23.6 SEGMENTING CONSUMERS OF BATH SOAP

Refer to 22.6 CRISA bath-soap IMRB Solution.xlsx.

BUSINESS SOLUTION

CRISA is an Asian market research agency that specializes in tracking consumer purchase behavior in consumer goods (both durable and non-durable). In one major research project CRISA tracks about 30 product categories (e.g. detergents, etc.) and within each category, perhaps dozens of brands. To track purchase behavior, CRISA constituted about 50,000 household panels in 105 cities and towns in India, covering most of the Indian urban market. The households were carefully selected using stratified sampling to ensure a representative sample; a subset of 600 records is analyzed here. The strata were defined on the basis of socio-economic status, and the market (a collection of cities).

CRISA has both transaction data (each row is a transaction) and household data (each row is a household), and, for the household data, maintains the following information:

- Demographics of the households (updated annually)
- Possession of durable goods (car, washing machine, etc. updated annually; an "affluence index" is computed from this information)
- Purchase data of product categories and brands (updated monthly).

CRISA has two categories of clients: (1) Advertising agencies that subscribe to the database services, obtain updated data every month, and use the data to advise their clients on advertising and promotion strategies. (2) Consumer goods manufacturers which monitor their market share using the IMRB database.

Key Problems

CRISA has traditionally segmented markets on the basis of purchaser demographics. They would like now to segment the market based on two key sets of variables more directly related to the purchase process and to brand loyalty:

1. Purchase behavior (volume, frequency, susceptibility to discounts, and brand loyalty), and

2. Basis of purchase (price, selling proposition)

Doing so would allow CRISA to gain information about what demographic attributes are associated with different purchase behaviors and degrees of brand loyalty, and thus deploy promotion budgets more effectively. More effective market segmentation would enable CRISA's clients (in this case, a firm called IMRB) to design more cost-effective promotions targeted at appropriate segments. Thus, multiple promotions could be launched, each targeted at different market segments at different times of a year. This would result in a more cost-effective allocation of the promotion budget to different market-segments. It would also IMRB to design more effective customer reward systems and thereby increase brand loyalty.

Data

File: IMRB_Summary_Data.xls

Sheet: DM_Data

The data in this sheet profile each household – each row contains the data for one household.

Member Identification	**Member id**		Unique identifier for each household
Demographics	**SEC**	1-5 categories	Socio Economic Class (1=high, 5=low)
	FEH	1-3 categories	Food eating habits (1=veg, 2=veg. but eat eggs, 3=non veg., 0=not specified)
	MT		Native language (see table in sheet)
	SEX	1=male 2=female	Sex of homemaker
	AGE		Age of homemaker
	EDU	1-9 categories	Education of homemaker (1=minimum, 9=maximum)
Demographics	**HS**	1-9 categories	Number of members in the household
	CHILD	1-4 categories	Presence of children in the household
	CS	1-2	Television available 1. Available 2. Not Available
	Affluence Index		Weighted value of durables possessed

Summarized Purchase Data

Purchase summary of the household over the period	**No. of Brands**		Number of brands purchased
	Brand Runs		Number of instances of consecutive purchase of brands

	Total Volume		Sum of volume
	No. of Trans		Number of purchase transactions; Multiple brands purchased in a month are counted as separate transactions
	Value		Sum of value
	Trans / Brands Runs		Avg. transactions per brand run
	Vol / Tran		Avg. volume per transaction
	Avg. Price		Avg. price of purchase
Purchase within Promotion	**Pur Vol No Promo - %**		Percent of volume purchased under no-promotion
	Pur Vol Promo 6 %		Percent of volume purchased under Promotion Code 6
	Pur Vol Other Promo %		Percent of volume purchased under other promotions

Measuring Brand Loyalty

Several variables in this case deal measure aspects of brand loyalty. The number of different brands purchased by the customer is one measure. However, a consumer who purchases one or two brands in quick succession then settles on a third for a long streak is different from a consumer who constantly switches back and forth among three brands.
So, how often customers switch from one brand to another is another measure of loyalty. Yet a third perspective on the same issue is the proportion of purchases that go to different brands – a consumer who spends 90% of his or her purchase money on one brand is more loyal than a consumer who spends more equally among several brands.

All three of these components can be measured with the data in the purchase summary.

Brand wise purchase	**Br. Cd. (57, 144), 55, 272, 286, 24, 481, 352, 5 and 999 (others)**		Percent of volume purchased of the brand
Price category wise purchase	**Price Cat 1 to 4**		Percent of volume purchased under the price category
Selling proposition wise purchase	**Proposition Cat 5 to 15**		Percent of volume purchased under the product proposition category

Assignments

1. Use k-means clustering to identify clusters of households based on
    a. The variables that describe purchase behavior (including brand loyalty). Variables used: *#brands, brand runs, total volume, #transactions, value, Avg. price, share to other brands, max to one brand.*
    b. The variables that describe basis-for-purchase. Variables used: Pur_vol_no_promo, Pur_vol_promo_6, Pur_vol_other, all price categories, selling propositions 5 and 14 (most people seemed to be responding to one or the other of these propositions – see sheet SellingProps).
    c. The variables that describe both purchase behavior and basis of purchase. All above variables were used.

Note 1: How should k be chosen? Think about how the clusters would be used. It is likely that the marketing efforts would support 2-5 different promotional approaches. *K = 2, 3 and 4 were tried. In deciding what k to use (and also how many variables to include), the following factors should be considered: How distinct are the clusters? Is good separation achieved? How consistent are they? If cluster #1 shows low values on one measure of brand loyalty, does it also show low values on other measures of brand loyalty? How simple are they to describe? Simple clusters are more interpretable by domain knowledge experts, easier to take action on, and are more likely to be statistically stable (i.e. not artifacts of random chance).*

Note 2: How should the percentages of total purchases comprised by various brands be treated? Isn't a customer who buys all brand A just as loyal as a customer who buys all brand B? What will be the effect on any distance measure of using the brand share variables as is? *As is, a customer who buys all brand A is very distant from a customer who buys all brand B, which is probably not what we want. Consider using a single derived variable. The single variable used here was "maximum share devoted to one of the main brands." (A main brand is one of the ones included as a variable.) One additional variable was also used – the proportion devoted to other than main brands. (This is not a clear cut decision. This proportion could be high because the consumer is a maverick type – buying esoteric brands. In such a case, we want to be including the "other" variable as a marker of a different type of consumer. On the other hand, it could be high because the consumer buys a lot of a secondary brand that just misses the main brand cutoff. In such a case, we may be mis-measuring – it would be better, though not possible given these data, to include such a brand in the main brand list.)*

2. Select what you think is the best segmentation and comment on the characteristics (demographic, brand loyalty and basis-for-purchase) of these clusters. (This information would be used to guide the development of advertising and promotional campaigns.) *The spreadsheet IMRB_Solution.xls presents 8 clustering schemes. The outputs 1, 2 and 3 present clusterings into 2, 3 and 4 clusters on the basis of brand loyalty and other purchase behavior variables. Outputs 5, 6 and 7 present clusterings into 2, 3 and 4 clusters on the basis of basis for purchase variables (price and selling proposition). Outputs 8 and 9 present clusterings into 2 and 3 clusters on the basis of both sets of variables. Which scheme to select is a matter of judgement. The initial clustering into two clusters on the basis of brand loyalty is selected here because it is simple, achieves good separation, and focuses on an area of prime interest to the client. Cluster*

1 is brand loyal, relatively low volume, responsive to middling-high price, and relatively less affluent and with bigger households. Cluster 2 is less brand loyal, higher volume, responsive to a slightly higher price, relatively more affluent and with smaller households and more access to television. Another contender was the 2-cluster model, using basis-of-purchase variables (achieves good separation on both counts). Also the 3-cluster model using all variables.

3. Develop a model that classifies the data into these segments. Since this information would most likely be used in targeting direct mail promotions, it would be useful to select a market segment that would be defined as a "success" in the classification model. *The likely candidate for an initial targeting of efforts would likely be cluster 2 – it is more affluent, relatively price insensitive, purchases in higher volume, and is not brand loyal. So a success is initially defined as cluster 2. Three models are presented – logistic regression, neural net, and classification tree. Looking at the lift charts, the neural net seems to do best.*

APPENDIX

Although they are not used in the assignment, two additional data sets are provided that were used in the derivation of the summary data.

IMRB_Purchase_Data is a transaction database, where each row is a transaction. Multiple rows in this data set corresponding to a single household were consolidated into a single household row in IMRD_Summary_Data.

The Durables sheet in IMRB_Summary_Data contains information used to calculate the affluence index. Each row is a household, and each column represents a durable consumer good. A "1" in the column indicates that the durable is possessed by the household; a "0" indicates it is not possessed. This value is multiplied by the weight assigned to the durable item. For example, a "5" indicates the weighted value of possessing the durable. The sum of all the weighted values of the durables possessed equals the Affluence Index.