Segmenting Consumers of Bath Soap

Brandon Winder

12/15/2020

## Introduction

This problem focused on creating good market segmentations that would allow IMRB to design more cost-effective promotions for bath soaps. This allows for multiple different promotions to be offered throughout the year targeting different clientele. The data set provided contains 600 households with 46 different variables from the Asian market research agency, CRISA. CRISA usually clusters markets based on purchasers demographics but would now like to use variables more directly related to the purchasing process. These key variables are purchase behavior and basis of a purchase. All of the variables of interest are broken up into categorical or quantitative below:

**Categorical:**

*SEC*: Socioeconomic class (1 = high, 4 = low)

*FEH*: Eating habits (0 = Not Specified, 1 = Vegetarian, 2 = Vegetarian but eats eggs, 3 = non-vegetarian)

*SEX*: Gender (0 = Not Specified, 1 = Male, 2 = Female)

*AGE*: Age of homemaker (1 = 24, 2 = 25-34, 3 = 35-44, 4 = 45+)

*EDU*: Education of homemaker (0 = Not specified, 1 = Illiterate, 2 = Literate, but no formal schooling, 3 = Up to 4 years of school, 4 = 5-9 years of school, 5 = 10-12 years of school, 6 = Some college, 7 = College graduate, 8 = Some graduate school, 9 = Graduate or professional school degree)

*CHILD*: Presence of children in household (1 = Children up to age 6 present (only), 2 = Children 7-14 present (only), 3 = Both, 4 = None, 5 = Not specified)

*CS*: Television availability (0 = Not Specified, 1 = Cable/Broadcast TV, 2 = Unavailable)

**Quantitative:**

*No. of Brands*: Number of brands purchased

*Brand Runs*: Number of times consecutive purchase of a brand

*Total Volume*: Sum of volume

*No. of Trans*: Number of transactions

*Value*: Sum of value

*Avg. Price*: Average price of purchase

*Pur Vol No Promo - %*: Percent of volume purchased without promotion

*Pur Vol Promo 6 %*: Percent of volume purchased with banded offer promotion

*Pur Vol Other Promo %*: Percent of volume purchased with other promotion

*Br. Cd. 57, 144*: Percent of volume purchased for brand Lux Beauty and Lux International White

*Br. Cd. 55*: Percent of volume purchased for brand Lifebouy

*Br. Cd. 272*: Percent of volume purchased for brand Cinthol Lime Fresh

*Br. Cd. 286*: Percent of volume purchased for brand Pond’s Mois.& Cleansing

*Br. Cd. 24*: Percent of volume purchased for brand Pears

*Br. Cd. 481*: Percent of volume purchased for brand Godrej Fair Glow

*Br. Cd. 352*: Percent of volume purchased for brand Hamam Herbal

*Br. Cd. 5*: Percent of volume purchased for brand Dettol

*Others 999*: Percent of volume purchased for other 999 brands

*Max Brand Loyalty*: Maximum percentage from the variables Br. Cd. 55, 272, 286, 24, 481, 352, 5.

*Pr Cat 1*: Percent of volume purchased for Premium Soaps

*Pr Cat 2*: Percent of volume purchased for Popular Soaps

*Pr Cat 3*: Percent of volume purchased for Economy/Carbolic

*Pr Cat 4*: Percent of volume purchased for Sub-Popular

*PropCat 5*: Percent of volume purchased for Beauty

*PropCat 6*: Percent of volume purchased for Health

*PropCat 7*: Percent of volume purchased for Herbal

*PropCat 8*: Percent of volume purchased for Freshness

*PropCat 9*: Percent of volume purchased for Hair

*PropCat 10*: Percent of volume purchased for Skin Care

*PropCat 11*: Percent of volume purchased for Fairness

*PropCat 12*: Percent of volume purchased for Baby

*PropCat 13*: Percent of volume purchased for Glycerine

*PropCat 14*: Percent of volume purchased for Carbolic

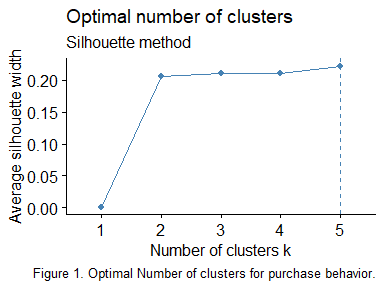
*PropCat 15*: Percent of volume purchased for Others

## Identify Cluster of Households

### For Purchase Behavior

The variables that describe purchase behavior (including brand loyalty) are: *Pur Vol No Promo - %*, *Pur Vol Promo 6%*, *Pur Vol Other Promo %*, *Brand Runs*, *No. of Trans*, *No. of Brands*, *Total Volume*, *Value*, *Others 999*, and *Max Brand Loyalty*. The variable *Max Brand Loyalty* was created to determine if a household was loyal to any specific brand or if they shopped for soaps sold by any brand. This variable was created from the eight brandwise purchase variables (Br. Cd. variables listed above) by selecting the maximum percent from each household. If this percent is greater than the percent for the variable *Others 999*, the household is loyal to one of the eight specific brand.

After selecting the variables of interest and normalizing them, the optimal number of clusters plot has been created using the silhouette method. This plot is shown in Figure 1.



Then the *k*-means clustering method using the optimal number of clusters was run. Figure 2 shows the cluster centroids plot with normalized data and Table 1 shows the centroid values with un-normalized data.

## [1] "The size of cluster 1 is: 165"

## [1] "The size of cluster 2 is: 79"

## [1] "The size of cluster 3 is: 143"

## [1] "The size of cluster 4 is: 55"

## [1] "The size of cluster 5 is: 158"

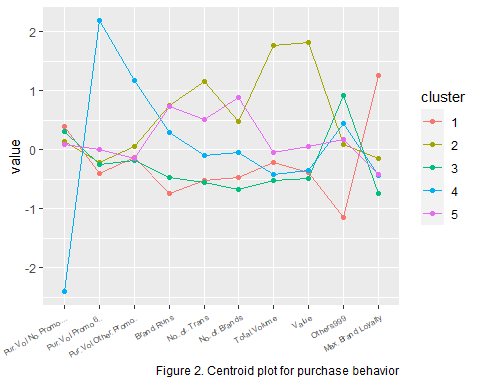


Table 1. Purchase Behavior Centroids

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
| Pur Vol No Promo - % | 96.04 | 92.93 | 94.92 | 62.45 | 92.31 |
| Pur Vol Promo 6 % | 1.55 | 3.34 | 3.00 | 25.69 | 5.36 |
| Pur Vol Other Promo % | 2.41 | 3.73 | 2.08 | 11.86 | 2.33 |
| Brand Runs | 7.98 | 23.52 | 10.82 | 18.71 | 23.42 |
| No. of Trans | 22.07 | 51.28 | 21.43 | 29.49 | 39.96 |
| No. of Brands | 2.88 | 4.39 | 2.58 | 3.56 | 5.03 |
| Total Volume | 10242.39 | 25668.86 | 7858.64 | 8665.09 | 11586.47 |
| Value | 992.38 | 2941.09 | 908.27 | 1031.46 | 1390.70 |
| Others 999 | 17.75 | 54.75 | 79.73 | 65.66 | 57.29 |
| Max Brand Loyalty | 73.14 | 33.05 | 15.88 | 24.70 | 25.10 |

The size of the clusters is not very evenly spread out, Cluster 2 & 4 have a lower number of households while Clusters 1, 3 & 5 have a higher number of households in them. Using the results from Figure 2 and Table 1 the following key variables can be seen:

Cluster 1 (Loyal Cluster): This is the most loyal clusters because it has the lowest percent volume purchased from the *Others 999* and the highest *Max Brand Loyalty*. It also has the lowest *Pur Vol Promo 6%* and *Brand Runs*. Additionally it has the highest *Pur Vol No Promo - %*.

Cluster 2 (High Volume Cluster): This has the highest value for *No. of Trans*, *Total Volume* and *Value*. It also does not have great brand loyalty because its *Max Brand Loyalty* is lower than the *Others 999* value.

Cluster 3 (Un-loyal Cluster): This is the least loyal cluster as it has the highest percent volume purchased in *Others 999* and the lowest *Max Brand Loyalty*. It also has the lowest *No. of Trans*, *No. of Brands*, *Total Volume* and *Value*.

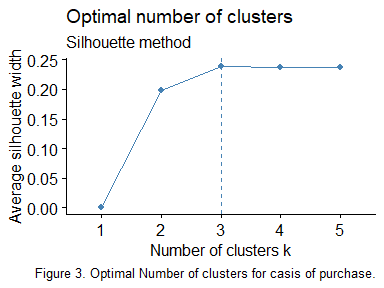
Cluster 4 (Promotions Cluster): This cluster has a very low *Pur Vol No Promo - %* and a very high *Pur Vol Promo 6%* and *Pur Vol Other Promo %*. Additionally it does not have good brand loyalty.

Cluster 5 (Many Brands Cluster): This cluster does not have great brand loyalty and has the highest value for *No. of Brands*.

### For Basis of Purchase

The variables that describe the basis of purchase are: *Avg. Price*, each price category (Pr Cat 1-4), and each selling propositionwise purchase (PropCat 5-15).

Selecting these variables and normalizing them was the first step taken. Then the optimal number of clusters plot was created to find the number of clusters to use for the *k*-means clustering. This plot is shown in Figure 3.



Then the optimal number of clusters from Figure 3 was used to run the *k*-means clustering. The cluster centroid plot with normalized data is shown in Figure 4 and Table 2 displays the centroid values with un-normalized data.

## [1] "The size of cluster 1 is: 462"

## [1] "The size of cluster 2 is: 84"

## [1] "The size of cluster 3 is: 54"

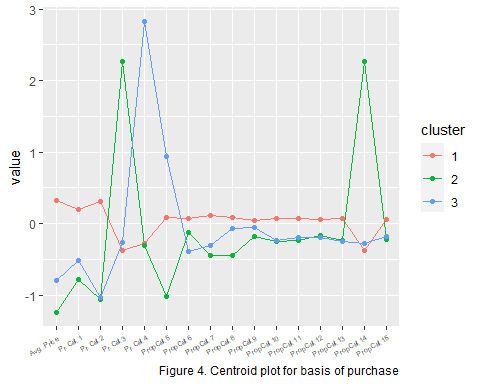


Table 2. Basis of Purchase Centroids

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cluster 1 | Cluster 2 | Cluster 3 |
| Avg. Price | 13.03 | 7.19 | 8.85 |
| Pr Cat 1 | 33.59 | 5.89 | 13.54 |
| Pr Cat 2 | 59.10 | 16.35 | 16.86 |
| Pr Cat 3 | 3.71 | 74.75 | 6.66 |
| Pr Cat 4 | 3.61 | 3.01 | 62.94 |
| PropCat 5 | 48.17 | 13.22 | 75.33 |
| PropCat 6 | 10.39 | 7.03 | 2.76 |
| PropCat 7 | 11.97 | 1.02 | 3.66 |
| PropCat 8 | 9.39 | 1.18 | 6.89 |
| PropCat 9 | 3.33 | 1.90 | 2.76 |
| PropCat 10 | 2.60 | 0.05 | 0.20 |
| PropCat 11 | 3.58 | 0.63 | 1.00 |
| PropCat 12 | 0.76 | 0.19 | 0.09 |
| PropCat 13 | 3.17 | 0.26 | 0.16 |
| PropCat 14 | 3.56 | 73.91 | 6.24 |
| PropCat 15 | 3.08 | 0.61 | 0.90 |

Cluster 1 is more than 5 times larger than the other two clusters. Additionally, from the results shown in Figure 4 and Table 2, the following information can be seen:

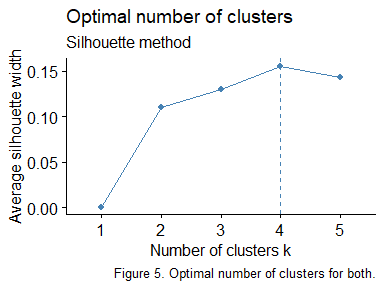
Cluster 1 (Premium/Popular Cluster): This cluster has the highest value for many of the variables, such as: *Avg Price*, *Pr Cat* 1 & 2, *PropCat* 6 through 13 and 15.

Cluster 2 (Economy/Carbolic Cluster): This cluster has very high values for *Pr Cat 3* and *PropCat 14*, while having a very low value for *PropCat 5*. It also has the lowest value for *Avg Price*, *Pr Cat* 1 & 4, and *PropCat* 7 through 11.

Cluster 3 (Sub-popular Cluster): This cluster has a very high value for *Pr Cat 4* and *PropCat 5*. Additionally it has the lowest value for *PropCat 6*.

### Both Purchase Behavior and Basis for Purchase

The optimal number of clusters plot using both the variables that describe purchase behavior and the basis for a purchase is shown in Figure 5.



The optimal number of clusters from Figure 5 is used to run the *k*-means clustering method with all the variables that describe purchase behavior and basis for purchase. The centroid plot with normalized data is shown in Figure 6 and the un-normalized centroid values are displayed in Table 3.

## [1] "The size of cluster 1 is: 220"

## [1] "The size of cluster 2 is: 72"

## [1] "The size of cluster 3 is: 60"

## [1] "The size of cluster 4 is: 248"

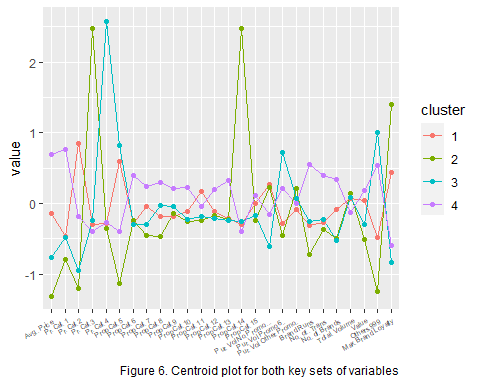


Table 3. Basis of Purchase and Purchase Behavior Centroids

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
| Avg. Price | 11.32 | 6.94 | 8.99 | 14.40 |
| Pr Cat 1 | 14.73 | 5.66 | 14.30 | 49.33 |
| Pr Cat 2 | 75.78 | 12.00 | 19.94 | 43.78 |
| Pr Cat 3 | 6.07 | 80.20 | 7.45 | 3.21 |
| Pr Cat 4 | 3.42 | 2.14 | 58.31 | 3.67 |
| PropCat 5 | 64.48 | 9.97 | 71.80 | 33.15 |
| PropCat 6 | 4.63 | 5.21 | 4.21 | 15.69 |
| PropCat 7 | 8.81 | 0.84 | 3.76 | 14.48 |
| PropCat 8 | 5.26 | 0.80 | 7.64 | 12.65 |
| PropCat 9 | 1.90 | 2.18 | 2.85 | 4.45 |
| PropCat 10 | 1.16 | 0.06 | 0.34 | 3.77 |
| PropCat 11 | 4.66 | 0.51 | 1.17 | 2.54 |
| PropCat 12 | 0.34 | 0.19 | 0.07 | 1.13 |
| PropCat 13 | 0.42 | 0.28 | 0.16 | 5.54 |
| PropCat 14 | 5.83 | 79.53 | 6.88 | 3.09 |
| PropCat 15 | 2.51 | 0.43 | 1.11 | 3.52 |
| Pur Vol No Promo - % | 94.55 | 93.99 | 84.10 | 89.38 |
| Pur Vol Promo 6 % | 2.70 | 1.16 | 12.07 | 7.29 |
| Pur Vol Other Promo % | 2.75 | 4.85 | 3.83 | 3.33 |
| Brand Runs | 12.54 | 8.18 | 13.15 | 21.43 |
| No. of Trans | 26.43 | 24.83 | 27.13 | 38.15 |
| No. of Brands | 3.51 | 2.86 | 2.80 | 4.18 |
| Total Volume | 12453.87 | 13062.36 | 12579.08 | 10942.64 |
| Value | 1377.40 | 891.87 | 1077.23 | 1494.18 |
| Others 999 | 38.06 | 15.08 | 82.08 | 68.29 |
| Max Brand Loyalty | 49.80 | 77.09 | 13.42 | 20.01 |

It is apparent that clusters 1 & 4 are much larger than clusters 2 & 3. Also from Figure 6 and Table 3, the following information can be seen:

Cluster 1 (Popular Fairness Cluster): This is a fairly loyal cluster as it has a higher value for *Max Brand Loyalty* than for *Others 999*. It also has the highest value of *Pr Vol No Promo - %*, *Pr Cat 2*, and *PropCat 11*.

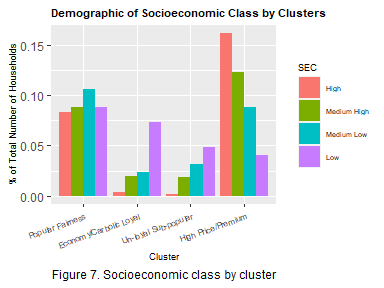
Cluster 2 (Economy/Carbolic Loyal Cluster): This is a very loyal cluster as it has the lowest value for *Others 999* and the highest for *Max Brand Loyalty*. It also has much higher values for *Pr Cat 3* and *PropCat 14*. It also has the lowest value for many variables such as: *Pur Vol Promo 6%*, *Pr Cat* 1, 2 & 4, *PropCat* 5, 7, 8, 10, 11, & 15, *Avg Price*, *Brand Runs*, *No. of Transactions*, and *Value*.

Cluster 3 (Un-loyal Sub-popular Cluster): This is the least loyal cluster with the highest value for *Others 999* and lowest for *Max Brand Loyalty*. This cluster also has the highest value in *Pur Vol Promo 6%*, *PropCat* 3 & 4. It also has the lowest value for *Pur Vol No Promo - %* and *No. of Brands*.

Cluster 4 (High Price/Premium Cluster): This is a fairly unloyal cluster with higher values of *Others 999* and lower *Max Brand Loyalty*. It has the highest values for many variables including: *Pur Cat 1*, *PropCat* 6 through 10, 12, 13 & 15, *Avg. Price*, *Brand Runs*, *No. of Trans*, *No. of Brands*, and *value*. Additionally it has the lowest values for *Pr Cat 3*, *PropCat 14*, and *Total Volume*.

## Characteristics of Clusters

From Figure 1, 3 and 5 the optimal number of clusters are 5, 3, and 4 respectively. Because the problem wants to look at both purchase behavior and the basis of purchases, the results from Figure 5 are the most relevant. Figure 5 shows that four clusters are optimal. Additionally, Figure 1 and 3 show that the average silhouette width for four clusters is only slightly lower than the optimal number of clusters. Therefore the best segmentation method is *k*-means clustering with four clusters. So, the demographics of the four clusters are inspected below.



From Figure 7, it appears that as socioeconomic class lowers so does the percent total of household in the High Price/Premium cluster. Also it shows that the percent total of households in the Popular Fairness cluster does not differ between socioeconomic class. Lastly, as socioeconomic class lowers the percent total of households increases for the Economy/Carbolic Loyal and Un-loyal Sub-popular clusters.

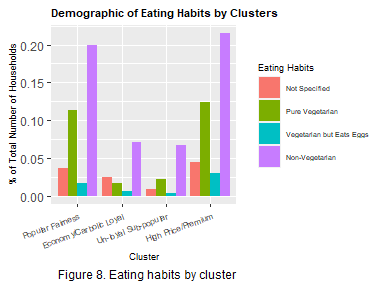


Figure 8 shows there is no clear relationship between eating habits and cluster membership. Non-vegetarian has the highest percentage in each cluster showing that it is the most likely eating habit. But this does not help classify any relationship between eating habits and clusters.

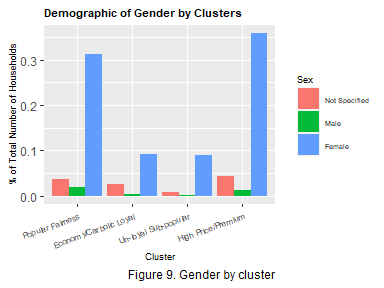


Figure 9 shows that there is a large percentage of females in each cluster showing that the data set is mainly female. The largest percentage of females is in the High Price/Premium cluster which is also not very brand loyal.

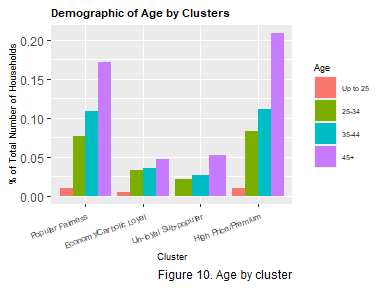
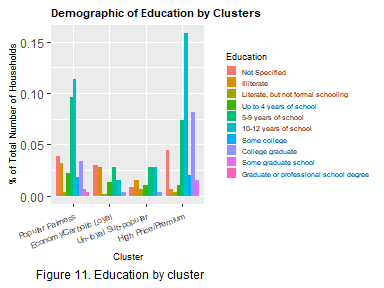


Figure 10 shows that for each cluster as age increases so does the percentage of total number of households. This shows that mostly people over the age of 45 are purchasing soaps. But, there is no significant relationship that can be found between the clusters and age.



The High Price/Premium cluster has the highest percentage of households with 10-12 years of school and college graduates. This suggest that with more education, a person is more likely to purchase premium soaps and also have less brand loyalty. Also the Economy/Carbolic cluster has a large proportion of illiterate homemakers. This suggest that the less educated the homemaker of the household is, the more brand loyalty they will have.

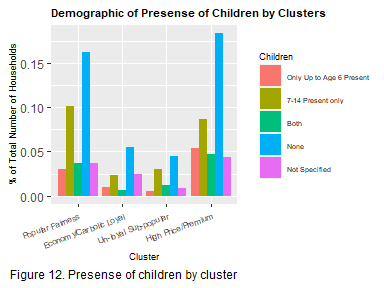


Figure 12 shows that the largest percentage is having no children present in the household for every cluster. The middle two clusters (Economy/Carbolic Loyal and Un-loyal Sub-popular) have the largest proportion of only children ages 7-14 present. But the Economy/Carbolic cluster has fairly good brand loyalty and the Un-loyal Sub-popular cluster has the worst brand loyalty. So, there is no good relationship between clusters and children preset in the household.

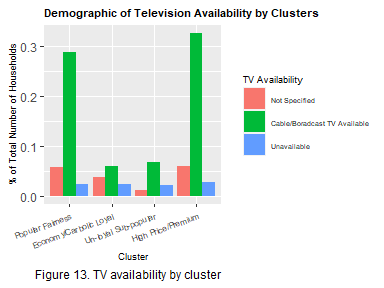


Figure 13 shows that each cluster has a large percentage of household that have television availability. The cluster with the largest proportion of no television available is the Economy/Carbolic Loyal cluster.

### Summary and Recommendations

The majority of the clientele is female whose household has television available. The High Price/Premium cluster has the largest proportion of college graduates and as the socioeconomic class lowers so does the percent of households in the cluster. Additionally this cluster is not very brand loyal and has a high number of transactions and brands. So, broadcasting any type of premium soap on television would be an effective way of promoting products specifically for this cluster.

The Popular Fairness cluster is more loyal and tends to purchase soaps without the use of promotions. Also the proportions within the cluster do not differ much between socioeconomic class. From this it is clear that marketing to this cluster could be done through TV ads specifically for popular soaps used for fairness.

Lastly the Economy/Carbolic Loyal cluster has a very low average price with high use of promotions categorized as other. Additionally this cluster has a large proportion of illiterate homemakers and as their socioeconomic class decreases their percent total number of households increases. So, this would be the best cluster to target direct-mail promotions to.

## Developed Model

The first model developed to classify this data into clusters uses a classification tree. First, a new data frame was created containing the purchase behavior, basis of a purchase, and the demographics that showed a relationship between clusters. Additionally, each brand purchase percentage was used instead of the max brand loyalty variable created so the classification tree uses the loyalty households have for specific brands. Then the data was split into a training set and a validation set to test the accuracy of the prediction. The classification tree that was created using these variables can be seen in the image below. The cluster membership for the validation data was predicted using this classification tree. The predictions were then compared to the actual cluster assigned to the household for the validation data. Table 4 shows the classification table.

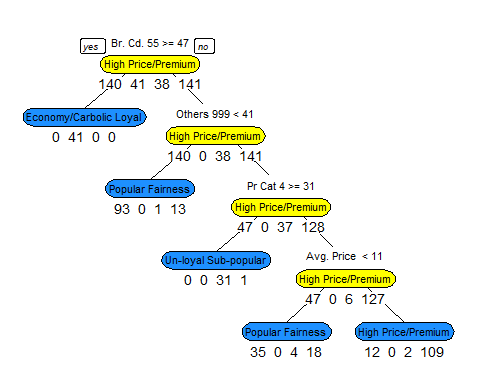


Table 4. Classification table using classification tree model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Popular Fairness | Economy/Carbolic Loyal | Un-loyal Sub-popular | High Price/Premium |
| Popular Fairness | 64 | 1 | 0 | 26 |
| Economy/Carbolic Loyal | 0 | 30 | 0 | 0 |
| Un-loyal Sub-popular | 1 | 0 | 22 | 1 |
| High Price/Premium | 15 | 0 | 0 | 80 |

## [1] "The overall accuracy using the classification tree is: 81.7%"

It is apparent that this model is not highly accurate at predicting cluster membership on the validation data. While each terminal node is fairly dominated by one cluster, the overall accuracy is only 81.7%. But, this model allows for the rules of the classification to be seen. For example, the recommendations above show that the Economy/Carbolic Loyal cluster would be the marketing segmentation classified as a success. So, from this classification tree the rule for a success would be if the percent volume purchased from the brand Lifebouy (Br. Cd. 55) is greater than or equal to 47, then the Economy/Carbolic Loyal cluster is selected. But, because this was not highly accurate with its predictions, the use of a random forest was applied. The classification table for this method can be seen in Table 5.

Table 5. Classification table using random forest model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Popular Fairness | Economy/Carbolic Loyal | Un-loyal Sub-popular | High Price/Premium |
| Popular Fairness | 76 | 0 | 1 | 6 |
| Economy/Carbolic Loyal | 0 | 31 | 0 | 0 |
| Un-loyal Sub-popular | 0 | 0 | 18 | 0 |
| High Price/Premium | 4 | 0 | 3 | 101 |

## [1] "The overall accuracy using the random forest method is: 94.2%"

The random forest model resulted in a much higher overall accuracy on the validation data shows that this model is preferred. Additionally, the model correctly predicted every households into the Economy/Carbolic Loyal cluster, which is classified as a success for direct-mail promotions. So, it is apparent that this random forest model can be used to target direct-mail promotions with fairly high efficiency.