Relevant Marketing Based on Operational Databases, Surveys, Online Sources, and Social Media Data – The Potential of Integrating External Information Sources into the Customer Database: A Simulation Study Evaluating Different Sources by Means of their Specific Characteristics

**Introduction**

Today, marketing budgets are monitored closely. ROIs are calculated and information is sought to be conveyed to the right person at the right point and the right time. Relevance is the keyword dominating all activities, and data is its foundation. With the vast amount of data collected today, relevant communication should be very easy to deliver. But if the data is not the right data, if it cannot be interpreted, if it must not be used or if it is not sufficient in order to differentiate target groups, relevance degenerates to simple gender-age-region-schemes. Oftentimes, available information is not sufficient for narrowing down the target group in order to be relevant for the customers. Even if detailed information is available, it is most commonly available only for a small, highly active portion of the customers. The big portion of occasional and inactive customers is not well describable by the criteria sought. Other information may well be available and useful, but may not be allowed to be used. When information is not ready to hand, database marketers make assumptions in order to target customers, which may or may not describe the target criterion well. Additional knowledge about customers is often available in aggregated form only, described by variables not present in the customer database. Although these segments are very well describable and campaign actions are clear, it is not possible to identify these segments in order to treat them individually.

Because of these limitations, ad space cannot be used efficiently, the ad burden is perceived high if the information is irrelevant, and costs per contact are high when comparing contacts with conversions. Data enrichment is the answer to many of the problems described above. With the help of data enrichment, variables never observed can be matched to individual customer profiles based on link variables. They are derived from internal or external sources, e.g. the company website, a customer survey, market-media studies, social media applications, or other available information. That way, definite variables can be used for targeting, rather than demographic target group descriptions, derivations, or common knowledge.

Values are matched for all customers – with corresponding probabilities and measures of fitness – so that not only very active customers can be differentiated. Because data is matched by groups rather than on a personal level, also sensitive information, e.g. a personal income level, is addable. Even aggregated data can be used if link variables are available that are also present in the customer database.

**Research question**

In the dissertation, the following question shall be answered.

*Which data enrichment sources are able to increase conversion probabilities?*

Previous works by Putten et al. (2002), Hattum & Hoijtink (2008), and Krämer (2010) suggested that data enrichment is able to increase conversion rates, at least under certain circumstances. However, it has never been considered what is special about data enrichment in database marketing and under which circumstances data enrichment results significantly increase conversion probabilities. In particular, it has never been assessed comprehensively which sources are suitable for data enrichment in database marketing. This, however, is a question crucial to practitioners before starting a data enrichment project.
Literature Review

The first approaches to data enrichment (or data fusion/statistical matching) in order to receive new variables were undertaken by Okner (1972) and Wendt (1977). Data fusion was developed and matured mainly in the two areas of media planning and official statistics. It has been very well described by Rässler (2002) and D’Orazio et al. (2006). Notable developments in the area of data fusion have been described by Kamakura & Wedel (1997), Gilula et al. (2006), Putten et al. (2002), and Hattum & Hoitink (2008).

Data enrichment involves two (or more) sources of information, where the overlap between the groups is said to be negligible, meaning there is no need to find the one correct match in the other source, but one or more individuals that are alike. They can be identical (record linkage), but do not have to, when referring to the idea of conditional independence (Rässler, 2002; Rodgers, 1984; Radner, 1980; Adamek, 1994). If a correlation exists between the link variables and the target variables in the donor source, it is possible to form groups of persons being alike as measured by their link variable values (Gilula, McCulloch, & Rossi, 2006; Rodgers, 1984; Rässler, 2002, p. 11). Target variables for the customer group can be calculated from them. A schematical illustration of data enrichment is given in Illustration 1.

Illustration 1: Schematical illustration of the data enrichment idea (derived from Putten, Kok, & Gupta (2002, p. 2)

Data enrichment results are not preferable to single source data, but enrichments are a notable alternative whenever single source data is not available or unreasonably difficult to obtain (Kamakura & Wedel, 1997; D’Orazio, Di Zio, & Scanu, 2006, p. 1). For database marketing applications, data enrichment is relevant for information on customers differing from the information available. The information available usually comprises profile data, transactions data, or data from different customer touch points, but lacks information on attitudes, needs and wants, motivations, and purchase behavior.

The idea of data enrichment in marketing has become more and more popular in the last years. Putten, Kok, & Gupta (2002) were among the first to introduce the idea of data enrichment in the context of database marketing. Putten et al. (2002) proposed a generalized model for data enrichment in marketing. Their idea was further developed by Hattum & Hoitink (2008) who conducted data enrichment in order to be able to segment customers into groups, explicitly addressing the problem of micro validity.

Theory

Let there be a population , of which a recipient unit and a donor unit are two samples. Let the link variables be present in all three samples and let the target variables only be present in and . Illustration 2 shows the scope of possible sources with examples and where to find them in the three-dimensional space of identity , size , and representation .
Every source has a data generation mechanism. A missing data indicator variable denotes whether data is present or missing, it is reasonable to ignore the data generation mechanism, if for all \( \theta \). Then, beliefs about \( \theta \) are unchanged when taking into consideration (Bernardo & Smith, 1994, p. 45). The weakest conditions under which the data generation mechanism can be ignored are if the missing data are missing at random, if the observed data are observed at random, and if there are no a priori ties between the data and the data generation mechanism (Rubin, 1976).

Illustration 2: Different data enrichment sources by degree of identity, size, and representation

Let \( \theta \) be a random variable denoting unknown parameters. If missingness does not depend on any of the data, the data is missing completely at random (MCAR). If the data generation mechanism is MCAR, it is ignorable:

\[
\text{(1)}
\]

If \( \theta \), no data is missing, which is why it can be regarded as missing completely at random (MCAR). If \( \theta \), data is missing randomly due to a representative source, which is also considered MCAR.

If missingness depends on only, which has been observed for all observations, the data is missing at random (MAR). If the data generation mechanism is MAR, it is ignorable:

\[
\text{(2)}
\]

If or \( \theta \) or \( \theta \), data is missing at random (MAR). It is an axiom that the data generation mechanism of the recipient unit is always MAR. Otherwise, the enriched information can be biased, or the customer indicator variable needs to be observable in the donor unit. If or or \( \theta \), the missingness depends also on a third variable that has not been observed and data is missing not at random (MNAR). However, as stated by Rubin (1976), the data generation mechanism is ignorable if there is no association between \( \theta \) and \( \theta \). In this case, \( \theta \) and \( \theta \) are distinct and one does not provide any information on the other (Rässler, 2002, p. 77; Schafer, 1997, p. 11).
Furthermore, the values to be fused are derived from \( X \), and from \( X \) alone. Therefore, a source not satisfying all previous assumptions can be used, if, and only if, \( Y \) and \( m \) are conditionally independent given \( X \):

\[
P(Y, m | X) = P(Y | X) P(m | X) \leftrightarrow P(Y | X, m) = P(Y | X)
\]

It is therefore possible to define a restricted class \( Q \) of conditions under which the data generation mechanism can be ignored.

\[
P(Y | X, m) = P_Q(Y | X) \quad \forall Q = \left\{ \begin{array}{l}
\mathcal{D} = \mathcal{P}, \mathcal{P}' \\
\mathcal{D} = \mathcal{R}, \mathcal{R}' \\
P(Y | m) = P(Y) P(m) \\
P(Y, m) = P(Y | X) P(m | X)
\end{array} \right\}
\]

The conditions will be regarded throughout to be the minimum acceptable conditions under which it is reasonable to perform data enrichment without incorporating the data generation mechanism into the model.

**Assumption:** Sources can be used, if the appearance of people in the source is at least conditionally independent from the target variables given the link variables.

A hypothesis test regarding the total correct classification rate (TCCR) of a data enrichment application can show that the data enrichment results of sources under the conditional independence assumption are significantly better than random. If \( H_0 \) (\( TCCR_{model} \leq TCCR_{chance} \)) is rejected, the results of the model are significantly better than what could have been achieved by a random chance model, and the assumption is assumed to be true.

Data enrichment results are seldom 100% accurate and precise. Because of the categorical nature and the confined number of link variables, the degree of precision can only have a certain extent. The exact value of \( Y \) is always determined by more, not observable variables. Because of that, it is assumed that the model will be more accurate, if it is built on the same people, rather than on people that are alike as measure by their link variable values. Only then, other confounding factors can be controlled.

**Assumption:** The higher the degree of identity between customer database and source, the more accurate the overall data enrichment results.

In order to test the assumption, a model lift function is created by calculating a relationship between the model lift and the identity \( \mathcal{R} \cap \mathcal{D} \). If the relationship between the model lift and \( \mathcal{R} \cap \mathcal{D} \) is positive, the assumption is assumed to be true.

The size of data enrichment sources influences the enrichment results, leaving all other parameters equal. Following the argumentation given, the surplus of observations in the donor unit does not add to the total knowledge, but dilutes it.

**Assumption:** The smaller the size of the source for a given level of identity, the more accurate the overall data enrichment results.

In order to test the assumption, a model lift function is created by calculating a relationship between the model lift and the size for given identity \( \mathcal{R} \cap \mathcal{D} \). If the relationship between the model lift and the size is negative, the assumption is assumed to be true.

Oftentimes, enrichment sources are representative samples, e.g. of the customer group or of the overall population. For representative samples, it is assumed that \( \mathcal{D} \equiv \mathcal{D}' \). Their data generation mechanism is random, which is why all information is preserved in the smaller samples.
Assumption: A representative sample leads to the same quality of results as the bigger population from which it was sampled.

In order to test the assumption, the total correct classification of a population needs to be compared to the total correct classification rate of its representative sample.

Method
In order to test the suitability of different sources for data enrichment in database marketing, a simulation experiment is carried out. In contrast to real world applications, the data enrichment is done for a database already containing the target variables. The data enrichment situation is simulated so that the results can be compared to the true values. The simulation can be carried out for situations that have not yet been established in the real world (Albright et al., 2011, p. 919), like the sources under the conditional independence assumption. The modification of a simulation is much easier and cheaper than for real world systems. Various approaches can be adapted and compared. Because of its reduction and simplicity, the implications of the modifications are easily analyzable and interpretable (Dekker, 1993). During the simulation, more than 4000 data enrichments are carried out for various sources and six target variables, using three different data enrichment methods (mode imputation, nearest neighbor hot deck, logistic regression). The different source types are generated from the full dataset using respective data generation mechanisms. The aggregated results are used to evaluate the assumptions described above.

The data for the simulation study is a real-world sample from the customer database of a renowned German company whose name is omitted due to data protection reasons. The real-world origin guarantees realistic distributions of variables and correlations among link variables, target variables, and the varying data generation mechanisms. The simulation is carried out using SAS 9.2. SAS is certified and is most frequently used for database marketing purposes, which is why it was preferred over SPSS and R.

Expected Results
This dissertation is going to shed light on the yet unstudied field of donor unit characteristics. Common sources for data enrichment are internal data (e.g. behavioral data from company website), internal market research (e.g. a survey conducted to answer particular questions), external market research (e.g. market-media studies, branch surveys), and other external sources (e.g. social media data, other available sources). It seems natural that information from internal market research is applicable to all customers. Also market-media studies covering the whole population are thinkable as a source. But data from social media, for example, applies only to a subgroup of the customers and it seems questionable whether this source can be used. With the results of the simulation, the assumptions will be verified and deductions for further research and practice will be made.

Current state of research
I have hitherto planned and executed the simulation based on a thorough literature review and am currently writing the theoretical parts of my thesis and analyzing the simulation results. I have gotten feedback from two of the main researchers in the field of data enrichment, Prof. Dr. Susanne Rässler and Dr. Pascal van Hattum. At the AMC Doctoral Colloquium, I hope to have valuable discussions regarding the aspects to research during the simulation study, the applicability of the enrichment results in practice, and the overall model as such. I will incorporate my findings in the finalization of the simulation and the dissertation text. Also, I would be interested whether my study is suitable for a journal article.


