Customer Segmentation

Individual Customer Approach Despite Optimal Use of Resources

Dr. Steffen Wagner

Your customers are individuals. They have diverse needs, financial backgrounds, and shopping behaviors. What conclusion should your company draw from this fact? It's clear that your customers can't be treated as one homogenous group. Of course, individualization is easier said than done, since communicating effectively with each customer is time- and cost-intensive. One doesn't need to have studied Business Administration to know that such costs and benefits are difficult to reconcile.

The solution is a compromise: customer segmentation.

Using statistical methods, your customer base can be divided into manageable, homogenous groups. A segment-specific approach allows for the bundling of resources, while at the same time achieving an individualized customer experience.

Each segment contains customers that are similar to each other on the basis of their needs and characteristics. For example, segments could include "Bargain Hunters", "Fashionistas", or "Technology Freaks". Advertising and CRM measures can be tailored to each segment, resulting in an optimal distribution of resources. Put simply, it's a "win-win".

This white paper provides a comprehensive overview of customer segmentation. The relevant prerequisites, procedure, and possible applications of such a project are discussed, and the essential points are highlighted.

The Objectives of Data-driven Customer Segmentation

Customer segmentation allows for the identification of data-driven homogenous groups that are both similar in terms of their characteristics and buying behavior within the group, and as distinct as possible from other groups. Figure 1 summarizes the goals of data-driven customer segmentation.

Insights about your Customer Base

In the course of doing business, it is likely that your company has hypothesized about your *typical* customer, or types of customers. Data-driven customer segmentation discovers which of these hypotheses can be confirmed by the data, and how large their shares of your total customer base are. We've often found that analysis results differ in interesting ways from one's expectation.

Individualized Communication

We know that every customer is different, and should therefore be treated differently. Customer segmentation allows for different strategies depending on the customer group; for example, the development of a segment-specific newsletter program. Specific and relevant content is more successful at avoiding annoying
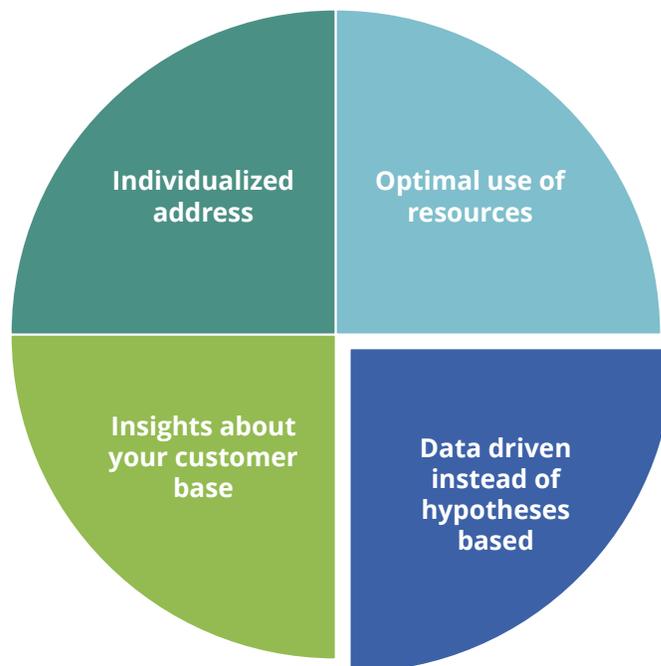

Figure 1: *Reasons for data-driven customer segmentation*

Figure 2: *Typical structure of order data: For every customer there is one or more orders consisting of one or more items. To use such data, aggregation at the customer level is necessary.*

the recipient, as they feel that the advertising company is accurately perceiving their interests and characteristics. Additionally, content can be curated on the basis of other group features, such as by timing newsletters in line with when a particular segment may be more inclined to make a purchase. It's even possible to avoid advertising to unprofitable groups altogether once they have been identified.

Data-driven, rather than Conjecture

In principle, customer segmentation can also be achieved by using previously-defined heuristics. For example, groups could be constructed on the basis of age, gender, or ordering behavior. However, this approach would only be effective if such arbitrary groupings were an adequate portrayal of true customer segments. A likely scenario is that these heuristically-defined groups are not truly selective and homogenous within themselves. If segmentation is data-driven, on the other hand, identified groups can be assured to match the actual customer structure.

Optimal Distribution of Resources

Last but not least, the insights from customer segmentation allow for the more efficient use of resources. For example, newsletters may be developed specifically for large segments, or newsletters may be restricted to only those groups where such advertising methods are effective. As another example, online advertising can be more targeted: as soon as the link between the customer account and a current session on another website is established via a cookie, the customer can be served the most effective advertisement medium for their segment. This both allows you to cut costs by only advertising where it is effective, and be more persuasive with the advertisements that are seen by your customers.

Database for Customer Segmentation

For customer segmentation you should take into account any data that may help to comprehensively characterize your customers. This could include demographic data, order and tracking history, or responses to customer satisfaction surveys, for example. To begin a project, order history data (such as the number of orders, products, and order totals) is typically sufficient. This information is customarily easily accessible via a CRM system. Customer segmentation can also be carried out purely on the basis of tracking data. The combination of both CRM data and tracking data is particular helpful for businesses with lower order frequencies, as it can help to obtain higher-frequency information, and also potentially uncover new customers.
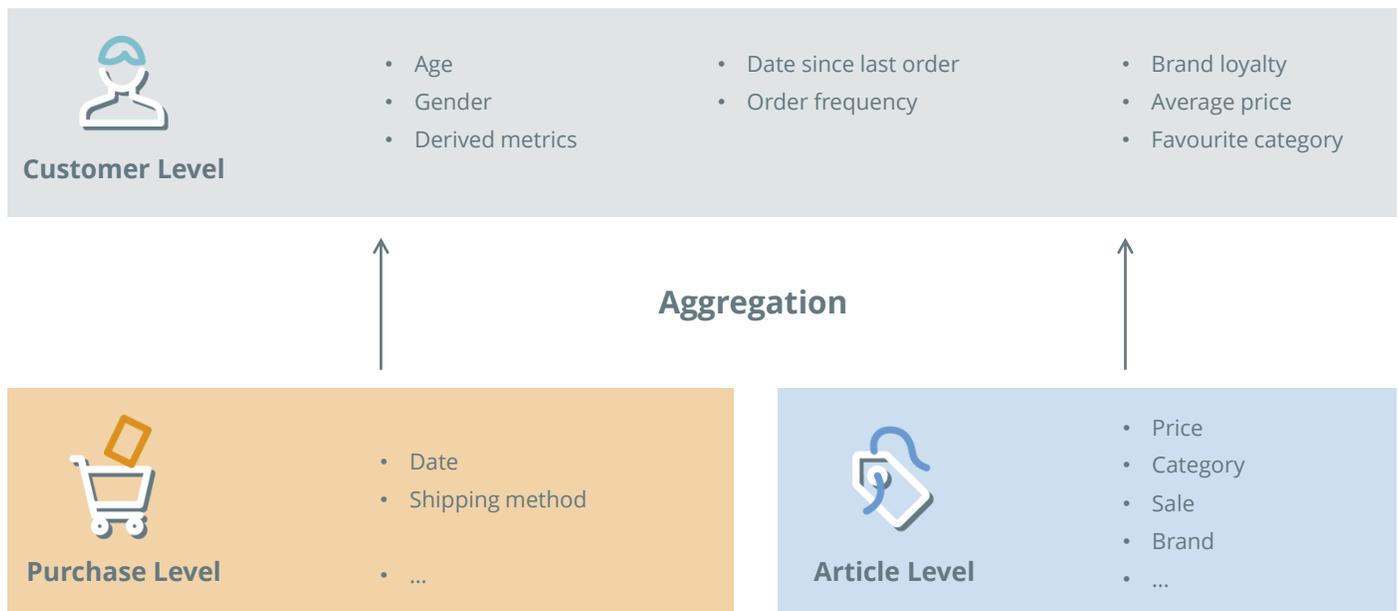
Customer Level	• Age • Gender • Derived metrics	• Date since last order • Order frequency	• Brand loyalty • Average price • Favourite category

Aggregation

Purchase Level	• Date • Shipping method • ...	Article Level	• Price • Category • Sale • Brand • ...

Figure 3: *Customer segmentation is preceded by the data aggregation step.*

Relevant Metrics

Any information that is available and necessary for the differentiation of your customers should be considered in the analysis. In order to ensure a comprehensive perspective, a kick-off meeting with experts from every relevant department is recommended. Inter-departmental exchange is necessary to ensure that no relevant information is overlooked. It's also helpful to discuss which business cases the customer segments will be relevant for at an early stage in the project, since this can be vital for interpreting the identified segments.

Metrics: Example

Some customer metrics are consistent for most B2C firms. For example:

- Age
- Gender
- Order frequency
- Shopping cart totals
- Products
- Return rates
- Use of vouchers/discounts
- Payment method

Selecting the Study Period

In order to ensure adequate data availability, the analysis period should be sufficiently long. However, it is not absolutely necessary to consider the entire customer history. The length of the required data history depends primarily on the typical customer order frequency. For example, travel bookings may require a longer order history (1 – 2 years), while grocery orders might need only a few weeks or months. It is also important to consider that findings may be irrelevant or distorted with data that is outdated, if the business model or composition of the customer base has changed considerably during that time.

Given that seasonality can have a sizable impact on order type and quantity, it is advisable to consider seasonal factors when choosing the time period (for example, holiday shopping). To avoid the influences of business fluctuations, the best course of action is to consider at least one complete "cycle" in the data.

Data Aggregation

To create customer segments, data must be customer-level. Some data are already available in this format, such as age, gender, or country. Other data may be available only at the order or article level.

Figure 2 illustrates the typical hierarchical structure of purchase order data. For aggregation, experience and coordination are required to develop a meaningful logic to map purchase order data to the customer level.

Figure 3 shows an example of how such purchase order data can be summarized at the customer level.
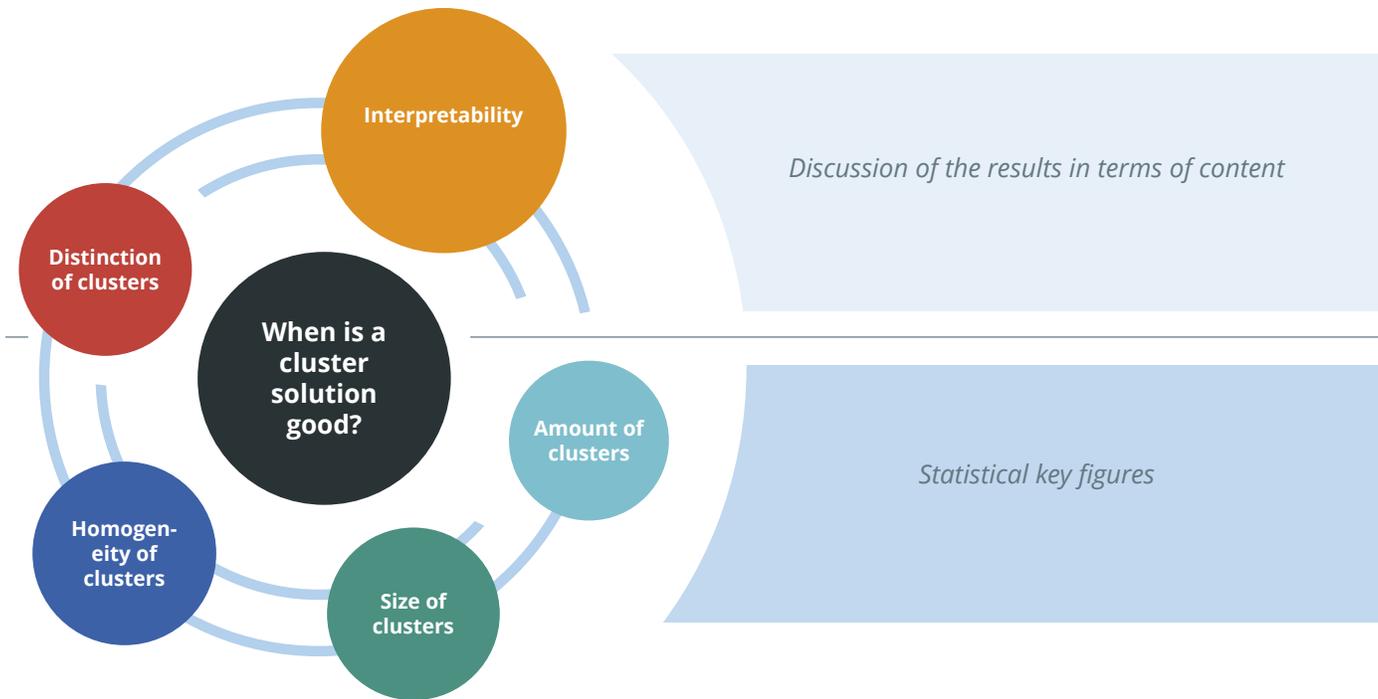
Figure 4: *Criteria for evaluating the suitability of a cluster analysis for customer segmentation. The most important factor is the interpretability, because this is key for the productive use of the customer segments in a company.*

Statistical Approach

"Cluster analysis" is *the* data-based method for identifying customer segments. This is an approach to uncovering the unknown, underlying connections between data points: in this case, your customers. There are many conceptually different methods for cluster analysis within the scope of statistical analysis, and the optimal solution is identified downstream.

Cluster Analysis Methods

There are several different statistical methods for performing a cluster analysis. They differ "under the hood" in the algorithm used to classify individual customers into groups. In general, the algorithms can be divided into three large groups: *hierarchical*, *partitioning* and *modell-based* approaches.

Hierarchical Methods

Hierarchical methods are based on the following two steps:

1. Determining the similarity measure between the individual customers.

2. From there, determining the ideal classification of customers into clusters/segments, so that the cus-

tomers within a cluster are similar, and each cluster is as different from the other clusters as possible.

There are two approaches for hierarchical cluster methods: agglomerative and divisive. Agglomerative methods start at the level of individual customers, combining the most similar ones step-by-step to create ever larger groups. Divisive procedures operate in reverse, beginning with the total quantity of customers and gradually subdividing into smaller and smaller subgroups. This consolidation or division is always based on the similarity of the customers or groups to each other. In both scenarios the hierarchical relation of the groups is evident, with each cluster consisting of two different subgroups. (see Fig. 5). This means that several solutions, each with a different number of segments, can easily be determined in parallel depending on the business task at hand.

Partitioning Methods

With partitioning cluster methods, customers are assigned to the center of the respective segment (called the medoid) to which they are closest. Since the segments - and thus their medoids - are initially unknown, a multi-step calculation with a randomly-selected initial center point is used to ensure that the iterative algorithm calculates the optimal solution, independent of

Customer ID

Figure 5: *A schematic representation of hierarchical clustering. At each step, one cluster is divided into two (divisive) or two are combined into one (agglomerative). The fewer clusters, the more heterogeneity within each segment.*

the random initial values. The medoids of the final segments represent the typical representative of that segment.

Model-based Approaches

A variety of algorithms are available for the implementation of model-based segmentations. The commonality between the approaches is the idea that customers belong to different groups, but that these groups cannot be observed directly. A customer's affiliation with an unobservable, latent group is the cause of the customer's behavior, and manifests itself in the available customer data. During the modeling process, the potential individual segments are explicitly defined, and then checked to determine which set of definitions can best model or reproduce the customer data.

Which Method is Correct?

Given the variety of cluster analysis methods, a key question is which is the correct one to use. The answer to this question is both simple and complicated. The simple, though general, answer is that the most suitable procedure is the one that manages to determine the *optimal* clusters from the available data. This broad answer is based on the fact that, as previously mentioned, cluster analysis is a structure-discovering, not structure-proving process. Therefore, the best method cannot be judged on a clearly defined metric, such as forecast quality. The evaluation is, rather, based on the following question:

When is Cluster Analysis Optimal?

The answer to this question serves not only to determine the cluster method to be used, but also the :*a priori* number of unknown segments.
The criteria used for the assessments are shown in Figure 4.

Key Statistical Figures

Criteria such as homogeneity within a segment or the number and size of the segments can easily be calculated. These criteria are, however, not independent from each other, but rather have inverse tendencies. Therefore, in order to determine the *statistical* optimum, indicators are used which, depending on the number of clusters, take into account the compactness and distance between segments. If these dependencies on the number of segments and different methods are considered, many potential solutions are identified from a statistical perspective.

Interpreting the Results

The most important characteristic of a good cluster analysis is its interpretability. If one can examine a cluster and create a clear picture of the corresponding customer group or group representative, this is the best pre-requisite for tailoring business practices to the cluster. If, on the other hand, the resulting clusters are diffuse or ambiguous, other solutions may provide better results.

Another discussion is the manageability of the size and number of segments. A solution with many segments may be easy to interpret, but may not be suitable for application it exceeds the capacity for an individualized approach. In contrast, if there are too few segments, individualization may be unprofitable.

Consideration of all of the above aspects is necessary to adequately judge a cluster analysis. It is important to note that this process is prepared by analyzing a comprehensive report on customer segment characteristics, but the actual discussion should be conducted together with all stakeholders to identify the optimal solution for your company.

Results of a Customer Segmentation

After completing the analysis and discussion of the resulting segments, you will have valuable information about the personas of your customers and their composition within your complete customer base. You can use these data-driven insights to understand your customers more accurately than with a purely hypothetical approach. Important learning can also be achieved by comparing the true and hypothesized groups.

In addition, each customer that was accounted for in the analysis will have a direct classification based on their corresponding segment. This allows you to immediately start implementing segment-specific actions and quantifying their benefits, e.g.~in the context of an A/B test. In addition, the analysis provides an algorithm that allows you to assign new or additional customers and customers with changed usage behavior to the identified segments. We recommend that you implement this algorithm in your BI system to ensure that the segment assignments are automatically available. If not all of the metrics used for the analysis are available for live operation (for example, if they originate from a survey), there is still the possibility for you to extend the results to new customers. This can be done through an additional classification algorithm that transfers the learnings from the segmentation project to a new, narrower database.

Conclusion

In close cooperation with you, we create a tailor-made solution that is data-driven and identifies your true existing customer segments. The combination of your domain knowledge and our data science expertise ensures that the identified segments become directly applicable for your business. You can immediately use the results of the analysis to optimize your business processes for specific segments, and an automated classification of new customers based on segmentation is also possible.

Your Contact Person

Dr. Steffen Wagner

Steffen is co-founder of INWT. He specializes in predictive analytics, online marketing, and customer relationship management. He holds a Ph.D. in physics and gives insights into his data science work as a lecturer in the joint master's program in statistics offered by a consortium of Berlin universities.

Contact

- **Tel.:** +49 (0)30 1208231-0
- **E-Mail:** steffen.wagner@inwt-statistics.de

INWT Statistics GmbH
Hauptstraße 8
Meisenbach Höfe, Aufgang 3a
10827 Berlin

+49 30 1208231-0

info@inwt-statistics.de

www.inwt-statistics.de