

MANAGING DATA QUALITY AND TIME-SPAN COVERAGE FOR CRM – COST/BENEFIT TRADEOFFS

Adir Even
Ben Gurion University of the Negev
P.O. Box 653, Beer-Sheva, 84105,
Israel
adireven@bgu.ac.il

Ganesan Shankaranarayanan
Boston University, School of Management
595 Commonwealth Avenue
Boston, MA 02215
Shankar@bu.edu

Paul D. Berger
Bentley College
175 Forest St.
Waltham, MA 02452
pberger@bentley.edu

- 1) Dr. Adir Even is an Assistant Professor of Information Systems at the Department of Industrial Engineering and Management at Ben-Gurion University, specializing in data management, data warehousing, and business-intelligence applications. He can be reached at adireven@bgu.ac.il.
- 2) G. Shankaranarayanan is an Assistant Professor of Information Systems at the Boston University School of Management. He specializes in data modeling, data management, and data quality management solutions for business applications. He can be reached at gshankar@bu.edu
- 3) Paul D. Berger is Professor of Mathematical Sciences and Marketing at Bentley College in Waltham, Massachusetts. His interests include mathematical modeling of optimization problems, database marketing, data analysis, and experimental design. He is a member of several journal editorial boards. He can be reached at pberger@bentley.edu

ABSTRACT

A large majority of work in database marketing deals with what to do with data when it is available. Some work deals with how best to obtain data. Many authors have pointed out how important it is that “every contact” with a customer and prospect needs to be recorded and that there is very little in the database/interactive marketing arena that can be done without data. We might even parrot the real estate mantra of “location, location, location” to describe the price/value of a home, with “data, data, data,” for its importance in making virtually all meaningful marketing decisions.

This paper focuses on an aspect of “data” that has not been visited frequently in the database/interactive marketing area – that of optimizing the time span of data and its quality from a *profit* point of view. Most decisions of these types are routinely made based on satisfying technical and functional requirements (i.e., what capacity do we have to collect/carry data, and what are the requirements for the data?), with the notion that “more is better” always prevailing, essentially with very little consideration, if any, of cost. It is reminiscent of the thinking of the 1980’s, a time when virtually any “quality improvement” of a product was thought to be a clear action to take, until Total Quality Management and Taguchi Methods changed thinking significantly, arguing (rightfully) that quality improvements need to be *economically* justified.

In earlier research, we proposed a model that links design characteristics of a dataset to economic outcomes: utility, cost, and net-benefit (essentially, profit.) In this study, we use the model to address key dataset-configuration decisions that can have significant profit implications when dealing with large datasets: (a) *Time span*: time-span is typically defined by a “cut off” age threshold, and data administrators may consider managing older records differently (e.g., discarding or archiving). Increasing the time span broadens historical coverage and, hence, increases utility potential. However, a larger time-span also increases the number of records, and hence, data acquisition and maintenance costs; (b) *Targeted quality level*: The presence of quality defects in a data resource reduces its utility. In this study, we measure quality by the proportion of non-missing data. Of course, data quality can be measured in many ways, in addition to reflecting its impact within a specific usage context. Given a certain (current) data quality level, organizations may consider targeting a higher level of quality. Improving quality increases the utility of the data for making key marketing decisions. However, the higher the targeted quality level, the higher the associated data-improvement and maintenance costs.

We apply the model to a large data set from a real-world data resource used for managing alumni relations. Perhaps more importantly, we *evaluate* our dataset model as an artifact that helps configure data resources for improving economic results. Our evaluation is set in a CRM context. We show that, in this context, all model variables can be measured and all assumptions in building the model are well supported. Also, we show that using the model to optimize the time span and data quality of the alumni dataset can help revise current data acquisition and maintenance policies leading to a major increase in net benefit. Further, we show that our evaluation results offer important insights for enhancing the model, itself.