Marketing and Customer Analytics

Individual Assignment
Segmentation

Segmentation Exercise:
This exercise builds directly off the Product Optimization Exercise we just did. The exercise calls for follow up work to characterize the segments. For this exercise, apart from the files from the Product Optimization Exercise, we need demographics data, which are provided in this file: demographics-full.xlsx. The demographics variable are: income (in thousands of dollars per year), age (in years), sports (=1 if the person is sports active, 0 if not), gradschl (=1, if the person has a master's degree or beyond, 0 if not).

If you did the Product Optimization Exercise correctly you have a file similar to this one: mugs-analysis-full-incl-demographics.xlsx
This file contains a correct execution of the core calculations in the Product Optimization Exercise and also contains the demographics data of the demographics-full.xlsx file.

This exercise has two parts, the first on product affinity based segmentation and the second on classical segmentation.

Part (A) Product affinity based segmentation:

For this part, we need to assume a market scenario, which we will take to be the one below:
Incumbents
1: $30, 3 hrs, 20 oz, Clean Easy, Leak Resistant, Brand A
2: $10, 1 hrs, 20 oz, Clean Fair, Spill Resistant, Brand B
Our Candidate
3: $ 30, 3 hrs, 20 oz, Clean Easy, Leak Resistant, Brand C

Brand C's product has the identical attribute level combination as Brand A's product. You may think that Brand C's marketing team is committing an error, but perhaps not because that attribute combination level is actually the one that maximizes expected profit per customer, as you may have discovered in the previous HW. To build your intuition, you should reflect on why it may be optimal for Brand C to mimic Brand A.

As in the previous exercise, we have the following attributes and attribute levels:
Price: $30, $10, $5
Time Insulated: 0.5 hrs, 1 hrs, 3 hrs
Capacity: 12 oz, 20 oz, 32 oz
Cleanability: Difficult (7 min), Fair (5 min), Easy (2 min)
Containment: Slosh resistant, Spill resistant, Leak resistant
Brand: A, B, C
and the following cost structure:
Time Insulated: 0.5 hrs costs $0.5, 1 hrs costs $1, 3 hrs costs $3
Capacity: 12 oz costs $1.00, 20 oz costs $2.6, 32 oz costs $2.8
Cleanability: Difficult (7 min) costs $1, Fair (5 min) costs $2.2, Easy (2 min) costs $3.0
Containment: Slosh resistant costs $0.5, Spill resistant costs $0.8, Leak resistant costs $1

You are given data on the preference parameters of 311 consumers in this file: mugs-

preference-parameters-full.xlsx. This is the same input file used in the Product Optimization Exercise.

Your tasks for Part (A):

1. Compute and report the characteristics of the affinity-based segment for Product 3 (our candidate as given above). Report the characteristics in terms of the following descriptors: IPr, Iin, ICp, ICl, ICn, IBr, I*pPr30, I*pPr10, I*pPr05, I*pIn0.5, I*pIn1, I*pIn3, I*pCp12, I*pCp20, I*pCp32, I*pClD, I*pClF, I*pClE, I*pCnSl, I*pCnSp, I*pCnLk, I*pBrA, I*pBrB, I*pBrC, and the demographics income, age, sports and gradschl. For this you need to compute the weighted average of all the columns in the "for-cluster-analysis" worksheet, weighted by the probabilities given in column BF of the "mugs-full" worksheet. You should not do this in Excel directly. You are to load the data into python or R and compute the weighted averages in R or python; this is to help you build familiarity with these languages and prepare you better for your job.
2. Repeat the step above for the product of brand A and the product for brand B. Compute also the overall mean for each descriptor (this done by a simple average of each descriptor across all 311 customers). Compute the log-lifts for all variables for the affinity based segment for each product and focus on the large positive and negative numbers to figure out how each segment is different from the other segments and the overall population average. (You will use these in the next step to come up with a verbal description that characterizes each segment.)

3. Use your findings from Step 2 above to come up with a verbal description and a persona story that characterizes each segment. The persona story gives a mental image to the marketing manager not only in terms of the descriptors in the dataset, but also in terms of plausible hypothesized characteristics that go beyond the descriptors available.

Steps 1 and 2 above are similar to what we saw in the lecture session's spreadsheet: mugs-analysis-limited-affinity-segmentation.xlsx. The same spreadsheet, with cells highlighted for the analysis of the competitive advantage of product 3 over product 2 is here: mugs-analysis-limited-affinity-segmentation-compare-2-for-customers-of-3.xlsx. The PDF file summarizing the key results in a compact manner **for the lecture's data (not the HW data)** is given here: mugs-analysis-limited-affinity-segmentation-display.pdf. As is clear from the numbers, in the lecture's data, Segment 3 is distinctive in that it has lower price importance, lower capacity importance, higher cleanability importance, higher income, more sports active and more educated.This allows us to form a psychological picture as we did in the lecture. You need to a similar analysis for the HW based on the HW data (which is different from the lecture data).

Hints on checking your answers: If you did your affinity scoring computations correctly, then you should find that for our customers (ie Product 3's customers), price is the most important attribute with an average within-segment mean of 25.96 and that the most distinguishing demographic variable, where the log-lift is furthest away from zero, is gradschl with a log-lift of 0.12737. My computation uses log to the base 10. If you use some other base, you will get a different number of course. It does not matter which base you consider because you are only looking to order the descriptors according to the absolute value of the log-lift to identify the descriptors where the segment id most different from the overall population or the other affinity segments, and these descriptors in turn are used to come up with the persona of the

segment.

Note: A common thumb rule is to take base 10 absolute log lifts greater than 0.04 to be noteworthy and greater than 0.08 to be very noteworthy.

Part (B) Classical segmentation:

Run a kmeans cluster analysis using the following variables as the segmentations basis variables: IPr, Iin, ICp, ICl, Icn, IBr, I*pPr30, I*pPr10, I*pPr05, I*pIn0.5, I*pIn1, I*pIn3, I*pCp12, I*pCp20, I*pCp32, I*pClD, I*pClF, I*pClE, I*pCnSl, I*pCnSp, I*pCnLk, I*pBrA, I*pBrB, I*pBrC. Important: Do not use the demographic variables for the kmeans analysis (the demographics are to be used in the profiling step, not as inputs into the cluster analysis).

You may recognize that the importance parameters like IPr, Iin, ... and the importance times preference terms (the utility contributors) like I*pPr30, I*pPr10, .. are of different magnitudes. Therefore, you may think you should rescale all the terms but please do NOT do that. Input the variables into kmeans() exactly as they are without rescaling. Obviously because the importance times preference terms are larger the cluster analysis will be influenced more by them but that is exactly what we want because the meaningful cluster analysis outputs are those driven by the utility contributors. The Importance parameters like IPr, Iin,.. are of smaller magnitude and used largely as "tie-breakers".

Usually kmeans routines take three arguments in addition to the dataset:

1. The value of "k", the number of clusters to look for. For this, you should try every integer from 2 to 10.
2. The number of runs to try, each run corresponding to a different set of starting seeds. This you should set to 50.
3. The maximum number of iterations (what in the slides I referred to as improvement attempts). This should you should set to be 100.

In the kmeans() function of R, these three arguments are named "centers", "nstart" and "iter.max" respectively. For python's "sklearn.cluster.KMeans" function, these three arguments are named "n_clusters", "n_init", "max_iter" respectively.

For each value of "k", compute the average within-cluster sum of squares averaged across all clusters, actually a weighted average with weights corresponding to the clustersize. In R, this is done as follows: If kmeans.out is the output from kmeans()
function, **ave_within_cluster_mean_sum_of_squares<-
weighted.mean(x=kmeans.out$withinss, w=kmeans.out$size, na.rm = TRUE)**.
In Python, it is done as follows:

```
# X is an np.array containing the data
kmeansModel = KMeans(n_clusters=k, n_init=50, max_iter=100)
kmeansModel.fit(X)
ave_within_cluster_mean_sum_of_squares = (kmeansModel.inertia_) / X.shape[0]
# alternatively you compute it manually as follows
# ave_within_cluster_mean_sum_of_squares = sum(np.min(cdist(X,
kmeansModel.cluster_centers_, 'euclidean')**2, axis=1)) / X.shape[0])
```

Because there is randomness in the behavior of kmeans, and so that we can compare outputs

across all students, we should initialize the seed to a fixed value across before EACH of the 9 calls to the kmeans function. Set the seed to **410014.** In R you do this using by evaluating the expression **set.seed(410014).** In python, you do this using by evaluating the expression **random.seed(410014).**

Create a plot with k from 2 to 10 on the x-axis and the average within-cluster sum of squares on the y-axis. Decide the best value of "k". Note that there is some subjectivity here and the answer may not be self-evident, but the main guiding principle is to pick the "k" where further increase does not achieve improvement in not just average-within-group-variance but also profit, sales, share and other such factors which the manager cares about. These latter aspects are hard to assess without experience and domain expertise: please do the best you can but do not spend more than 10-15 minutes thinking about on this part.

Recall that we discussed in class that there are two ways to pick the value of "k" in cluster analysis.

- The machine learning way: This is based on the plot of average-within-group-variance (like slide 26) and looking at the elbow point where there is not much reduction in average-within-group-variance going from "k" to "k+1". The machine learning way take only a few seconds but will not necessarily lead to the right number of products for maximizing profits.
- The marketing way: This is more tied to product creation decisions. It is more correct in that it likely leads to high profits but it is also more difficult because it involves qualitative judgment. If you follow the marketing way, you can describe your work via Challenge Task 2 and I will give you extra credit towards class participation, commensurate with the quality of your submitted work. If you do not want the extra credit then you can just use the machine learning way, but if you want the possibility of this extra credit then read on for more info on how to implement the marketing way. Note first that there is no single right answer and so the grading will be based not on the value of k chosen but the reasoning that went into your answer. Some comments that may help you: you should try to form a qualitative assessment of the value of "k", beyond which there is little benefit to further segmentation and creating products customized for the segments. One can do this by assessing, when going from "k" to "k+1" segments, whether a segment emerges that has distinct persona and preferences and creating a distinct product to cater to that segment would lead to substantial increment in overall profit, sales or share. If there is no substantial increment in these when going from "k" to "k+1" then "k" is better than "k+1". Related to this are my comments in class about Slide 20: ordinarily using k=4 and splitting the people in the lower right into two segments (relative to k=3, Slide 16) may not be worthwhile because the additional product created to target the additional segment would have only small differences with, and would cannibalize, the product created for the people in the lower right with k=3. However, if the people in the lower right are rich and willing to pay for those small differences, it would in fact lead to substantial increment in overall profit, sales or share by going from k=3 to k=4, in which case k=4 is better than k=3. Whether or not a segment is rich, willing to pay, growing, etc, can be assessed by looking at the segment characteristics.

Once you have decided the best value of "k", for the kmeans cluster analysis output for that best value of k, do the following.

1. Compute and report the characteristics of all segments. Report the characteristics in terms of: IPr, Iin, ICp, ICl, Icn, IBr, I*pPr30, I*pPr10, I*pPr05, I*pIn0.5, I*pIn1, I*pIn3, I*pCp12, I*pCp20, I*pCp32, I*pClD, I*pClF, I*pClE, I*pCnSl, I*pCnSp, I*pCnLk, I*pBrA, I*pBrB, I*pBrC, and also do profiling in terms of the demographics income, age, sports and gradschl.
2. Compute the log-lifts for all variables for all segments and focus on the large positive and negative numbers to figure out how each segment is different from the other segments and the overall population average. (You will use these in the next step to come up with a verbal description that characterizes each segment.)
3. Use your findings from Step 2 above to come up with a verbal description and a persona story that characterizes each segment. The persona story gives a mental image to the marketing manager not only in terms of the descriptors in the dataset, but also in terms of plausible hypothesized characteristics that go beyond the descriptors available.

Steps 1 and 2 above are similar to what we saw in the lecture session's spreadsheet: classical-segmentation-example.xlsx, illustrated on slides 32 to 35. The kmeans and profiling output for k=3 is given in this spreadsheet: classical-segmentation-example-done.xlsx

Hint on checking your answers: If you chose k=4 and did your profiling correctly, then you should get the loglifts for "sports" to be -0.46, 0.32, 0.10, and 0.12. My computation uses log to the base 10. The ordering of the clusters may be shuffled.

What and how to Submit

Submit the following items online via LMS. The page-lengths given below are suggestive only. Feel free to exceed the suggested page lengths if you want to.

1. For Part A: A 2-page PDF file containing the following: the final numbers asked for in Part A's (1) and (2) and the verbal descriptions in (3).
2. For Part B: A 3-page PDF file containing the following: the final numbers asked for in Part B's (1) and (2) and the verbal descriptions for (3), and also the plot justifying your chosen value of "k".
3. Your R or Python code for Parts A, B in the zip format I specified earlier.