Table of Contents

Introduction

To communicate and engage more effectively with the customers customer segmentation is used which splits different characterized customers into the different group. The case describes about the need of the Delta Bank to do customer segmentation in order to know the customers who had and had not subscribe for the long-term bank deposits. As per the data set given segmentation has been done with the help of existing customer profiles which includes Marital status, age, education and career of theirs and on the other side segmentation also done with the help of marketing campaign data of the existing customers which includes contact type, number of contact, mortgage, and personal loan if they have any. On the basis of segmentation, we will identify potential customers who are likely to subscribe for the Delta banks long-term deposit plans.

Demographics based segmentation

The variable Subscribed has been set as a target variable, all the demographic variables as input variables and rejected the other variables demographic segmentation have been derived which is as below.

Data Exploration:

Only Age is the continuous variable so must check the distribution of the variable but from the below graph it seems almost normal distribution. So, does not require the transformation of the variable Age. Other type of transformation is handled by the clustering node itself.

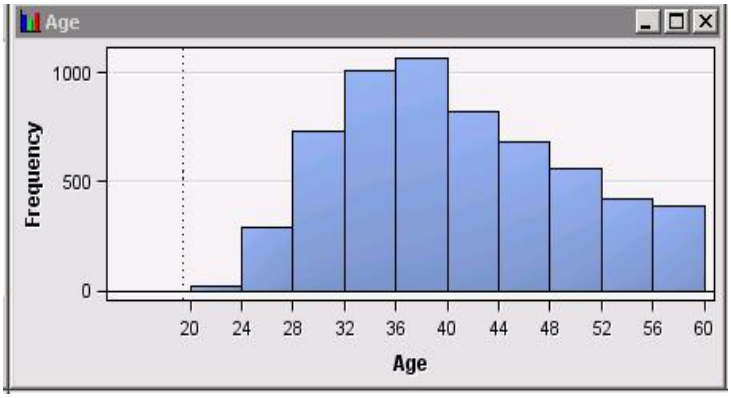


Figure 1

After using the StatExplore node it has been seen that the four variables have consisted zero missing values so missing value handling is not required. One interesting thing to note here is the variable worth for the given dataset of the selected variables with respect to the target variable. Age is the worthiest variable and Marital_Status is the least worthy among all four but still the worthiness of Age with respect to the target variable is only 0.006.

Key demographics segments profiles:

After adopting the 5-7 number of clusters we found that 5 clusters are the best representative of the clustering segmentation. Whose profiling is described below.

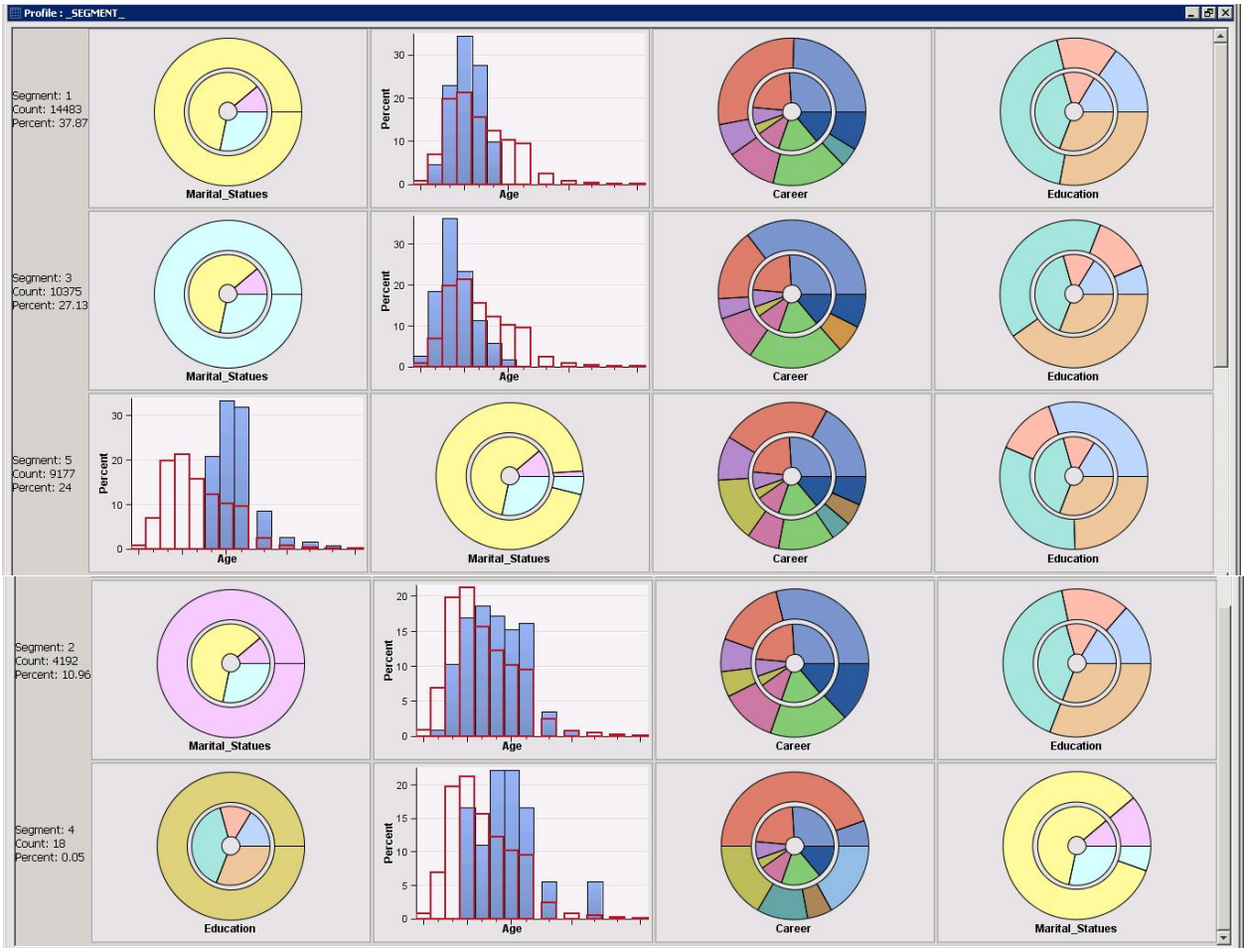


Figure 2

The above identified segments have been mapped to the segments in Australian community based on the Roy Morgan values segment.

Segment 1 – (Conventional family life)

This segment contains people who have age distribution less than normal distribution which states youngsters and who are married, who have more than usual careers in Blue Collar jobs and almost half of the population studied secondary education. Because of their age and marital status, they represent the conventional family life segment of Roy Morgan. Usually they tend to have worried about their children's future and paying mortgage for the house they bought results in seeking greater financial securities. This segment contains 37.87% of all the customers given in the dataset.

Segment 3 – (Young Optimism)

This segment contains people who are youngsters and are not married which states that they might be still studying or have just started their professional careers. The segment has more customers who are in their universities and more customers have administration jobs than the usual population. Based on their demographics this segment tends to map with the Young Optimism segment of the Roy Morgan segments. They might be in the university now, but they are planning to go overseas for the career advancement. Having new experiences and creating the right image of themselves is the primary goal for this segment. This segment contains 27.13% of the entire customers.

Segment 5 – (Traditional family life)

This segment contains people who are middle aged to slightly older aged and are mostly married. This segment contains people from all the fields and have studies from primary school to university education. Due to the age and marital status of the customers the segment represents the Traditional family life of the Roy Morgan segments. This segment is a counter part of the conventional family life which holds same values in terms of security and giving better to the families of theirs, but nowadays the life stays empty-nested in the customer of this segments. This segment contains about 24% of the entire customers.

Segment 2 – (A fairer deal)

This segment contains only divorced people of all ages but have studied something in their lifetime unlike the divorced people in the final segment. This segment contains people with all kinds of education from primary to university. Based on their demographics they tend to have map with A fairer deal segment of the Roy Morgan segments. This segment people mostly think that the life has not treated fairly to them because of the struggles they are facing right now in terms of financial and mental. They tend to have escape from the with extra activities to get rid of their frustration. This segment contains 10.96% of the entire customers.

Segment 4 – (Basic needs)

This segment contains customers who are all illiterate and have middle to old age. Despite of no education they seem to have blue collar, self-employed and retired as their current career status. This segment represents the Basic needs Roy Morgan segment value. They seem to enjoy what they are doing now despite of living on securities and pensions they are happy of what they have and do not look for more now. This segment has count of only 18 which represents only 0.05% of the dataset.

Most important variables based on each segment:

Based on the variables worth derived from the segment profiles we have below table generated where variables worth more than 0.1 or is most worthy in the segment have been written below. The worthiness of variables shows the discriminant capabilities of the variables where below are the most discriminant/important variables among all.

Segment	Important Variables
1	Marital_Status, Age
2	Marital_Status
3	Marital_Status, Age

4	Education
5	Age

Differences between Demographic Segment of subscribed and non-subscribed customers:

Initial looking at the subscribed and non-subscribed customers gave almost similar customer segmentations as the whole dataset segmentation.

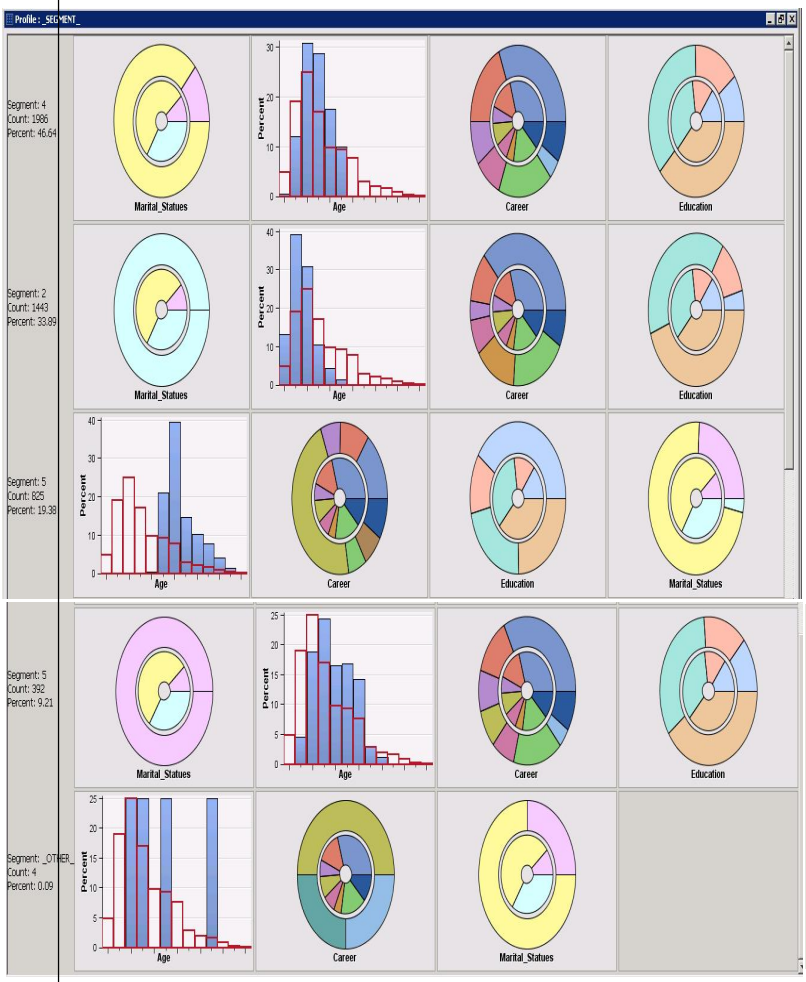


Figure 4 Subscribed

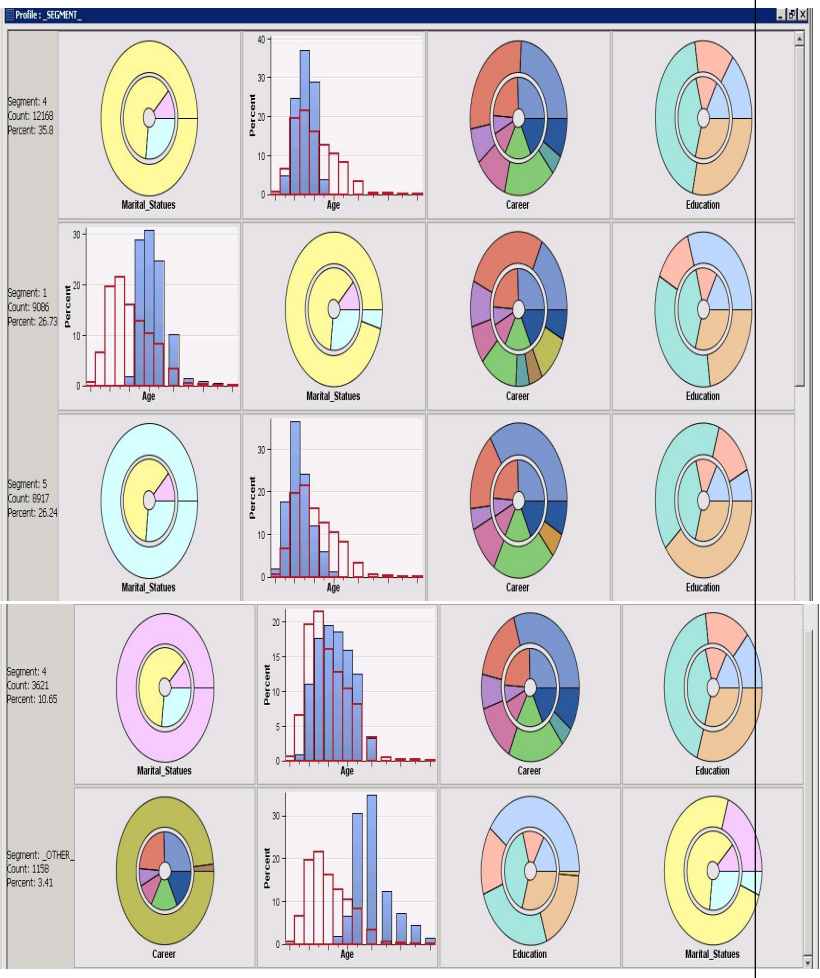


Figure 3 Non-Subscribed

Based on the above graphs we can see that conventional family life segment which is segment 2 in subscribed group and segment 5 in non-subscribed group is same in both the groups with slightly age difference between them reported but the percentage distribution of yes and no shows that this segment is more likely to subscribe to the long-term deposit plan of the bank.

In young optimism segment which is segment 5 in subscribed and segment 2 in non-subscribed is almost identical in both the groups of subscribed and non-subscribed. And the percentage distribution of these group in both the segments shows that they are more likely to subscribe to the long-term plan.

But in the traditional life segment which is segment 3 in subscribed group and segment 6 in non-subscribed group we can see that customers who are retired now in that segment are more likely to subscribe to the long-term deposit plan of the bank. Also, in the final segments of both the groups which are Basic needs segments considered to have subscribed to the long-term deposit plan if they are housemaid only which is unlike in the retired customers of that group. These three segments have been seen to have not subscribe more likely than the above two segments.

Behavioral based segmentation

By setting Subscribed as a target variable and all the behavioral variables as input variables we have derived the 5 clusters as below.

Data Exploration:

Any variables in the behavioral dataset are not of a continuous. So, no transformation of variables is needed as clustering handles other type of transformation itself. After using the StatExplore variable in the dataset we see that not a single variable here consists missing values. So, missing value handling is also not needed.

Key behavioral Segments:

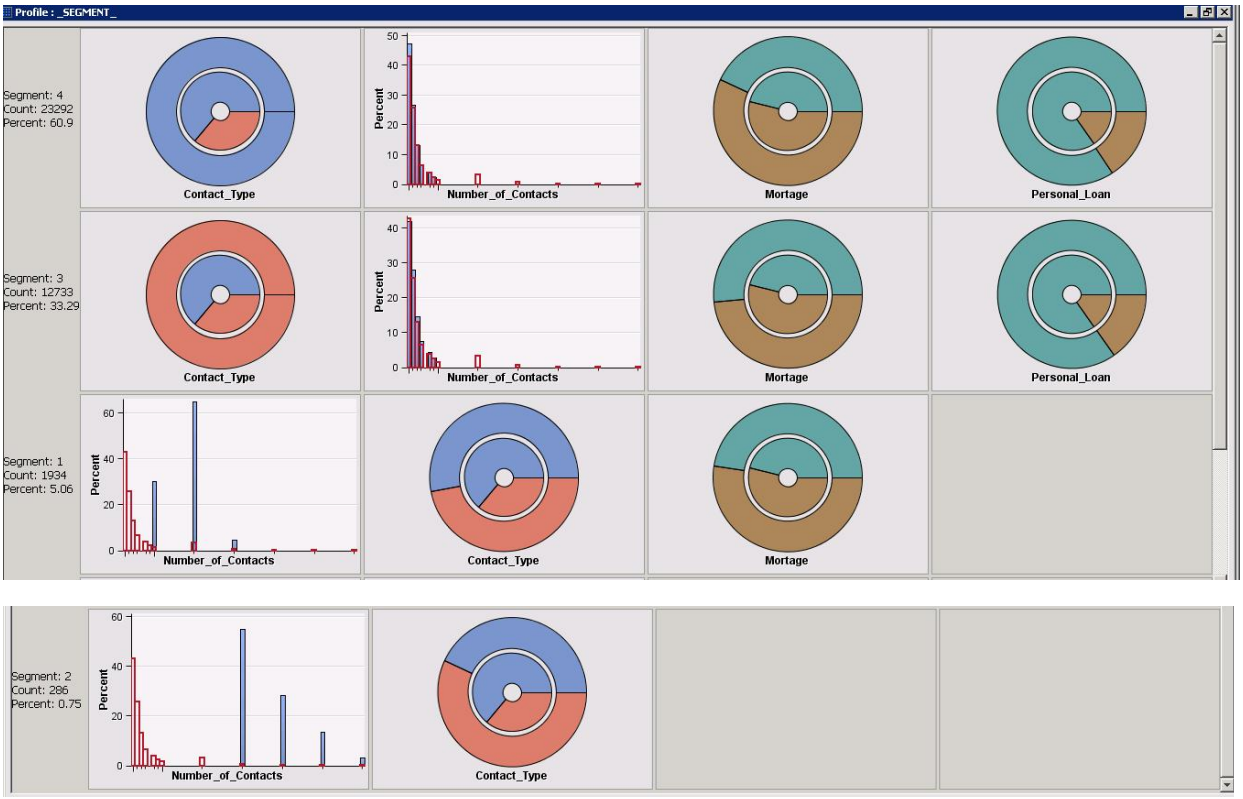


Figure 5

After adopting the 4 number of clusters above 4 clusters have been identified whose profiling is described below.

Segment – 4 (Cellular)

This segment contains variables who prefers contacted via cellular phones and typically responds in smaller number of contacts only. This segment customers have been seen having mortgage and personal loan in some customers. This segment contains 60.9% values from all the customer segments.

Segment – 3 (Telephone)

This segment is typically similar to the above segment customers but instead of having contacted via cellular they preferred to have contacted by telephones. This segment contains 33.29% values from all the customer segments.

Segment – 1 (Number of contact - Medium)

This segment contains values of all the customers who gave response in typically after 6 times to 12 times contacted regarding the plan of the bank. This segment contains only 5.05% of the values from the entire customer segments.

Segment – 2 (Number of contact - More)

This segment contains values of all the customers who gave response in typically more than 21 times contacted regarding the plan of the bank. This segment contains only 0.75% of the values from the entire customer segments.

Most important variables based on each segment:

Based on the variables worth derived from the segment profiles we have below table generated where variables worth more than 0.1 or is most worthy in the segment have been written below. The worthiness of variables shows the discriminant capabilities of the variables where below are the most discriminant/important variables among all.

Segment	Important Variables
4	Contact Type
3	Contact Type
2	Number of Contacts
1	Number of Contacts

Differences between Demographic Segment of subscribed and non-subscribed customers:

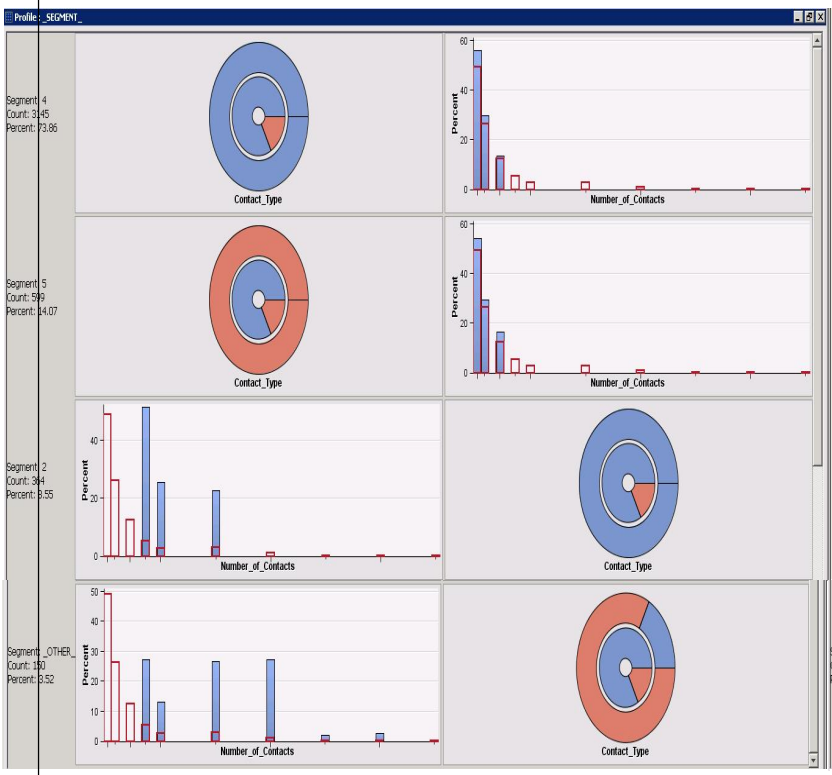


Figure 7 Subscribed

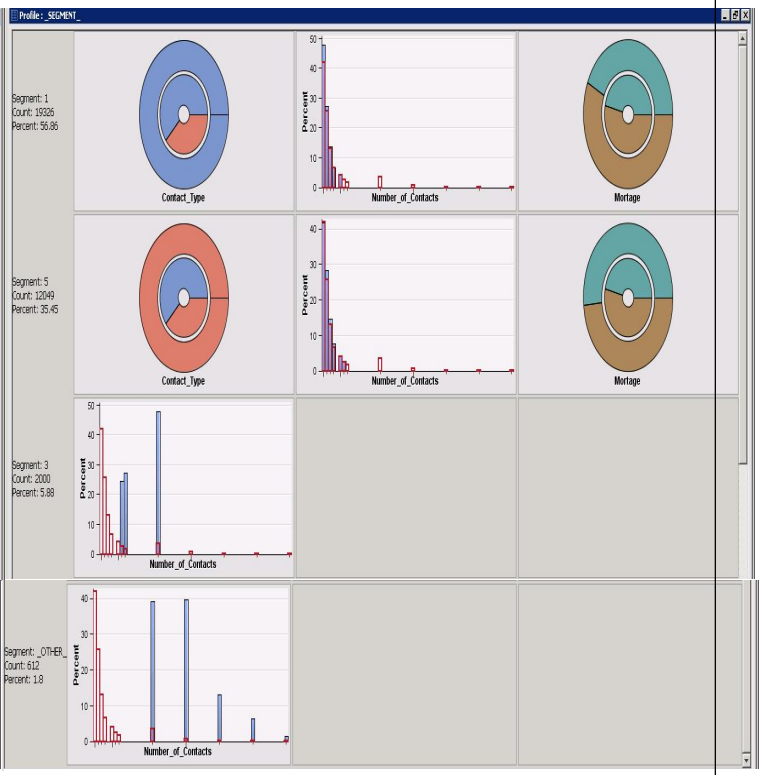


Figure 6 Non-Subscribed

Based on the above two graphs it can be seen that the first two clusters are almost identical to each other. Only in the last two segments we have seen the difference which states that number of contacted is more in the non-subscriber's group than the subscriber's one.

Also, the percentage of population distribution is different in both the groups as there are 73.86% of subscribers belongs to the group – 1 while only 56.86% non-subscribers belongs to the group-1. While reverse flow has been seen in the second group as there are only 14.07% subscribers present with respect to 35.85% non-subscribers in the same group.

Preparing data for the cross-clustering task:

Using the save data node of the SAS enterprise miner we have saved whole clustered data of Behavioral and Demographics dataset in the csv file format.

Using table function of R, we have done the cross-cluster analysis by setting demographics cluster as rows and behavioral clusters as columns for both subscribed and non-subscribed groups.

Associations between two groups:

	1	2	3	4
1	0.0190090	0.0025624	0.1425284	0.2145901
2	0.0056478	0.0007844	0.0381226	0.0650543
3	0.0135965	0.0019087	0.0747026	0.1810694
4	0.0000261	0.0000000	0.0000784	0.0003661
5	0.0122892	0.0022225	0.0775003	0.1479409

Figure 9

	1	2	3	4
1	727	98	5451	8207
2	216	30	1458	2488
3	520	73	2857	6925
4	1	0	3	14
5	470	85	2964	5658

Figure 8

After doing the cross-cluster analysis we have identified two types of table described as above. Figure-9 is the probability distribution and Figure-8 is the frequency table for the cross-clusters. The frequency table shows that cluster 4 in the demographics has frequency of only 18 values which is so much less as compared to the entire datasets. And in the behavioral one we can see that cluster 1 and 2 has so much less frequency than the other two. So, they are not of much value as they represent small percent of entire population.

After getting the probability of each associations from in the cross-clusters we can see that segment-1 from demographic dataset highly associated with the segment-3 and segment-4 of the behavioral dataset. Segment-3 from demographic dataset is highly associated with the segment-4 of the behavioral dataset and segment-5 from demographic dataset is highly associated with the segment-4 of the behavioral dataset.

Highly associated combined segments:

1.) After seeing the characteristics of the identified highly associated segments, we have first segment derived which involves the below characteristics:
   - i.) Marital status – Married
   - ii.) Age – Youngsters
   - iii.) Education – Mostly Secondary
   - iv.) Number of contacts – Less than 6
   - v.) Contact Type – Any (Because clusters 3 and 4 from behavioral dataset are highly correlated with the segment – 1)

2.) The characteristics of the second segment gives the second combined segments with the following characteristics:
   - i.) Marital status – Single
   - ii.) Age – Youngsters
   - iii.) Education – All kinds of from primary to university
   - iv.) Contact type – Cellular
   - v.) Number of contacts – less than 6

vi.) Career - Administration

3.) Seen from the characteristics of third type of combined segments we have below conclusion derived:
   i.) Marital status – Married
   ii.) Age – Middle age
   iii.) Education – From Primary to University
   iv.) Contact type – Cellular
   v.) Number of contacts – less than 6

Relationship between combined segment and outcome:

To identify the relationship between this two we have calculated the lift of the combined segments with the below formula using R. Lift gives the percentage of customers who are likely to subscribe in the combined segments.

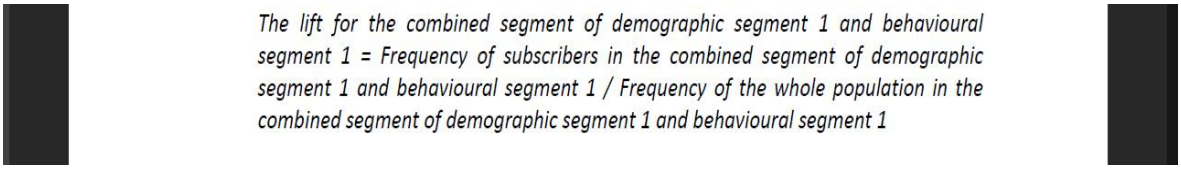
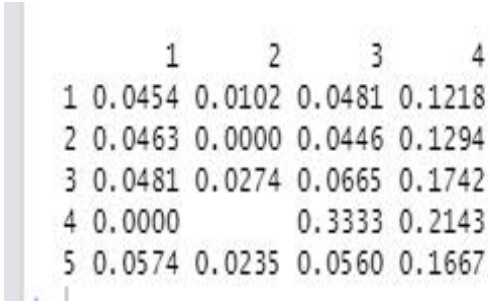


Figure 10

After calculating lift in R we have below results identified.



	1	2	3	4
1	0.0454	0.0102	0.0481	0.1218
2	0.0463	0.0000	0.0446	0.1294
3	0.0481	0.0274	0.0665	0.1742
4	0.0000		0.3333	0.2143
5	0.0574	0.0235	0.0560	0.1667

Figure 11

Based on lift whoever has the most lift value is the most ideal segments which have positive outcomes with respect to the subscribed. We can see that segment-4 from behavioral dataset has most important worth related to all the segments. Segment – 4 in demographic and segment – 3 in behavioral dataset is the worthiest among all.

Based on the above result and result of cross-cluster analysis we conclude that there is some kind of relationship between the outcome variable and combined segments as there are almost equal clusters identified. Only difference is segment-1 from demographic clusters and segment-3 from behavioral clusters identified using cross-cluster analysis does not have more ratio of subscribers to non-subscribers so that segment may not be worthy for the future references.

In the lift part Segment-4 from behavioral and segment – 2 and 4 from the demographic segments have been identified as ideal segments which are not identified by the cross-cluster analysis. After further analysis it can be seen that the frequency of segment – 4 of demographic and behavioral is

not much so having less frequency gives the lift value to be more only after having 1 subscribed customer in it. Same trend has been seen in the segment – 4 (Demographics) and segment – 3 (Behavioral) as there is only one customer among three who has subscribed. So, having only 3 customers in these segments is not of much worth.

Key segments related to the outcome variable:

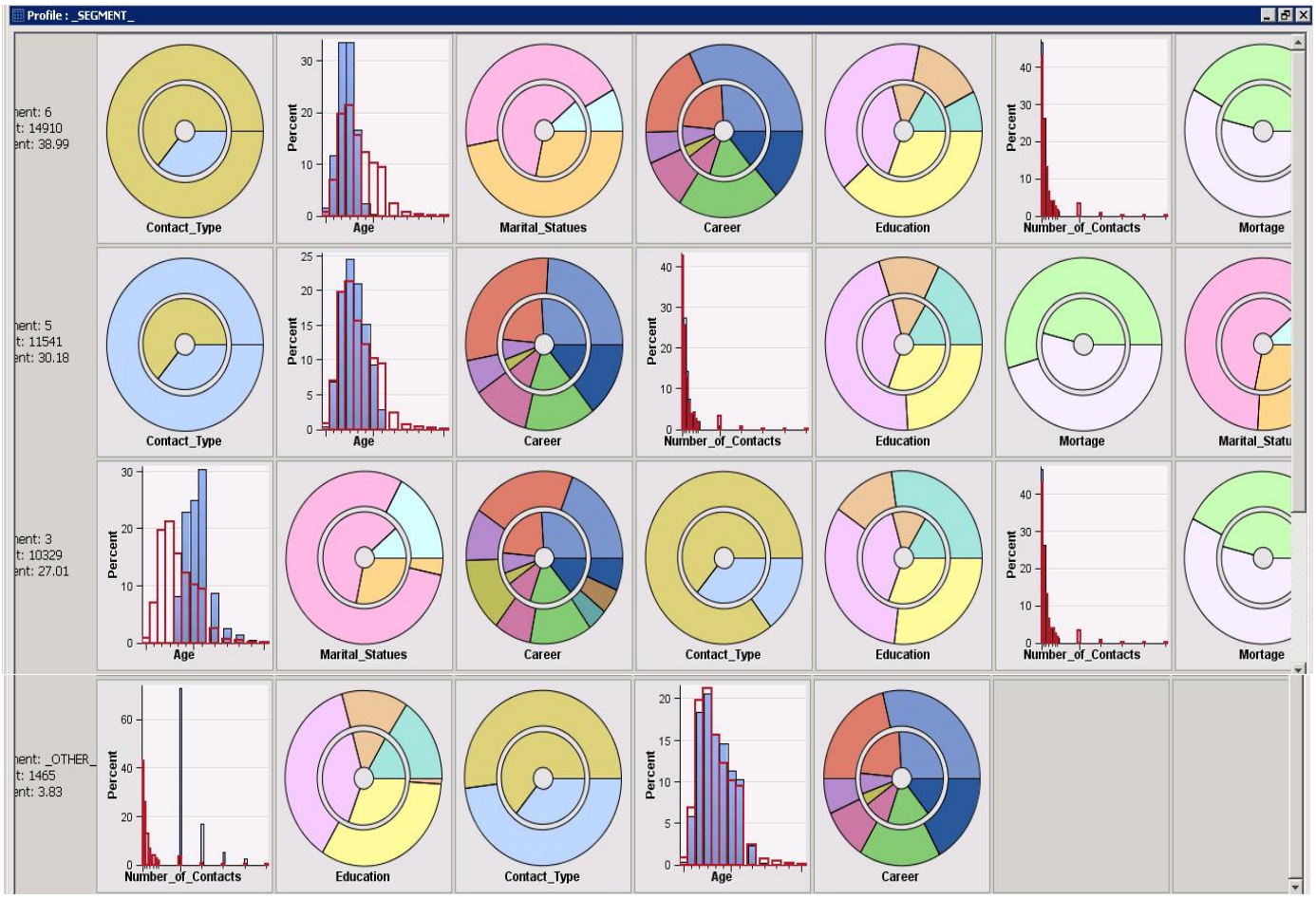


Figure 12

Based on the above graph 4 main types of clusters have been identified whose profiling is described below.

Segment 6

This segment contains customers who prefers to have contacted via the cellular phones. They are the youngster with mix of married and single customers with few young aged divorcees and responds in less than 6 times contacted related to the plan. Education of this group's customers is from primary to university. This segment represents 39% of the entire population.

Segment 5

This segment contains customers who prefers to have contacted via telephones. They are the youngsters with few of them young to middle age. This segment contains customers who have blue collar job more than the population distribution one with education of secondary school and generally responds within contacted 6 times. This segment represents 30.18% of the entire population.

Segment 3

This segment contains customers who are of middle to old age and mostly married. They usually prefer to have contacted via the cellular rather than telephone and responds within contacted 6 times related to the long-term deposit plan of the bank. This segment represents 27.01% of the entire population.

Segment 1,2 and 4

This segment contains customers who responds after contacted more than 12 times. That is the only unique identity of this segment as all the other variables in this segment tends to represent the entire population itself. This segment represents 3.83% of the entire population.

Most important variables based on each segment:

Based on the variables worth derived from the segment profiles we have below table generated where variables worth more than 0.1 or is most worthy in the segment have been written below. The worthiness of variables shows the discriminant capabilities of the variables where below are the most discriminant/important variables among all.

Segment	Important Variables
6	Contact Type, Age
5	Contact Type
3	Age
1,2 and 4	Number of Contacts

Conclusion

Based on the above results we have arrived at the conclusion that doing separate clustering first with respect to the target variable and then conducting the cross-clustering does not tends to have increased the clustering results. So, instead of conducting separately, combined segmentation can also be done which represents the final answer.

Based on the lift calculation and the cross-cluster and combined segments generation we arrive at the conclusion that 3 customer segments identified in the task 3 are the best representative of the entire population and also best representative of the prediction about which customer segments most likely to be subscribe to the long-term deposit plan of the Delta bank.