Linking Your Database With Your Segmentation

Michael Conklin – 56Stats LLC

Abstract

In this document I outline the major considerations needed to successfully link a market segmentation with an internal customer database. I will illustrate how the specific use cases that are anticipated for the the resulting linkage impact the feasibility and desirability of performing the linkage. I also cover the major statistical methods for creating the linkage and their advantages and disadvantages for specific use cases.

Introduction

Market Segmentation is an important tool for marketers in many different industries. The purpose of segmentation is to identify subgroups of buyers in a market that will respond to different marketing efforts in different ways. John Wanamaker famously said "I know that half of my advertising dollars are wasted…I just don't know which half". Market segmentations are an attempt to make marketing efforts more efficient by identifying groups within the market who will be more likely to respond to specific marketing efforts and targeting those efforts to those groups.

Segmentations of the market can be very simple. If you have a product that is intended to be used only by women, then you can segment the market by gender and concentrate your marketing efforts with messages and media that have a high proportion of women. Most segmentations, however, are much more complicated than just targeting a single gender. If a marketer can understand a segment's attitudes and desires it is much easier to craft a message that will generate a strong response from those potential customers. In today's media environment of micro-targeting it is even possible to specifically target your messages to the people most likely to respond.

Segmentations are usually done on the whole market. That way a marketer can understand whom to target with specific products or messages. Internal databases of customers are, by definition, a subset of the market. They consist of the part of the market who are already

your customers. So what is the benefit of being able to classify your current customers into the market segments you have identified in the overall market? There are actually several benefits.

- Understand how your customers differ from the market as a whole by seeing how the relative size of the segments in the market differ from the sizes among your customers.
- Evaluate how new segment targeting efforts are bringing in new customers from the targeted segments.
- Identify target segments that are under-represented in your customer base and generate strategies to address those segments.
- Understand the relative profitability of different market segments to be able to prioritize new customer acquisition programs.
- Identify segments that have a history of repeat purchases to maintain long term growth.

All of the above considerations become especially important if your market share is small enough that you have very few customers in your segmentation study. Assuming that you have, or are considering doing, a segmentation study to segment your total market into actionable segments– how do we go about making sure that we can map those segments on to your existing customer database.

Linking a CRM Database and a Segmentation

The biggest impediment to linking a CRM database and a segmentation study is that the two data sets were created for entirely different purposes. The segmentation is designed to be a representation of the whole market with a detailed understanding of behaviors, attitudes and demographics that are believed to be important in determining who will be interested in purchasing the product category of interest. The CRM database is designed to facilitate Customer Relationship Management. The people represented in the database are either customers or likely prospects. When a customer or prospect is entered into the database it is critically important to know how they would like to be contacted, how they heard about the product or service and how interested they are in the product/service being offered. It tends to be counter productive to collect a large amount of demographic, attitudes and behavioral data (similar to data collected in a segmentation survey) since collecting that information is costly (in time and money) and does not have a direct immediate impact on the customer relationship.

But, with little or no data in common between a segmentation study and a CRM database it is very difficult to build a model that identifies segment membership accurately in the CRM database. This can be remedied if you haven't yet done your segmentation study. In this case, you can add questions to your segmentation study that are equivalent to fields that you capture in your CRM data. You can also include some of your CRM customers in your survey. This

latter strategy should not be done instead of including CRM questions in your segmentation study since this will limit the CRM info only to the subset coming from your CRM database. The remaining records will have missing data in those fields.

Building your segmentation

When building your segmentation you will have to select a set of base variables that will be used to create segments. Segments are typically created by finding groups of respondents that are similar to each other on these base variables and as different as possible from people in other segments. The important concept here is that the base variables define the segmentation. Sometimes, for this reason, users want to limit the base variables to the items in the segmentation survey that they care about the most. However, if the goal is to eventually link the segmentation to a database then you should always include as many variables as possible that are present in the CRM database. The concern is that these variables change the segmentation. My experience is that the segmentation changes very little but the gain in the ability to link to the CRM database is large.

A second strategy that is sometimes used is to conduct a second survey with a sample of respondents from the CRM database and assign them to segments using a typing tool that was created from the original segmentation. These typing tools are developed by finding some minimal number of questions from the original segmentation that can reasonably predict segment membership. The typing tool questionnaire is much shorter than the original segmentation questionnaire and the resulting data provides segment membership assignments and the complete CRM data for those respondents. One should be very careful with this approach. Typing tools are rarely validated. While an error estimate can be obtained during typing tool development (all models have some error), there is a more insidious problem. The subset of questions used in the typing tool are asked in a different context than the original segmentation study. The fact is that respondents answer these questions differently when they are the only questions being asked instead of being embedded in a much longer questionnaire. This is especially problematic if the questions represent attitudes. Therefore we recommend using the original segmentation data to build the CRM linkage model.

Choosing an Algorithm for the Linkage Model

There are several considerations in choosing an algorithm for a linkage model.

- Simplicity of implementation: Simpler algorithms such as linear discriminant models can easily be directly programmed in tools like Excel or with SQL statements. This makes implementation in the database simple. The disadvantage is that simple models may sometimes have much larger error than more modern machine learning algorithms.

- Control of overfitting: Overfitting can happen with any algorithm. It is imperative that when creating the model it is tested on new sets of data. The basic procedure is to create the model on some subset of the segmentation data and then predict the remaining data to evaluate the accuracy. If accuracy is assessed on the same data used to fit the model then you will get an inflated estimate of model accuracy. There are many methods for controlling for overfitting including cross-validation, leave one out validation, and bootstrap aggregation.
- Feature creation: The features or predictors in a model can be much more than the actual fields present in the CRM database. Combinations of those fields can be quite useful in prediction. In many machine learning algorithms multiple models with multiple sets of features are used to predict the outcome. These models include recursive partitioning models (tree based models), random forests (many hundreds of tree models - each fit on different subsets of data), gradient boosting models (a sequence of models is built where each successive model is built only on the data that the previous model got wrong). All of these models must control for overfitting because they add many parameters to the model which directly leads to overfitting.

All of the above mentioned techniques can be useful. In addition, there is really no reason to limit your models to simple models that can be programmed in Excel or SQL. It is quite simple to set up a small internal server to provide an API to these models that the database can query to get the desired result.

Measuring Error in the Linkage Model

One of my favorite quotes is from George E. P. Box, a statistician famous for the development of the Box-Jenkins Time Series model. "All models are wrong. Some models are useful". In this case, every linkage model will have some error. The question is how much error and does that error affect the usefulness of the model.

Estimating the error can be done using techniques similar to cross-validation. Build the model on a subset of your segmentation data, using only the CRM fields that you have in the segmentation, and measure the error in predicting the remainder of your segmentation data. Do not make the mistake of setting an arbitrary cutoff such as "my model must be 80% accurate". Usefulness also needs to be taken into account and the usefulness of a model depends on the specific use case.

In fact, it may be that multiple models are needed for different use cases. For example, if the desire is to target one specific segment identified in the segmentation study then a model built to predict that single segment would be more useful than a model that maximizes accuracy across all of the segments. On the other hand, if one wishes to track over time how the mix of segments in your customer or prospect base changes then a model that is more accurate across all of the segments is more useful.

In addition, a low accuracy model might still be very useful. A model that improves response from 1% to 2% to some marketing effort can really pay off. This is especially true with today's ability to develop target audiences for digital advertising.

Checklist for building a linkage model

- Identify fields in the CRM database that are, or can be, added to or included in the Segmentation study.
- Develop and understand your use cases. How will the segment identifiers in the CRM database be used?
- Determine whether the tagging of the database needs to be a one time or annual exercise or does it need to be real time or on demand.
- Test a variety of algorithms to balance the accuracy and implementation details against the requirements of the use cases.
- Choose your final model(s) based on the goals of your use cases.
- Implement the model(s) either with a SQL/Excel Script or developing a more sophisticated server/API solution.