of quality should be targeted? (3) Should all data attributes be treated the same, or should the firm consider different quality improvement policies for different attributes?

These decisions introduce cost-benefit trade-offs and impact the firm’s profit. If strictly following a technical/functional approach, “Great-Store” might attempt to improve all the records in its datasets along all possible attributes and target perfect
quality. Our study suggests that a strict technical/functional approach for DQM is likely to be economically suboptimal. It proposes a framework for assessing such DQM decisions for maximizing economic performance. The objective of this framework is to maximize net-benefit, defined as the difference between value and cost.

This study addresses the issue of data quality improvement in data environments that support marketing, such as Customer Relationship Management (CRM). Customer data is vulnerable to data quality defects. Two common types of quality defects in CRM environments are: a) Missing attribute values: some attribute values may not be available when initiating a customer profile (e.g., income and credit score). The firm may leave these blank and update them later.Existing profiles may be enhanced with new attributes (e.g., email and mobile number), and the corresponding values are initially null. b) Failure to keep attribute values up-to-date: some attribute values may change over time (e.g., address, phone number, and occupation). If not maintained, the data becomes obsolete and the firm loses the ability to reach associated customers. In this study, we associate the presence of quality defects with benefits and costs.

An important DQM decision in marketing databases is the time-span targeted for quality improvement. When managing a large number of records, would it make economic sense to maintain all records at a high quality, or would it be better to focus on a subset? We link this question to the age of the data – can we maximize profit by improving only the newer records (e.g., profiles associated with recent purchases)? The decision that we seek to evaluate is, what time-span of data (e.g., recent 50%, recent 80%, or, the entire dataset) should be improved? Improving the quality of a larger time-span offers a larger high-quality dataset, and thus a more detailed perspective of the business. The drawback is that it costs more to achieve.

The next section provides the background for developing the profit maximization framework. The framework is then used to develop a quantitative model for the assessing and optimizing DQM decisions. The model’s application is subsequently illustrated. Concluding remarks and directions for further research are presented at the end.

A MODEL FOR OPTIMIZING DATA QUALITY DECISIONS IN MARKETING DATABASES

Our model is based on the framework for configuring tabular datasets proposed in Even et al. (2007) and is influenced by Rust et al (2004), and Ballou and Pazer (1995, 2003). We adapt the model for managing data quality in marketing databases and supplement the model’s development with illustrative examples that demonstrate its use. We first describe the general formulation of the key constructs, which is then extended to a more specific model.

The General Framework - Utility (U), Cost (C), and Net-Benefit (B)

The general framework links configuration decisions (represented by $X$, a vector of decision variables) to economic outcomes – utility, cost, and net-benefit.

**Utility (U):** The utility measure reflects the business contribution of a dataset, measured monetarily. We assume $I$ possible utility-contributing usages, indexed by $[i]$. The effect of DQM decisions (the vector $X$) on utility is represented as a set of utility functions, one per usage, and the overall utility is assumed sum-additive. Based on this we model utility as:

$$U(X) = \sum_{i=1}^{I} U_i(X)$$

where

- $U$ – The overall utility
- $X$ – The vector of data management decisions
- $I$ – The total number of usages
- $U_i$ – Utility of usage $[i]$

**Cost (C):** Costs of DQ improvements are driven by error detection and correction efforts, payments to data vendors, software programming, and administrative overheads. We represent the cost components as a parameterized function that translates the effect of DQ decisions ($X$) to monetary measurements. We group these into two – a variable component, $C^V$, which reflects costs that are affected by data quality decisions (the vector $X$), and a fixed component, $C^F$, that is independent of factors such as infrastructural hardware and networks, and managerial overhead. That is, we combine all $J$ costs for a given configuration into a “total” fixed cost and a “total” variable cost:

$$C(X) = \sum_{j=1}^{J} C_j(X) = C^F + C^V(X)$$

where

- $C$ – The overall cost
- $X$ – The vector of decision variables
- $J$ – The total number of cost components
variable available. We also assume that our dataset is currently at some quality-ratio (a set of attribute-level decisions). We would like to focus on more recent data. The time-span decision \( T \) defines a cut-off age—we will focus our efforts on records younger than this age. We represent the time-span as a \([0, 1]\) continuous variable representing the proportion of the maximum time span covered by the dataset \( T^0 \). So, \( T=1 \) implies the entire dataset, \( T=0.5 \) the most recent half of the dataset, and so on. Obviously, a larger \( T \) implies a larger number of records targeted for DQ efforts, and hence, higher costs.

**Quality \( Q \):** Quality is measured as a ratio in the range of 0 (bad) to 1 (good), which reflects the proportion of non-defective data elements (Pipino et al., 2002). Targeting a higher quality-ratio increases the dataset’s value, but may also increase cost. In our model, we consider two levels of measuring the quality-ratio, and accordingly, develop two variations of our model:

(a) **Record-level measurement** – This assumes that all attributes are treated alike with respect to quality improvement. The decision variables are, \( T \), how much of the dataset to target for quality improvement and, \( Q \), the target quality-ratio to which the dataset’s quality is to be improved. We assume that we have the ability to cover, \( T^\delta \), the largest time-span of data available. We also assume that our dataset is currently at some quality-ratio \( Q^\delta \), which can be determined. Our decision variable \( Q \) reflects the targeted quality-ratio. We may choose to leave the dataset at its current quality \( (Q=Q^\delta) \), improve it to perfection \( (Q=1) \), or target a quality-ratio in-between these two. (b) **Attribute-level measurement** – a tabular dataset has multiple attributes and each record contains a set of values (including null) for these attributes. You may specifically define the quality-ratio for each attribute (a set of attribute-level decisions). We would be interested in improving as many attributes as possible as more high-quality attributes implies a richer dataset. However, not all attributes are equally important from a database-marketing perspective. Targeting a large set of attributes may increase DQM costs, resulting in a sub-optimal decision. We therefore may target a small subset of attributes and ignore others. To model this decision, we assume a set of \( M \) attributes (indexed by \( [m] \)). We define the quality-ratio for attribute \( [m] \) as a \([0, 1]\) proportion of records in which \( [m] \) has non-defective values. Our dataset has a given quality-ratio, \( Q^\delta_n \) for attribute \( [m] \). Accordingly, we define a set of quality-ratio decision variables \( \{Q_{m}\}_{n=1,M} \) each reflecting the targeted quality-ratio of attribute \( [m] \) \( (Q^\delta_n \leq Q_m \leq 1) \).

We first develop a record-level model for optimizing time-span \( T \) and quality-ratio \( Q \) and then extend it to include attribute-level quality-ratio decisions \( \{Q_{m}\} \). In both, we assume that \( Q \) does not change with \( T \).

A Record-Level Model – Optimizing Time-span \( T \) and Quality-ratio \( Q \)

The time-span \( T \) is a record-level decision — records younger than the cut-off age defined by \( T \) will be included for quality improvement and older records will remain at the original quality level, \( Q^\delta \). We determine the targeted quality-ratio \( Q \) at the record level — the targeted proportion of non-defective records in the subset that we choose to improve. Records that we ignore (between \( T \) and 1) are at the original quality-level, but, still present (and usable).

We assume that the utility for usage \( [i] \) has a certain cap \( u_i \), which is reached when the entire dataset is included \( (T=1) \), and maintained at perfect quality \( (Q=1) \). Utility grows with the time-span \( T \) but, since more recent data is assumed to be more valuable, it increases at a decreasing rate. We use the Cobb-Douglas production function to represent this behavior — \( T \), bounded between \( 0 \) and 1, raised to the positive power of a degradation parameter \( a_i \). This parameter, which reflects the degradation in utility for usage \( [i] \), is between \( 0 \) and 1, as we assume that utility increases at a decreasing rate. The utility per use decreases with a higher proportion of defective records (lower \( Q \), bounded between \( Q^\delta \) - the current quality-ratio - and 1). Again, we formulate this using the Cobb-Douglas production function - \( Q \) raised to the power of a degradation parameter \( \lambda_i \). In this case, we don’t limit the assumption to a concave effect; hence, \( \lambda_i > 0 \).
Records within the targeted time-span will be improved and their utility will increase. The quality-ratio of records with relative age between \( T \) and \( l \) will not be improved and their utility will remain low. We represent \( U_i \), the utility of usage \([i]\) (Figure 1a) as:

\[
U_i(T, Q) = u_i \left( T^{\alpha_i} Q^{\lambda_i} \left(1 - T^{\alpha_i}\right) \left(Q^S\right)^{\delta_i} \right)
\]

where
\( U_i \) – The utility of usage \([i]\)
\( T, Q \) - The time-span covered and quality-ratio, respectively – the decision variables
\( u_i \) – The utility cap of usage \([i]\), at \( T=1 \) and \( Q=1 \)
\( \alpha_i \) – A positive time-span sensitivity factor of usage \([i]\), \( 0 < \alpha_i \leq 1 \). The smaller the value of \( \alpha_i \), the less dependent the usage is on older data.
\( \lambda_i \) – A positive quality sensitivity factor of usage \([i]\). The greater the value of \( \lambda_i \), the more sensitive the usage is to loss of quality.

The values of \( u, \alpha, \) and \( \lambda \) can be solicited using decision calculus. Decision calculus has been successfully applied to evaluate the parameters of a sales response function to advertising and to determine the optimal time limits for email price promotions (Hanna et al., 2005), among many other applications.

Like utility, costs also increase with the number of records and with the quality level, \( Q \). We define \( N^S \) as the total number of records in the dataset. Assuming that the number of records grows in a more-or-less fixed rate over time and has an identical variable-cost per record, the variable cost for the dataset will be linearly proportional to \( N^S T \). Improving the quality, beyond the guaranteed \( Q^S \) (the given quality-ratio), will also increase variable cost. We split the variable cost into two components: \( c_{V_0} \), the variable costs that are not affected by the target quality-ratio (e.g., the time to retrieve and display the record), and \( c_{V_q} \), costs that increase with targeted quality-ratio (e.g., the time to audit and correct data). We assume that this increase is proportional to the ratio between the targeted \( Q \) and the given \( Q^S \) to a certain power, where the power parameter \( \delta \) reflects the cost sensitivity to quality improvement. The greater the \( \delta \), greater is the increase in cost as we approach perfect quality. We envision that \( \delta > 1 \) as a convex cost structure that is typical (examples abound in level of customer service and manufacturing contexts) with quality. Under these assumptions, the overall cost (Figure 1b) is:

\[
C(T, Q) = c^F + \left( c_{V_0} + c_{V_q} (Q/Q^S)^\delta \right) N^S T
\]

where
\( C \) – The overall cost
\( T, Q \) - The time-span and the quality-ratio, respectively – the decision variables
\( N^S \) - The total number of records in the dataset
\( Q^S \) – The given quality-ratio, guaranteed at no additional cost
\( c^F \) – The fixed cost parameter
\( c_{V_0} \) – The variable cost parameter, reflecting costs per record that do not depend on \( Q \)
\( c_{V_q} \) – The variable cost parameter, reflecting \( Q \)-dependent costs per record, at \( Q=Q^S \)
\( \delta \) – Cost sensitivity parameter to the quality improvement.

The net-benefit, considering \( T \) and \( Q \), can be derived by consolidating (4) and (5):
\[ B(T, Q) = \sum_{i=1}^{I} u_i \left( T^\alpha Q^\lambda_i + \left(1 - T^\alpha\right) \left( Q^S \right)^\lambda_i \right) - \left( c_F + c^V_0 + c^V_\theta \left( Q / Q^S \right)^\delta \right) N^S T \] 

This extended formulation of the objective function in (3) can be stated as choosing time-span and targeted quality-ratio such that the overall net-benefit is maximized. Though the model is non-linear we can achieve tractable solutions with certain simplifications - e.g., fixing values for one decision variable and optimizing only the other. In our formulation, time-span is limited by \( 0 \leq T \leq 1 \), and targeted quality is bounded by \( Q^S \leq Q \leq 1 \). The net-benefit has apparent tradeoffs with \( T \) and \( Q \). We expect the net-benefit to (in each bullet, all other parameters are assumed the same):

- Increase with \( I \) and \( \{u_i\} \) – more usages and higher per-use utility increases net-benefit.
- Decrease with \( \{a\} \) – a lower degradation parameter implies that near-optimal (net-benefit) can be achieved, (even) with relatively smaller time-span coverage.
- Decrease with \( \{\lambda\} \) – higher quality sensitivity implies higher utility decline as quality degrades, hence, lower net-benefit.
- Decrease with \( \{c\} \) – higher costs decrease net-benefit.
- Decrease with \( \{\delta\} \) – a higher cost sensitivity to quality-ratio decreases net-benefit.
- Increase with \( Q^S \) – a higher initial quality-ratio reduces the potential extent of quality improvement, thus lowering costs and increasing net-benefits.

**Illustrative Example - Part 1 – Configuring Time-span and Quality-ratio:** The following prototypical example is stylized to illustrate the use of the framework. Numerical values were confirmed to be acceptably realistic by marketing executives.

Marketing specialists at “Great-Store” use the firm’s customer list and sales transactions to promote new products. Currently, the database covers 10 years of activity with average of 100,000 new customer records per year. A preliminary quality assessment revealed that \( Q^S = 0.5 \). The customer dataset and associated sale transactions are used for: a) managing direct mail promotions – targeting the entire customer base (10 years of data, \( T = 1 \)) is expected to yield $4 million in revenue, assuming perfect data quality \( (Q = 1) \). Customers associated with more recent purchases are expected to contribute more to this revenue. The sensitivity of revenue to time-span is estimated at 0.32 (i.e., 60% of the revenue will come from customers who are associated with activity in the last 2 years, or 20% of the dataset). The sensitivity of the revenue to quality defects is estimated at 0.74 (i.e., the current quality of 0.5 reduces the maximum possible revenue (at \( Q = 1 \)) by 40%). b) Managing phone-based promotions – targeting the entire customer base \( (T = 1) \) is expected to yield $2 million at \( Q = 1 \). Customers associated with recent purchases are expected to contribute more to the revenue. The estimated sensitivity of revenue to time-span is 0.57 (40% of revenue will come from customers who make up 20% of the dataset). The estimated sensitivity of revenue to quality defects is 1.73 (the current quality of 0.5 will reduce the maximum possible revenue (at \( Q = 1 \)) by 70%).

We estimate an average cost per record of $0.06 to identify errors, (time to retrieve, observe, and validate data, regardless of whether changes are made). An additional cost of $0.08 per record is sensitive to the targeted quality-ratio, with a sensitivity factor of 7 – i.e., improving the quality-ratio from \( Q = 0.5 \) to \( Q = 1 \) will raise this cost 128 times, to about $10.24 per record, on an average. In addition, the estimated fixed cost of data-quality improvement is $100,000. To maximize net-benefit, two questions should be addressed: (a) what proportion of the recent records \( (T) \) should be targeted for quality-improvement and, (b) what quality-ratio should be targeted for the subset of records, \( (N^S \cdot T) \), covered?

In terms of model parameters, the problem is:

- \( I = 2 \), as two usages are identified. For the first, the maximum utility is \( u_1 = 4,000,000 \), the sensitivity of utility to time span is \( a_1 = 0.32 \), and the sensitivity of utility to quality is \( \lambda_1 = 0.74 \). For the second, the maximum utility is \( u_2 = 2,000,000 \), the sensitivity of utility to time span is \( a_2 = 0.57 \), and the sensitivity of utility to quality is \( \lambda_2 = 1.73 \).
- The dataset covers 10 years of activity and grows linearly at 100,000 records a year; hence, \( N^S = 1,000,000 \), and the number of records covered as a function of \( T \) is \( 1,000,000 T \).
- The estimated fixed cost is \( c_F = 100,000 \), and the estimated variable cost per record that does not depend on the targeted quality-ratio is \( c^V_0 = 0.06 \).
- The estimated variable cost is \( c^V_\theta = 0.08 \), with a sensitivity factor of \( \delta = 7 \), and a given quality of \( Q^S = 0.5 \).

Using these parameters, we evaluate 9 different DQM policies (Table 1), using the model (Eq. 6).
Even et al.  Tradeoffs in Managing the Quality of Marketing Data

The net-benefit ($B=\$3.366M$) at the optimum (policy-$H$, $T=0.14$, $Q=0.8$), is higher than “doing nothing” (policy-$A$) - although no loss is associated in this case (as no cost is involved). Policy-$C$ and policy-$D$ can be interpreted as applying “heuristics” – target a small subset of newer records ($T$ chosen arbitrarily), and improve quality to some level. Such policies appear reasonable. However, they are unlikely to maximize net-benefit, though, in our example, they are fairly close. The net-benefit from optimizing $T$ alone for a maximum $Q$ (policy $G$) is lower than the net-benefit from jointly optimizing both. The same holds for optimizing $Q$ for minimum or maximum $T$ (policies $E$ and $F$, respectively).

Maximizing both $T$ and $Q$ (policy-$B$) causes a substantial net-loss. DQM literature suggests that near-perfect quality-ratios can be obtained when proper quality improvement methodologies and applications are implemented. However, these are likely to be non-optimal for profitability. The usefulness of our proposed model is critical to understand such approaches.

### Addressing Attribute-Level Decisions

To refine the model, we now assume that the sensitivity of each usage to quality varies with attributes. Some usage may require specific attributes at nearly perfect quality for that usage to be profitable, i.e., the usage is moderately sensitive to the quality of these attributes. A different usage may tolerate “less-than-perfect” quality of these attributes. A different usage may not need certain attributes and is insensitive to quality defects in these attributes.

We assume that the utility per use increases with $Q_m$, the proportion of records with non-defective values in attribute $[m]$, which is bounded between $Q^*_m$ (the given quality of attribute $[m]$) and $1$. We formulate this effect using the Cobb-Douglas production function - $Q_m$ raised to the power of a degradation parameter $\lambda_{q,m} \geq 0$. The overall effect of quality on usage $[i]$ is modeled as the geometric average of the effect of the quality of the attributes: $\left( \prod_{m=1..M} Q_m^{\lambda_{q,m}} \right)^{1/M}$ . Our quality improvement efforts will add records within the targeted time-span (relative age between $0$ and the chosen $T$) and the utility associated with these records will increase accordingly. The quality of records with relative age between $T$ and $J$ will not be improved (i.e., will remain at $Q_m^n$) and, accordingly, the utility associated with these records will remain low. We now enhance the utility model (4) to include these attribute-level considerations. The overall utility is given by:

$$U(T, Q_m) = \sum_{i=1..I} U_i \left( T^{\alpha_i} \left( \prod_{m=1..M} Q_m^{\lambda_{q,m}} \right)^{1/M} + \left( 1 - T^{\alpha_i} \right) \left( \prod_{m=1..M} Q_m^{\delta_{q,m}} \right)^{1/M} \right)$$

Examining attribute-level decisions on cost, we assume (like the record-level model) that the variable cost per record has a baseline component $c^V_0$, which does not depend on the candidate attributes. However, we break the $c^V_m$ component in (5) to incremental fixed cost factors $c^V_m$, each reflecting the cost added when the corresponding attribute $[m]$ is included. We assume that this incremental increase is proportional to the proportion between the targeted quality in attribute $[m]$ $Q_m$ and the given $Q^*_m$ to a certain power. The power parameter, $\delta_{q,m} \geq 1$, reflects the sensitivity of cost to quality improvement in attribute $[m]$. Correspondingly, the cost and the net benefit expressions (Eq. (5) and (6), respectively) can be extended as:

$$\begin{align*}
C(T, Q_m) &= c^F + c^V_0 + \sum_{m=1..M} c^V_m Q_m \left( Q_m / Q^*_m \right)^{\delta_{q,m}} N^ST \\
B(T, Q_m) &= \sum_{i=1..I} U_i \left( T^{\alpha_i} \left( \prod_{m=1..M} Q_m^{\delta_{q,m}} \right)^{1/M} + \left( 1 - T^{\alpha_i} \right) \left( \prod_{m=1..M} Q_m^{\delta_{q,m}} \right)^{1/M} \right) - \\
&\left( c^F + c^V_0 + \sum_{m=1..M} c^V_m Q_m \left( Q_m / Q^*_m \right)^{\delta_{q,m}} \right) N^ST\end{align*}$$

<table>
<thead>
<tr>
<th>Decision Policy</th>
<th>$T$</th>
<th>$Q$</th>
<th>$U$ (SM$)$</th>
<th>$C$ (SM$)$</th>
<th>$B$ (SM$)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A - Do Nothing – accept current $Q$ for all records</td>
<td>1</td>
<td>0.5</td>
<td>2.997</td>
<td>0</td>
<td>2.997</td>
</tr>
<tr>
<td>B - Maximize $Q$ for all records</td>
<td>1</td>
<td>1</td>
<td>6.000</td>
<td>10.400</td>
<td>-4.400</td>
</tr>
<tr>
<td>C - Maximize $Q$ for an arbitrary $T=0.1$</td>
<td>0.1</td>
<td>1</td>
<td>4.142</td>
<td>1.130</td>
<td>3.012</td>
</tr>
<tr>
<td>D - Target $Q=0.75$ for an arbitrary $T=0.2$</td>
<td>0.2</td>
<td>0.75</td>
<td>3.742</td>
<td>0.385</td>
<td>3.358</td>
</tr>
<tr>
<td>E - Optimize $Q$ for an arbitrarily small $T=0.1$</td>
<td>0.1</td>
<td>0.83</td>
<td>3.741</td>
<td>0.377</td>
<td>3.363</td>
</tr>
<tr>
<td>F - Optimize $Q$ for all records</td>
<td>-1</td>
<td>0.66</td>
<td>3.907</td>
<td>0.710</td>
<td>3.197</td>
</tr>
<tr>
<td>G - Optimize $T$ for a maximum $Q$</td>
<td>0.03</td>
<td>1</td>
<td>3.657</td>
<td>0.353</td>
<td>3.303</td>
</tr>
<tr>
<td>H - Optimize $T$ and $Q$, simultaneously</td>
<td>0.14</td>
<td>0.80</td>
<td>3.778</td>
<td>0.411</td>
<td>3.366</td>
</tr>
</tbody>
</table>

Table 1. Optimizing Time Span ($T$) and Quality-ratio ($Q$)
Illustrative Example - Part 2 - Optimizing DQM at the attribute level: Managers at “Great-Store” want to examine some specific attributes more closely and consider different quality improvement policies for each attribute. Four key attributes are considered: 1) Years-of-Education, 2) Neighborhood-Ranking, 3) Credit-Status, and 4) Value-of-Property Owned. The current quality-ratio for each is 0.60, 0.45, 0.40, and 0.5, respectively. Each attribute contributes a different added margin to quality improvement costs (an average of $0.07, $0.05, $0.06, $0.08 per record, respectively), and is associated with different cost-sensitivity to the increase in the targeted quality (2, 4, 5, and 7, respectively).

The first usage, Mail-Based Promotion (see illustrative example, part 1), is highly sensitive to the quality of Years-of-Education and Neighborhood-Ranking (sensitivity parameters of 1.9 and 1.7, respectively), less sensitive to the quality of Credit-Status, and even less sensitive to the quality of Value-of-Property-Owned (sensitivity parameters – 0.7, and 0.1, respectively). The second usage, Phone-Based Promotion is very sensitive to defects in Years-of-Education (sensitivity parameter of 2.0), less sensitive to the quality of Neighborhood-Ranking and Credit-Status (sensitivity parameters – 1.3, and 0.9, respectively), and even less sensitive to the quality of Value-of-Property-Owned (sensitivity parameters of 0.2).

To optimize net-benefit, the following questions need addressing: (a) what time span should be targeted for quality improvement? (b) What data quality level should be chosen for each candidate attribute? We first translate the given inputs in terms of model parameters. A few parameters are taken from the description in part 1:

- Two usages \(I=2\), the first usage with maximum utility of \(u_1 = 4,000,000\) and sensitivity to time span of \(\alpha_1=0.32\). The second with \(u_2 = 2,000,000\) and \(\alpha_2=0.57\).
- 10 years of data with a linear growth in the number of records over time: \(N(T)=1,000,000T\)
- Fixed cost: \(c_F = 1,000,000\), and base variable cost: \(c_V^0 = 0.06\)
- The other attribute-level parameters are inferred from the description above and summarized in Table 2 – sensitivity of utility to quality \((\lambda_{i,m})\), given quality \((Q_m)\), added marginal cost \((C^0_m)\), and and sensitivity of cost to quality \((\delta_m)\).

To understand the impact of various parameters, we evaluate 8 policies (Table 3) using the optimization model (Eq. 9). The net-benefit gained by optimizing both (policy-H, \(B=3.681M\)), is superior to “doing nothing” (policy-A, \(B=3.037M\)). Similar to part 1, targeting perfect quality in all attributes and records will result in a significant net-loss (policy-B), and applying heuristics (C and D) is sub-optimal. Maximizing quality (targeting \(Q_m=1\)) can yield positive and relatively high net-benefit (compared to the optimum), only if applied to a small subset of the most recent records. In C, ~61% of the optimal net-benefit could be gained by maximizing quality for the most recent 10% of the dataset, and in policy G, ~79% of the optimum could be gained by setting \(Q_m=1\) for the recent 5% of the dataset.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>(\lambda_{1,m})</th>
<th>(\lambda_{2,m})</th>
<th>(Q_m^1)</th>
<th>(C^0_m)</th>
<th>(\delta_m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years of Education</td>
<td>1.9</td>
<td>2.0</td>
<td>0.60</td>
<td>0.07</td>
<td>2.00</td>
</tr>
<tr>
<td>Neighborhood Ranking</td>
<td>1.7</td>
<td>1.0</td>
<td>0.45</td>
<td>0.05</td>
<td>4.00</td>
</tr>
<tr>
<td>Credit Status</td>
<td>0.7</td>
<td>0.9</td>
<td>0.40</td>
<td>0.06</td>
<td>5.00</td>
</tr>
<tr>
<td>Value of Property Owned</td>
<td>0.1</td>
<td>0.2</td>
<td>0.50</td>
<td>0.08</td>
<td>7.00</td>
</tr>
</tbody>
</table>

Table 2. Attribute-Level Model Parameters

To optimize net-benefit, the following questions need addressing:

- Two usages \(I=2\), the first usage with maximum utility of \(u_1 = 4,000,000\) and sensitivity to time span of \(\alpha_1=0.32\). The second with \(u_2 = 2,000,000\) and \(\alpha_2=0.57\).
- 10 years of data with a linear growth in the number of records over time: \(N(T)=1,000,000T\)
- Fixed cost: \(c_F = 1,000,000\), and base variable cost: \(c_V^0 = 0.06\)
- The other attribute-level parameters are inferred from the description above and summarized in Table 2 – sensitivity of utility to quality \((\lambda_{i,m})\), given quality \((Q_m)\), added marginal cost \((C^0_m)\), and and sensitivity of cost to quality \((\delta_m)\).

<table>
<thead>
<tr>
<th>Decision Policy</th>
<th>(T)</th>
<th>(Q_1)</th>
<th>(Q_2)</th>
<th>(Q_3)</th>
<th>(Q_4)</th>
<th>(U)(SM)</th>
<th>(C)(SM)</th>
<th>(B)(SM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A - Do Nothing - accept current (Q_) for each (m), for (T=1)</td>
<td>1.00</td>
<td>0.60</td>
<td>0.45</td>
<td>0.40</td>
<td>0.50</td>
<td>3.037</td>
<td>0</td>
<td>3.037</td>
</tr>
<tr>
<td>B - Maximize - target maximum (Q_) for each (m), for (T=1)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>6.000</td>
<td>17.673</td>
<td>-11.673</td>
</tr>
<tr>
<td>C - Maximize all ((Q_)) for an arbitrary (T=0.2)</td>
<td>0.10</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>4.115</td>
<td>1.857</td>
<td>2.258</td>
</tr>
<tr>
<td>D - Target arbitrary (Q_)=0.75 for all (m), for an arbitrary (T=0.2)</td>
<td>0.20</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>3.641</td>
<td>0.762</td>
<td>2.879</td>
</tr>
<tr>
<td>E - Optimize each ((Q_)) for an arbitrarily small (T=0.1)</td>
<td>0.10</td>
<td>1.00</td>
<td>1.00</td>
<td>0.68</td>
<td>0.51</td>
<td>3.897</td>
<td>0.340</td>
<td>3.557</td>
</tr>
<tr>
<td>F - Optimize each ((Q_)) for all records</td>
<td>1.00</td>
<td>1.00</td>
<td>0.78</td>
<td>0.48</td>
<td>0.50</td>
<td>4.659</td>
<td>1.070</td>
<td>3.399</td>
</tr>
<tr>
<td>G - Optimize (T) for a maximum ((Q_)) for all (m)</td>
<td>0.01</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>3.418</td>
<td>0.316</td>
<td>3.102</td>
</tr>
<tr>
<td>H - Optimal - Optimize (T) and ((Q_)), simultaneously, for all (m)</td>
<td>0.39</td>
<td>1.00</td>
<td>0.93</td>
<td>0.57</td>
<td>0.50</td>
<td>4.392</td>
<td>0.711</td>
<td>3.681</td>
</tr>
</tbody>
</table>

Table 3. Optimizing the Time Span \((T)\) and the Quality Level \((Q)\)

The differences in net-benefit between optimizing \(T\) alone (policy-G), optimizing \(|Q_m|\) alone (policy-E and policy-F), and optimizing both simultaneously (policy-H) are relatively small. However, we notice significant differences between attributes in policies that attempted to differentiate their levels of quality (policies E, F, and H). The quality of Years-of-Education was
maximized under all policies. This can be explained by the fact that both usages are sensitive to defects in this attribute, and the cost of improving it is relatively low (low cost sensitivity). The quality of Neighborhood-Ranking was maximized only in policy-E (only a small subset of records was targeted), and the quality of Credit-Status was improved only to an extent by all policies. Both usages are sensitive to the quality of these two attributes; however, the cost of improving these attributes increases significantly when high quality-ratios are targeted (high cost-sensitivity parameters). Finally, the quality of Value-of-Property-Owned was improved minimally as both usages were relatively insensitive to defects in this attribute.

CONCLUSIONS

Our study suggests that DQM decisions should align with profit-maximization. It contributes by proposing a profit-maximization framework for evaluating such decisions. The framework is demonstrated through developing optimal DQM policies for a tabular-dataset in the context of database marketing. The model illustrates cost/benefit trade-offs with two key DQM decisions: the subset of data covered and the targeted quality level in this subset. As demonstrated, “more is not necessarily better” - increasing the number of records monitored, maintaining a larger number of attributes, and approaching perfect quality may have functional and technical merits, but may not be optimal in terms of profitability.

We believe that the proposed framework can effectively complement existing DQM methodologies and techniques. In reality, the economic performance associated with data quality in marketing data can be affected by many technical factors (e.g., the firm’s IT infrastructure), as well as business-related issues such as commitment to clients, information privacy, and legal constraints. This study is a step in incorporating economic considerations into managing the quality of marketing data. Examination of other factors and modeling their effect on utility and cost offers a range of opportunities for future research.

REFERENCES