C Meyer – Machine Learning Final Project

**Introduction**

To the Advertising Director,

As you know, CRISA has been preparing for its upcoming marketing campaign in the Indian urban market. As part of this initiative, a representative sample of 600 households has been selected so that an analysis of their demographic features and shopping trends can inform key campaign strategy. Based on an examination of purchase behavior and basis, a market segmentation using *k*-means clustering has been performed which divides these customers into one of four categories. This report will summarize the process and results of this analysis.

**Methods**

Given our company’s increasing interest in brand loyalty, customers’ purchase histories were used to derive a loyalty score by taking the proportions of all within-brand purchases across the eight brands that we track. This maximum proportion across columns was taken as a direct proxy for brand loyalty. Additionally, consecutive brand purchases have been included to capture a second dimension of brand loyalty. The hyperparameter *k* was determined through an application of silhouette plotting to maximize between-cluster distance while also minimizing within-cluster variance; this was done to promote differentiation between segments. In total, three rounds of *k*-means clustering were carried out: first using customers’ purchase behavior, second using their purchase basis, and third using both. Prior to clustering, data were normalized to eliminate bias from differences in unit scaling. After clustering, results were displayed in pre-normalized format to maximize interpretability.

**Results**

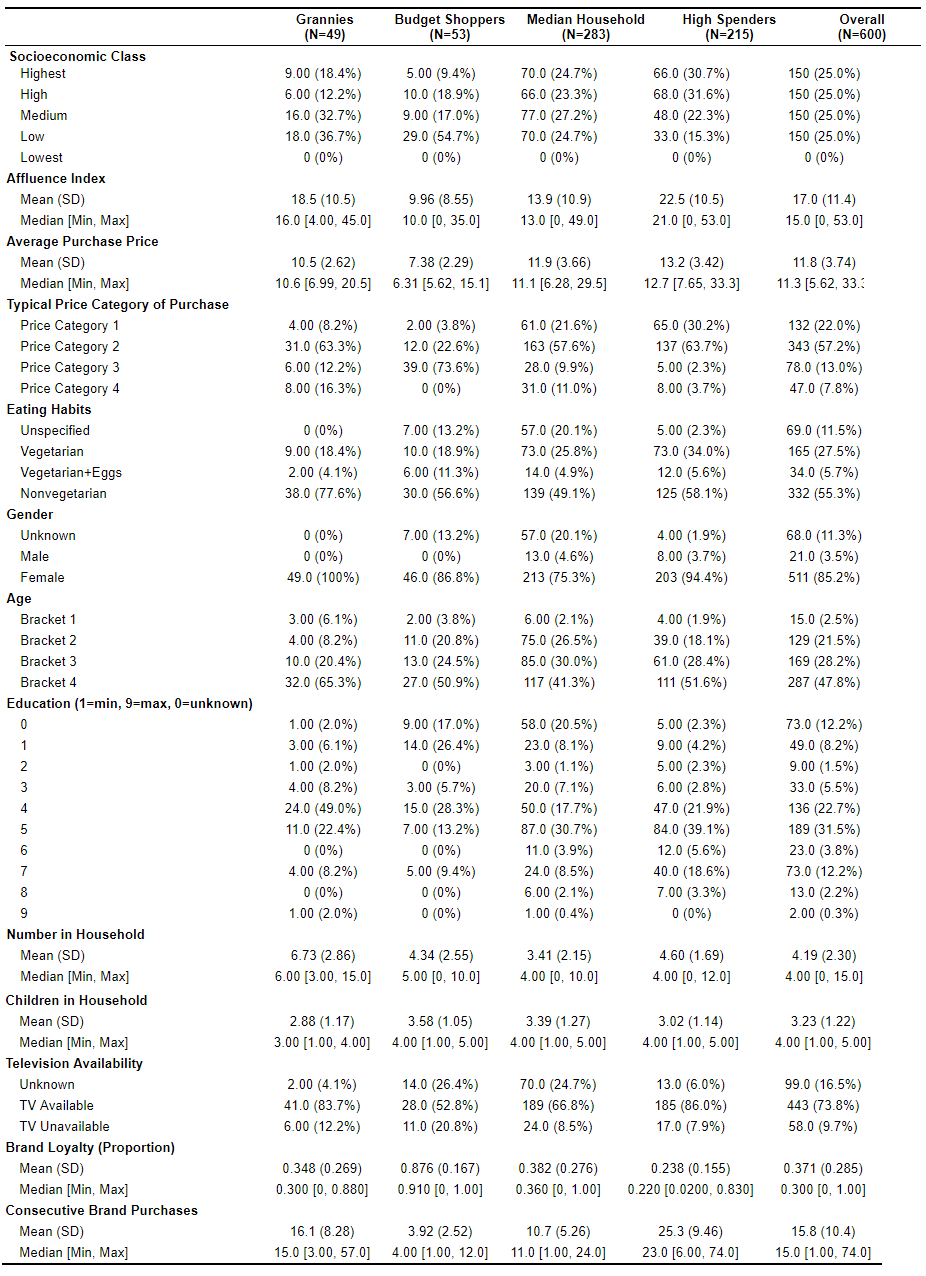
The first round of clustering was conducted on customer’s purchase behavior, which resulted in three broad patterns. The first group (*n*=283) demonstrated above-average brand loyalty but tended to be less affluent and made medium-sized purchases. The second (*n*=89) showed low brand loyalty overall but tended to be more affluent and more prone to brand runs. They displayed a high number of transactions, which were more likely to be from the promo category of products. The third group (*n*=228) displayed higher brand loyalty and tended to make larger purchases with higher values. There was considerable similarity among these groups on other measures.

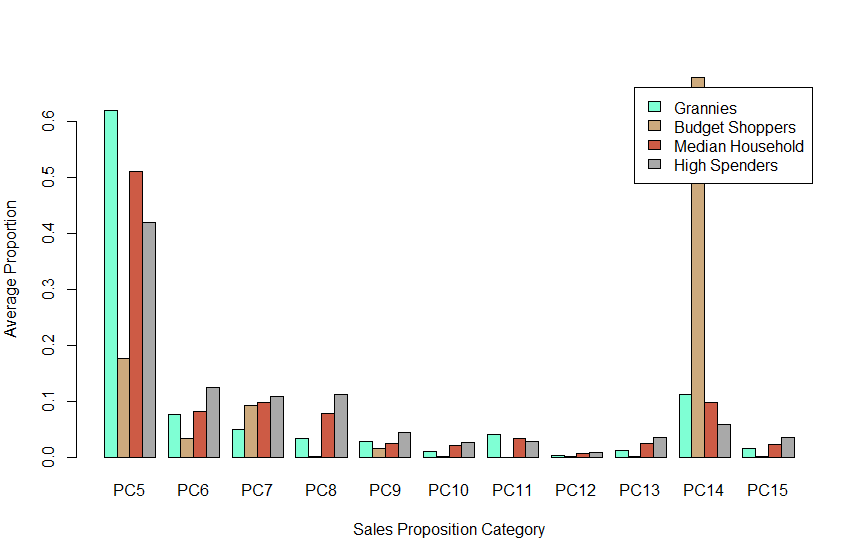
In the second clustering schema, customers were classified using information on their basis for purchases. This produced two buckets of customers, but the majority by far (*n*=522) fell into the first. This group tended to make purchases from price category 2 (55%) or 1 (31%) based on proposition categories 5 (51%) and 15 (28%). The second group (*n*=78) very strongly preferred price category 3 (77%) and proposition category 14 (77%).

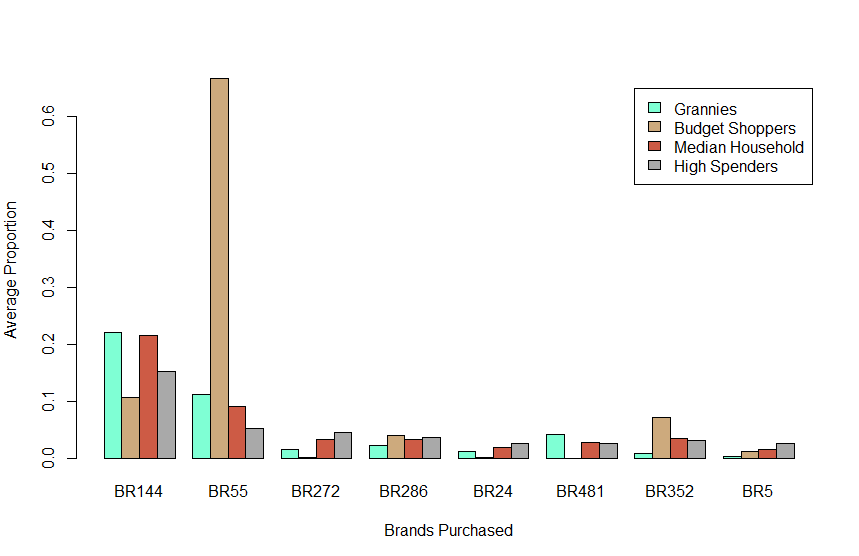
Neither of these features alone was sufficient for our purposes, so the final algorithm was run using both sets of customer features. This resulted in four trends which combined features from both areas above. Here, the biggest group (*n*=283) was who we are calling the Median Households. This group was socioeconomically diverse and somewhat brand loyal. Less than 40% of their purchases were made from a consistent brand, and half of their purchases came from proposition category 5. The second group (*n*=215) were our High Spenders, who tended to come from slightly larger households and higher socioeconomic categories with correspondingly higher levels of affluence and mean purchase price. The High Spenders were the least brand loyal group by brand proportion, though they did have the highest consecutive brand purchases, suggesting that they tend to cycle through brands. Third, a segment (*n*=49) of our customer base tended to be older. This group we are calling the Grannies, as not only were they older, all of them were women, and they typically had large households with fewer children.

These three groups had roughly analogous shopping behaviors, all preferring items at price category 2. All were likely to have access to televisions, though the Median Household’s access was limited relative to the other two. All three had middling brand loyalty, ranging from .24 to .38. They had slightly different patterns in sales propositions, however. The Grannies tended to be more singly focused on proposition category 5 (over 60%), whereas the Median Households and the High Spenders had a more balanced portfolio of secondary categories, more evenly distributed across PC6, PC7, and PC8 at around 10% each. Additionally, although the Grannies had similar brand behavior to the Median Households, the Grannies generally ran higher consecutive brand runs. Thus, while all three groups cycled between brands, the Median Households switched most often, followed by the Grannies, then the High Spenders. Brand 144 was the most commonly purchased for these three groups, but this preference was not strong overall. These three differed in price category trends, however; all three preferred category 2, but the Grannies leaned towards categories 3 and 4 as secondary while the Median Households and the high Spenders preferred category 1 over 3 and 4.

The fourth cluster deviated substantially from the other three. This group (*n*=53) has been named the Budget Shoppers, as they were considerably more likely to come from a low socioeconomic class, to have an affluence index below 10, and to make purchases with a typical value of 7.4. This group strongly preferred items in price category 3 followed by category 2, had more children, and were less likely to have access to a television. However, this group displayed very strong brand loyalty overall. Although their consecutive brand purchases were quite low, nearly 90% of their purchased products came from the same brand. Their purchase basis was also very different, preferring proposition category 14 almost exclusively, followed by categories 5 and 7 as secondary motivators. Another notable difference can be seen between clusters’ television ownership; while the High Spenders and Grannies overwhelmingly had access, only half of the Budget Shoppers were known to have one alongside two thirds of the Median Households.







**Conclusions**

Market segmentation revealed four groups of customers, of which three had several similarities that will streamline the marketing process. First, the Median Households, Grannies, and High Spenders, which were the vast majority of our customers, were motivated to make purchases largely based on proposition 5. This should therefore be our primary focus in this category. The second focus should be on proposition 14, which was the primary concern of the Budget Shoppers and a secondary concern for the Median Households and the Grannies. Second, our advertisements should highlight products meeting price category 2, which was the most common for the first three groups and the second most common for the fourth. Our materials should cover a range of products, including Brand 144 and Brand 55 as the most preferred brands among our customers. We should highlight these products in ways that are consistent with propositions 5, 6, 7, 8, and 14. Our advertisements should be tailored to meet these segments’ preferences as outlined below, and this campaign should be diversified to meet our full audience, many of whom lack access to televisions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Customer Segment Strategies |  |  |  |  |
|  | Median Households  (*n*=283) | High Spenders  (*n*=215) | The Grannies  (*n*=49) | Budget Shoppers  (*n*=53) |
| Socioeconomic Class | Mixed | High-Medium | Medium-Low | Low |
| Price Category (Primary) | 2 | 2 | 2 | 3 |
| Price Category (Secondary) | 1, 4 | 1 | 4, 3 | 2 |
| Preferred Brand | 144 | 144 | 144 | 55 |
| Brand Loyalty (Proportion) | 38% | 24% | 35% | 88% |
| Brand Loyalty (Consecutive) | 11 | 25 | 16 | 4 |
| Sales Proposition (Primary) | 5 | 5 | 5 | 14 |
| Sales Proposition (Secondary) | 14, 7, 6, 8 | 8, 6, 7 | 14, 6, 7 | 5, 7 |

A brief presentation summarizing these findings has been prepared and will be delivered at your convenience. I will be available to answer any questions during this presentation, but as always, if there is any additional information which you may find helpful in the meantime, do not hesitate to ask.