Market Segmentation Analysis

Contributor: Shreyash Chacharkar

Abstract

The process of market segmentation analysis involves multiple steps to gain an understanding of distinct market segments and tailor marketing strategies accordingly. The main purpose of this analysis is to get insightful information about market via structured methods. It involves 12 steps Deciding (not) to segment, Specifying the ideal target segment, Collecting data, exploring data, extracting segments, profiling segments, describing segments, selecting target segment, customizing the marketing mix, and evaluating and monitoring.

.

Step 1: Deciding (not) to Segment:

- Implementing market segmentation is a major decision that necessitates a long-term commitment due to its potential to reshape organizational operations. While market segmentation costs for research, design, and advertising promise profitability and customer engagement, the projected sales growth justifies these expenses...

- Many implementation barriers exist when deciding on segmentation, such as a lack of proactive leadership, championing, commitment, and involvement in market segmentation projects, organizational culture, a lack of training and expertise, and a lack of financial resources and planning. Strategies like increasing senior management involvement, reducing resource constraints, and cultivating a segmentation-friendly culture are required to address these challenges.

Step 2: Specifying the Ideal Target Segment

1. Knock-Out Criteria:

- A market segment's eligibility for further evaluation based on its attractiveness is determined by the knock-out criteria, which include homogeneity, distinctiveness, size, alignment with organizational strengths, identifiability, and reachability.

- These requirements serve as fundamental prerequisites and senior management, the segmentation team, and the advisory committee must all be aware of them. Shorter is preferable.

2. Attractiveness Criteria:

- Unlike knock-out criteria, attractiveness criteria are not binary; each market segment is rated on various criteria to gauge its attractiveness. The segmentation process's choice of target segments is guided by the overall evaluation of these criteria.

3. Implementing a Structured Process:

- A structured approach involves utilizing a segment evaluation plot that assesses segment attractiveness against organizational competitiveness.

Step 3: Collecting Data

1. **Segmentation Variable:**

    ● Empirical data for segmentation studies can be obtained from surveys, observations, scanner data, and experiments. It serves as the foundation for descriptor and segmentation variables, as well as intuitive and data-driven market segmentation.

    ● Commonsense segmentation divides consumers into groups based on a single characteristic (such as gender), with descriptor variables (such as age, vacation duration) providing specific segment descriptions.

    ● In order to identify organic or man-made market segments for customized strategies, data-driven segmentation uses a variety of variables.

2. **Segmentation Criteria:**

    ● **Geographic Segmentation**: Focuses on consumers' residence location, useful for targeting tourists from neighbouring countries' tourists. This may lead to bias.

    ● **Socio-Demographic Segmentation**: Uses criteria like age, gender, income, and education. Suitable for certain industries but may not fully explain preferences.

    ● **Psychographic Segmentation**: Groups based on psychological criteria, like beliefs and interests. Benefits and lifestyle segmentation are popular. Complex but useful for understanding behaviour.

    ● **Behavioural Segmentation**: Groups by relevant similarities, often using purchase behavior. May lack data on potential customers who haven't purchased before.

3. **Data from Survey Studies:**

    ● *Choice of Variables*: Ensuring the proper selection of variables is essential; data-driven segmentation may affect due to the presence of noisy or masking variables, this can introduce complexity to the segmentation process. Therefore, it is of utmost importance to pose distinct and essential questions, while also sidestepping redundant inquiries.

    ● **Response Options:** Response options like binary, nominal, and metric data types prove effective, the complexity arises with ordinal data. Opting for binary or metric options ensures both simplicity and accuracy in the segmentation process.

    ● **Response Styles**: Response styles wield influence over outcomes, with extreme answers and agreement bias being notable examples. Algorithms struggle to discern authentic beliefs from these response styles. To counter this, one can mitigate bias by conducting additional analyses or by excluding responses affected by such styles.

- **Sample Size**: Adequate sample size is essential to achieve accurate segmentation. In general, larger samples enhance the precision of segment identification. The required sample size varies according to market characteristics and data attributes. While larger samples can alleviate specific challenges, they might not address all issues comprehensively.

4. **Data from Internal Sources:**
   - Internal data like scanner, booking, and online purchases are used for segmentation. Watch for bias towards existing customers, possibly not representative of new ones.

5. **Data from Experimental Studies:**
   - Experimental data, especially from choice experiments, informs segmentation. Consumer preferences for attributes aid segment identification. Attribute impact on choice is a useful segmentation criterion.

Step 5: Extracting Segments

3. Model-Based Methods in Market Segmentation:

3.1. Finite Mixture Models:

- Model-based clustering involves analyzing data without taking into account independent variables (x) and fitting statistical distributions to it.
- The goal is to segment consumers based solely on segmentation variables (y),.
- A sum of weighted distribution functions, each corresponding to a segment, is used to represent the finite mixture model.
- Due to its capacity to model correlations between variables, a mixture of multivariate normal distributions is a popular choice for metric data.

3.1.1. Normal Distributions:

- Covariance between variables can be modeled using a multivariate normal distribution.
- The variances and covariances between two segmentation variable pairs are contained in the covariance matrix. Segment-specific parameters are the mean vector and covariance matrix.
- The number of parameters increases along with the segmentation variables.
- BIC values for various covariance structures are taken into account when choosing a model.

3.1.2. Binary Distributions:

- For binary data (variables are 0 or 1), finite mixtures of binary distributions are employed.
- It is possible to capture associations between variables by segmenting respondents based on their propensity to engage in various activities, and segments then explain the association between variables.
- Information criteria like AIC, BIC, and ICL help to select the number of segments.
- The probabilities of seeing a 1 for each variable for each segment are represented by parameters.

3.2. Finite Mixtures of Regressions:

- Finite mixtures of distributions are similar to distance-based clustering methods.
- These methods assume the existence of multiple segments in the data, each following a different regression relationship.
- The target variable y is explained by a set of independent variables I. Different market segments have different regression relationships.
- Usual steps of modelling are followed here like model fitting, visulisation result, summarising coefficient.
- Coefficients' point estimates, standard errors, z-test statistics, and p-values are presented.
- The approach allows for comparison of different models using information criteria like AIC, BIC, or ICL.

4. Algorithms with Integrated Variable Selection

4.1. Biclustering Algorithms:

- Biclustering simultaneously clusters consumers and variables.
- extracts market segments where consumers have the same value of 1 for a set of variables by concentrating on binary data.
- It rearranges rows and columns to create a rectangle with 1s at the top left. Assign observations falling into the rectangle to a bicluster. Remove assigned rows and repeat until no more biclusters can be found.
- Useful for data with many segmentation variables. Captures niche markets.
- Retains original data without transformation.

4.2. Variable Selection Procedure for Clustering Binary Data (VSBD):

- This is a method for binary data clustering, focusing on relevant variables.
- It selects a subset of observations and searchs for the best subset of a small number of variables and gradually add variables to minimize within-cluster sum-of-squares. Finally Stop when increase in sum-of-squares exceeds a threshold.
- It is based on the k-means algorithm, optimizes clustering solution.
- Recommends using multiple random initializations to enhance robustness.

4.3. Variable Reduction: Factor-Cluster Analysis:

- It is a two-step procedure involving factor analysis followed by clustering.
- When the original number of segmentation variables is too high relative to sample size and a validated psychological test battery's factors are relevant.
- It discards original data, use factor scores to extract market segments. Identify number of factors and threshold for retaining factors.
- It can cause loss of information due to factor analysis. Transformation of data changes the nature of information. Difficult interpretation of segment profiles based on factors.

5. Algorithms with Integrated Variable Selection

5.1. Cluster Indices

5.1.1. Internal Cluster Indices:

- These indices focus on aspects of compactness and separation of clusters within the solution. They help provide insight into whether the segments within the solution are distinct and well-separated or not.
- **Compactness Measurement**: This type of index assesses how similar the members of the same segment are. It calculates the sum of distances between each segment member and their segment representative (centroid).
- **Separation Measurement:** This type of index evaluates how different segments are from each other. It measures the weighted distances between the centroids of segments.
- **Combined Indices**: Some indices combine both compactness and separation measures to provide a comprehensive evaluation of the segmentation solution.

5.1.2. External Cluster Indices:

- External cluster indices assess the quality of a market segmentation solution by comparing it with external information or a reference solution.
- **Comparison with Known Solution:** External cluster indices are used when a known correct segmentation solution is available.
- **Comparison with Repeated Calculations:** If the true segment structure is unknown, multiple segmentation solutions can be produced by applying clustering algorithms repeatedly or by varying the data.

- **Correction for Agreement by Chance:** The correction factor aids in addressing the problem of segment sizes affecting index values. It adjusts the index values based on what would be expected by chance given the segment sizes.

5.2. Gorge Plots:
- Gorge plots are a particular kind of visualization used in market segmentation analysis to assess how distinct and similar segments are. They also aid in understanding how consumers relate to segment representatives.
- Gorge plots display how these values are distributed within each segment.
- Gorge plots offer insights into the effectiveness of the segmentation solution by making the degree of similarity and separation within segments more visible.

5.3. Global Stability Analysis:
- Global stability analysis is an approach used to evaluate the stability of market segmentation solutions across repeated calculations, especially when dealing with data lacking clear, well-separated segments.
- It generates new data sets using bootstrapping techniques, extracting multiple segmentation solutions using various algorithms, and comparing the stability of these solutions through similarity measures like the adjusted Rand index.

5.4. Segment Level Stability Analysis
5.4.1. Segment Level Stability Within Solutions (SLSW):
- Segment Level Stability Within Solutions (SLSW) is a to evaluate the stability of market segmentation solutions at the segment level and focused on assessing the stability of individual market segments within a segmentation solution.
- The SLSW is calculated by generating bootstrap samples, creating segmentation solutions for each sample, and then determining the agreement between the original segment and the segments in each bootstrap sample. If a segment consistently retains its identity across the bootstrap samples, it has high SLSW, indicating stability.
- This approach helps in identifying segments that are stable and can be relied upon for subsequent marketing actions. The SLSW concept is particularly useful when organizations are interested in targeting specific segments for their strategies.

5.4.2. Segment Level Stability Across Solutions (SLSA)

- The goal of Segment Level Stability Across Solutions (SLSA), which assesses the recurrence of a market segment across segmentation solutions with various numbers of segments, is to assess the stability of market segmentation solutions.
- In order to determine the SLSA, a variety of partitions (segmentation solutions) with different numbers of segments are taken into account.
- High values of SLSA indicate that a segment is stable and naturally occurring across different segmentation solutions.

Step 7: Describing Segments

1. Developing a Complete Picture of Market Segments:

- In step 7, segments are described using additional information about segment participants that is distinct from the segmentation variables.
- In order to describe segments, a variety of factors are used, including psychographics, demographics, media exposure, and attitudes toward particular goods and brands.
- Understanding segments and customizing marketing strategies benefit from clearly defined segment descriptions.
- Descriptive statistics and visualizations help highlight differences between segments in terms of descriptor variables, making the process more accessible than traditional statistical testing and tabular presentations.

2. Using Visualisations to Describe Market Segments:

2.1. Nominal and Ordinal Descriptor Variables

- **Data Preparation:**
  In order to learn more about the people or things contained within the segments, start by identifying the descriptor variables in of dataset. Additionally, make segment membership data accessible.
- **Segment Sizes:** Understand the size of your market segments. Make a list of the number of people or entities in each segment, and this will give a good idea of how respondents are distributed across the different segments.
- **Cross-tabulation and Data Enrichment:** Include segment membership as a categorical variable within the descriptor data frame. This enables you to compare descriptor variables to segment membership in cross-tabulation.
- **Visualizing Segment Differences:**
  Stacked bar charts are helpful to grasp the composition of each segment in terms of a specific descriptor variable However, keep in mind that comparing proportions between segments of different sizes can be difficult.

- **Mosaic Plots for Better Comparisons**: Mosaic plots are an effective solution for visually comparing proportions across segments, particularly when segments have different sizes.
- **Incorporating Inferential Statistics**:
  Mosaic plots can integrate inferential statistics by color-coding cells to highlight significant deviations from expected frequencies under the assumption of independence. Colors indicate positive and negative differences and reveal statistically significant associations.
- **Exploring Variable Associations:** Use the same guidelines to investigate associations between segment membership and other descriptor variables**.** Visualizing the relationships can reveal trends and help identify crucial traits of various segments.

.

2.1.1.   Metric Descriptor Variables

Graphical Exploration of Metric Descriptor Variables:

- **Histograms and Conditional Plots:** Histograms or conditional plots can be helpful exploring the distribution of a metric descriptor variable across different segments. They assist in understanding whether there are noticeable differences in the variable's distribution among segments.
- **Box-and-Whisker Plots in Parallel**: This graphical tool displays the distribution of a metric descriptor variable for each segment which eventually allows for an insightful comparison of the variable's characteristics across segments. By comparing medians, ranges, and outliers, it helps identify segments with distinct properties and statistically significant differences.
- **Segment Level Stability Across Solutions (SLSA) Plot**: Using color-coded nodes, this plot showcases the variation in the metric descriptor value for each segment. . It highlights segments consistently exhibiting certain metric values across different solutions, providing a deeper understanding of the segments' stability and characteristics.

3. Testing for Segment Differences in Descriptor Variables

- **χ2-Test (Chi-Squared Test):**
  The χ2-test evaluates the association between two categorical variables, such as segment membership and another nominal or ordinal variable. This test quantifies if there's a significant relationship between these variables. The resulting p-value indicates the

likelihood of observing the given association under the assumption of no true association. A small p-value (< 0.05) suggests significant differences.

- **t-Test:**

  When comparing the means of a metric descriptor variable across multiple segments, a t-test is employed. Pairwise t-tests are often used for this purpose. These tests reveal whether observed differences in means are statistically significant. To avoid inflated error rates due to multiple comparisons, p-values should be adjusted using methods like the Holm procedure. The t-test output provides compact yet valuable information about differences between specific segments.

- **ANOVA (Analysis of Variance):**

  ANOVA is an effective tool to test for significant differences in means among more than two groups. It assesses whether the observed variations between group means are statistically significant. The F-statistic, derived from ANOVA, is compared to a critical value to determine significance. A low p-value (< 0.05) indicates at least two segments exhibit different mean levels of the metric descriptor variable.

- **Summarizing and Visualizing Results:**

  Segment characteristics can be summarized in tables that display mean or median values of metric descriptor variables for each segment. These summaries provide a quick overview, and p-values from ANOVA indicate whether differences are statistically significant. Parallel boxplots can also visualize these differences, offering insights into the distribution and direction of variation.

- **Tukey's Honest Significant Differences (HSD) Plot**:

  For a comprehensive view of pairwise differences in means, Tukey's HSD plot is employed. This plot compares mean differences between all pairs of segments, showing point estimates and confidence intervals. If the confidence interval crosses zero, the difference is not significant. Otherwise, it signifies significant differences between segments.

4. Predicting Segments from Descriptor Variables.

4.1. Binary Logistic Regression

- A dependent variable is assumed to follow a Bernoulli distribution with a success probability in binary logistic regression. This approach predicts the likelihood of segment membership for individuals based on their characteristics.

- In this process, regression coefficients express changes in the log odds of success. These odds relate to the probability of being in a particular segment. The model estimates coefficients for each descriptor variable, indicating how they affect the log odds.

- The significance of these coefficients helps identify critical variables for segment identification.

- Model fit is evaluated through metrics like the deviation and the AIC (Akaike Information Criterion). Model selection methods like stepwise variable selection aid in choosing relevant independent variables, preventing overfitting.

4.2. Multinomial Logistic Regression

- Multinomial logistic regression is a statistical technique used to predict segment membership when dealing with multiple market segments. This method is suitable for scenarios where the dependent variable is categorical, representing different segments, and follows a multinomial distribution.

- Model assessment involves checking the significance of dropping individual variables using tests like the LR Chi-square test and analyzing the residual deviance. This can be visualized using mosaic plots and boxplots of predicted probabilities for different segments.

- The effects of independent variables on segment probabilities can be visualized using effect plots. These plots show how predicted probabilities change with varying values of the independent variables.

- The interpretation of coefficients in multinomial logistic regression involves understanding how they influence the log odds of being in a certain segment compared to the baseline segment.

4.3. Tree-Based Methods

- These methods are part of supervised machine learning and have distinct advantages. However, these methods can yield unstable results due to their sensitivity to small data changes.

- CARTs employ a stepwise process to construct a model, recursively partitioning data into groups based on independent variables. The goal is to create groups that are as pure as possible concerning the outcome variable. The resulting tree structure visualizes the splitting steps, nodes, and terminal nodes.
- Parameters such as split criteria, stopping criteria, and final prediction rules influence the tree construction process.
- Classification trees are suited for binary outcomes but can be extended to categorical outcomes with multiple levels. The process involves selecting a split variable, determining split points, and creating terminal nodes with associated predictions.
- Visualizations, like tree plots, enhance the interpretability of the models. They display the splits, nodes, and terminal nodes, along with the proportion of observations belonging to different categories.