Chapter 8

Limitations of data augmentation and outlook

Data augmentation sources are manifold and oftentimes easily available. The main question arising from these sources is whether using them for augmentation purposes will lead to an increase in information and a better basis for decision making. The decision for or against data augmentation is simplified with our guidelines. However, they do not substitute a thorough case specific examination and pretesting phase.

In order to correctly use data augmentation in database marketing, it is necessary to consider what data augmentation is capable of and what it is not. In this chapter, our findings are enhanced with hints on how to manage expectations regarding data augmentation results. Data augmentation is a tool for providing information in an area of database marketing where existing customer data is sparse. It should always be regarded as such. We contemplate the limitations of data augmentations and the successional decision whether data augmentation is the right tool to answer questions in database marketing. To understand these limitations is an important factor in the appreciation of data augmentation results.

B. Hüttenrauch, *Targeting Using Augmented Data in Database Marketing*,
DOI 10.1007/978-3-658-14577-4_8

We were able to answer the research question and to provide comprehensive findings regarding external sources in database marketing. However, our study has also raised new questions. We have started to examine the source data mechanism. More research should be dedicated to this topic and how different source data mechanisms can be approached in practice. Certain steps in our proposed data augmentation process can be enhanced by deeper exploration. Our case study is a comprehensive example of data augmentation settings. However, some parameters have been fixed in order to establish comparability. These parameters vary in practice. More data augmentation use cases should be regarded in order to support our findings. Eventually, we point out the difference between predictive and uplift models, the latter focusing on how customers are motivated and activated through direct marketing communication.

8.1 Limitations and alternatives

When conducting a data augmentation, managing expectations is an important part of the project. Any analyst making secondary use of the data should have a clear picture on the meaning of the available data. The augmented values are directly saved in the customer database and become an analysis basis themselves. But using it like a primary data source is dangerous, because it dilutes the analyst in knowing something not actually known (Dempster & Rubin, 1983), and data might be used for conclusions not possible from augmented data. Data augmentation can enrich the existing data in a way not possible otherwise and its results can serve as important decision criteria. However, data augmentation results are only approximations, and decisions based on the results should be treated accordingly.

Essentially, data augmentation is only an option if no complete dataset is available, or if it is impractical to collect all data from a single customer (Adamek, 1994). Sometimes, there are better possibilities to reach a particular marketing goal. Much information can already be found by mining the donor or recipient unit only. It is not always the best solution to mix up sources in order to receive valuable insights (Putten et al., 2002b).

First and foremost, all additional information is valuable to database marketing analysts. In practice, targeting formulas are oftentimes not very sophisticated. Knowledge on the majority of customers is limited and predictions of conversion probabilities are associated with great uncertainty. The benefits of data augmentation oftentimes outweigh the risks, so that data augmentation has become a popular tool (Adamek, 1994). There is a trade-off between the economical contribution of the new information and the cost related to the augmentation project. Because marketing is usually afflicted with not knowing very much on most of the customers, any additional information is valuable and money is attributed to this cause. Whenever target group decisions can be facilitated and direct marketing campaigns become more efficient, a data augmentation project is worth the effort. This should be kept in mind when comparing the usefulness of data augmentation to the challenges and limitations to be stated below.

The knowledge achievable from data augmentation is limited. Because of the categorical nature of variables, data augmentation information is usually not very precise. One could even argue that it is misleading to say that data is really augmented. It combines the already existing link variables in a way that conversion probabilities for certain target variables can be predicted. Not the target values are the new information, but the model combining the link variables pointing to the target value with a certain probability. The information has thus already inherently been present in the data. This combination of variables is the value of the data augmentation results.

The augmented data is always only as good as the data in the source. Data describing human behavior and preferences is volatile and may age quickly. These facts should be considered before approaching a data augmentation project and should be evaluated carefully for any source. Obstacles related to the quality and usefulness of the data are:

- *Data quality:* The results depend directly on the data quality of both recipient and donor unit. Some values might be deficient or missing and some values might not convey the intention of the person from whom it was collected.

- *Availability of link variables in the recipient unit:* In order to perform data augmentation, some existing knowledge on the customers is necessary. In fact, the more information is already known, the better more information can be augmented. This is not necessarily in accordance with the marketing strategy for data augmentation, which often involves receiving a better picture of customers on whom not so much information is available already.

- *Comparability of link variables:* Every data source contains variables collected according to specified concepts and definitions, saved in a specific format and scale. Link variables can only be used, if concepts and definitions are similar, and if formats and scales can be adjusted to be the same.

- *Correlation between link and target variables:* If the link variables are not able to predict the target variables, data augmentation is not possible.

- *Meaning and interpretation of variables:* Some variables might not have the meaning that the database marketing analyst ascribes to it. The donor unit is often a source whose data was collected for a different purpose. There are hard facts like age or gender, which are universally understood. When talking about interests, the situation is not as clear. If a woman records to be interested in shoes by adding this interest in her Facebook profile, it might mean a lot of things to her. The database marketing analyst would interpret the interest in shoes as the intention to buy shoes. While there certainly is a correlation between being interested in shoes and buying them, the two notions have a slightly different meaning. ”Borrowing” someone else’s data for a different project may lead to bias because of differences in interpretation one is not aware of (Ozimek, 2010).

- *Usefulness of target variables:* Even if the meaning of a variable is captured correctly, the variable might still not be a direct hint for the conversion probability of an offer. Targeting is the application of a mixture of variables, of which every variable is expected to have predictive power regarding the conversion. However, this predictive power cannot be tested upfront. It can only be observed after having conducted a marketing campaign. If a person is interested in shoes, and also likes to buy shoes, it might still not mean that this intention is influenced by a campaign. The conversion probability concept is more complex and thus manipulating it is not trivial.

It cannot be expected from data augmentation to exactly reproduce every single value. It has been shown in the past that many statistics judging data augmentation results were within acceptable limits of the real values (Baker et al., 1989). Sometimes, it makes sense to illustrate the uncertainty associated with data augmentation by introducing uncertainty bounds (D’Orazio et al., 2006, p. 97ff). Additionally, data augmentation is only able to identify groups of customers that can be targeted by different marketing mix strategies. It is still far from a one-to-one marketing solution (Hattum & Hoijtink, 2008a).

The data augmentation approach as described here is a micro approach. The most likely target value is augmented to the customers. Only the target value’s probability for people with the according link variable class in the source is available. Data augmentation results are not useful for aggregated statements, because the best value is sought for every customer, which does not necessarily adhere to the overall macro validity. Statements like ”half of our customers are interested in this product” are not valid. It would always have to be put into the context of the source. The correct interpretation of the data might be difficult to explain to external parties, e.g. general management. Likewise, no correlations should be made between the augmented variables and other existing variables. The augmentation results are a tool for enabling better segmentations and target group selections. If aggregated statements shall be made, market research is a better alternative.

Data augmentations rely on some assumptions that cannot be tested in practice. For example, if the customer database is assumed to be MAR (as described in chapter 3.1.2) and no auxiliary source is available to confirm or object this assumption, the results rely on the correctness of this assumption. Several theoretical, empirical, and simulation studies have shown that there are risks associated with data augmentation (Adamek, 1994; Rodgers, 1984). As in every research model, there are several steps in the data augmentation process where decisions have to be made by the researcher. Every poor decision can compromise data augmentation results.

The data augmentation results should always be used in combination with existing data. The uncertainty inherent in the augmentation results is too strong for decisions to be based solely on these results. Data augmentation results are only meaningful, if the current information available for decision making is sparse. For example, when introducing new products or cross selling other product categories, data augmentation results can enhance the decision basis. For categories or products for which much is already known, data augmentation results derived from an external source might not have an additional informative value. But whenever other, more substantiated information is available, data augmentation results can be used to improve decisions and to have more variables to base a decision on. They can also be used to validate preliminary decisions.

8.2 Further research opportunities

This study is a starting point in the exploration of data augmentation with external sources in database marketing. More questions arise from our findings. These are delineated below. Furthermore, other use cases relevant to the practice are given, broadening the field of data augmentation research in database marketing.

8.2.1 Ignorability of the source data mechanism

MCAR source data mechanisms are easily differentiable, while MAR and MNAR source data mechanisms are not observable. Whether a source is MAR or MNAR depends on the conditional association of source and target variable given the link variables.

It has previously been suspected that a clear distinction between conditionally dependent and independent sources is difficult, because of the categorical nature of the variables and the complexity of the association resulting from a big number of link variables and the various classes related to them. In a conditional independence test, partial two-way cross-sectional tables are built and independence is tested for source and target variable for every class. A major problem encountered has been the problem of sparse data in many of these classes. Another problem has been the frequent disagreement of the χ^2 test with aggregated total measures and the CMH test used. Because a MNAR source data mechanism did not always compromise the augmentation results, we eventually concluded that conditional independence can be assumed. It should be validated whether the assumption of conditional independence, as suggested in this study, is admissible when no auxiliary data is available. It was sufficient in our research context, but could prove differently when validated in a different context.

The process of assessing the ignorability of the source data mechanism is complex and hardly performable in practice. Even if we had been able to definitely prove or disprove conditional independence, the same test cannot be performed in practical applications, where the source data mechanisms, as well as the elements not observed are not known. More research is necessary in this field.

8.2.2 Deeper exploration of the augmentation process

In order to get a deeper understanding of data augmentation in marketing, further research is suggested at three points of the data augmentation process. The first enhancement concerns other and more complex data augmentation methods. The second enhancement is related to the uncertainty assessment of the target values, once a data augmentation has been carried out. Eventually, a study researching on the external evaluation of the augmentation results would be valuable in order to complement our study.

The methods presented here do not reflect the state of the art in statistical matching. They are hands-on ad hoc approaches to data augmentation, as it is frequently found in companies. This choice is not only for simplification purposes, but also because the practical application of these methods is more likely in database marketing than other, more complex methods. Nevertheless, the use of more complex methods is possible and their benefits regarding effectiveness and efficiency should be evaluated in order to get a more comprehensive picture of possible data augmentation methods. Examples are likelihood-based inference methods, as for example the EM algorithm, or Markov Chain Monte Carlo methods, as for example Gibbs sampling or the Metropolis-Hastings algorithm (Schafer, 1997, p. 2). These methods involve complex simulations of posterior distributions (Schafer, 1997, p. 4). It has not been feasible for us in our case study context to apply methods which would require a simulation each. Methods of these fields could also be useful in gaining more insight into the conditional association assessment. To contrast the effectiveness of such methods against our approaches would be a valuable extension to our work.

The probabilities augmented in our study are the direct observable probabilities as derived from the source. Its interpretation always takes into consideration the source derivation. These probabilities do not include any kind of uncertainty assessment related to the fact that MAR sources are not identical to or representative of the customer population. If possible, it would be desirable to get a more comprehensive measure, including an uncertainty part as derived from the source characteristics. It might prove difficult, because probabilities reflecting all uncertainty might easily decrease to very low numbers not catching differences between target values anymore. Further research is necessary in this area before such a comprehensive measure can be developed.

Eventually, an external validation of this study in terms of actual conversions and return on marketing investment is desirable to complement our conceptual model. Many factors influence the ROMI and it is not easy to conduct a comprehensive study including various sources, like it has been done here. To establish comparability when real sources are used is rather difficult. The external validation is much more focused on practical applications, and several case studies can be combined in order to perform an overall external validation of our conceptual model.

8.2.3 Further augmentation opportunities and use cases

In this study, we have collected, specified, and structured the features of data augmentation in database marketing for the special case of external data. The data augmentation approach as proposed in this study is built on the use case of optimal target group selection in direct marketing. Although being highly relevant for database marketing practice, it has never been comprehensively assessed in a scientific context. In order to give guidance to database marketing analysts, basic rules, relevant aspects, and cruxes of data augmentations are stated. These are verified with a suitable case study. More examples are needed in order to build a comprehensive picture for marketing in general. Other use cases are thinkable and many of them have different properties in terms of variable scales, inference requirements, and recency expectations. The augmentation opportunities comprise, but are not limited to, the following use cases.

Our conceptual model and all our methods are based on the assumption that the typical sources for data augmentation applications contain categorical variable scales. This might be different, if branches like the financial sector were regarded, where many variables are metric. In this case, it is possible to reduce these variables to categories. However, this would result in a loss of information in terms of accuracy. In these cases, it might be reasonable to develop a conceptual model for metric variables in order to receive more meaningful data augmentation results. Also, many sources contain a mixture of variable scales. In our approach, we propose to harmonize all variables to the same scale. A comparison of such methods to the proposed standard methods would be a valuable extension to our work.

In our data augmentation proposition, we chose a univariate pattern approach. It means that all target variables are augmented individually, so that the most accurate results are achievable for every variable. However, by doing so, no inference can be made between different target variables augmented from the same source. It means that there have to be made as many augmentations as there are target variables. This can prove time consuming, because every model has to be established and adjusted to best fit the respective target variable. Hence, marketing problems exist where the advantages of a multivariate pattern approach outbalance the accuracy advantages of the univariate pattern approach. Such multivariate pattern approaches are more complex than univariate pattern approaches, because different target variables have to be explained at the same time. Only methods using statistical twins, such as the nearest neighbor method, can guarantee that all target values are taken from the same donor. Interrelations existing between these target variables are preserved and it is possible to attempt to make inferences between these variables after having augmented the data. Establishing such approaches and comparing them to univariate pattern approaches would be a meaningful enhancement to our work.

Our approach is based on a classical database marketing structure, where unique customers are the elements the analyses are based on. While this is true for most channels in the offline world and in email marketing, other channels have different identifiers, such as cookies, online accounts being used by multiple persons, or other structures not directly relatable to unique persons. If it is not possible to convert these structures to a unique-person-element structure, classical data augmentation sources cannot be properly used. At the same time, there are possible sources in the internet world consisting of these non-classical structures. Likewise, these sources cannot be used for data augmentation, if no transformation is possible. Both the identification of unique persons in these structures and the usage of other units for data augmentation purposes (both as recipient units and donor units) are interesting and seminal fields which can be regarded.

Eventually, all data augmentation structures should be transferred to automatic processes in which new information is generated in real-time. One time approaches, especially as derived from dedicated surveys, have little information value, because the information is not valid for a long time period. Most marketing problems as described here, regarding target group selections in direct marketing, return frequently. An automated augmentation process has several advantages. Firstly, both the augmented information and the link variables information is augmented on are always up-to-date. Secondly, the ROMI increases with every reuse of a source that has once been connected to the customer database with data augmentation. Of course, the automation process has afflicted costs itself that need to be evaluated in the decision process. But the maintenance effort is low when compared to a whole new data augmentation. Finally, the data augmentation automation infrastructure can be used for new sources, again leading to economies of scale.

8.2.4 Uplift models

Our models assume that the fact that someone has a certain characteristic, i.e. a certain target value, is a good predictor for that customer to convert after having received a marketing communication. Uplift models reach one step further. They try to predict for which customers the probability of conversion is maximally *increased* when being contacted by a marketing campaign. It does not necessarily prefer those customers with a high affinity to an offer, if these customers would have bought an offered product anyways. Rather, it tries to identify those customers whose propensity significantly increases through a marketing campaign.

Marketing campaigns using uplift models might not reach as high conversion rates as those based on our approach of data augmentation. Nevertheless, the delta between a potential control group and a target group chosen by an uplift model is higher. When adding up general sales not motivated by marketing actions, the sales leads generated by the marketing action, and the costs of the marketing campaign, uplift models yield higher revenues than simple conversion probability models.

Uplift models are more complex than simple conversion models. The general conversion probability needs to be separated from the uplift in conversion probability induced by a marketing campaign. It needs to respect many volatile factors. However, if mastered, uplift models have a high profit, because the marketing budgets can be directed at exactly the customers that respond best and the overall earnings can be maximized.

Data augmentation results can also be used to specify uplift models. This is a logical extension to our work. Uplift models need to be trained well in order to make the separation between the general conversion probability and the uplift in conversion probability induced by a marketing campaign. Such information cannot be found in an external source. This makes it difficult to *augment* the probability for a significant uplift. But the augmented information can be used to build an uplift model *within* the customer database.