Customer Segmentation and Profiling

Using SAS Enterprise Miner and Rstudio

By Sazee S. (Maninderpreet Singh Puri)

Table of Contents

1. Introduction- Case Study

The Case Study is based on the use of predictive analytics for future bank to identify target customers who are most likely to subscribe long-term deposits. The customer analytics project is using existing customer profile and marketing campaign data. The dataset we are using consists of eight variables; '*Age', 'Default_Credit', 'Education', 'Housing_Loan', 'Job', 'Maritial_Status', 'Personal_Loan',* and *'Subscribed'* (with *'Subscribed'* being the target variable) and consists of 30477 observations. We are using clustering-based segmentation and profiling in SAS Enterprise miner to find segments seeking knowledge and insights relating to:

- The demographics-based segments and their profiles.
- The representative behavioural profiles for each segment.
- How the produced segments can be mapped to a broader concept of segments in Australian community.

We are defining segments in conjunction with Roy Morgan value segments which represent market segments based on consumer behaviour in Australian community. Also, we are performing a cross cluster analysis – demographics to behavioural segments using R Studio and SAS Enterprise miner to identify the common key segments (with customers subscribe for long-term deposits).

The first step in a data analytics task is data pre-processing. The data used in our analysis is free from missing values. We have only one variable *'Age'* with Level as *Interval,* and it is not skewed (I checked the Skewness is 0.98 and Kurtosis is 1.24 which is close to zero, so it is not skewed). Rest all variables are binary (0 and 1) and Nominal. So, we don't need any transformations for our variables. I added the customer index (the row number) using the '*Transform Variables'* node so that we identify the customers uniquely and perform cross cluster analysis in Task 3. Also, I set *'Subscribed'* to *Role* as *Target*.

2. Customer segmentation based on demographics data

In this Task, I used '*Age', 'Job', 'Marital_Status',* and *'Education'* and set all other variable to *Use* as *No (*in both *Cluster* and *Segment Profile* nodes*,* except *_SEGMENT_* and *_SEGMENT_LABEL_* which will be used in cross cluster analysis*)* as we are only trying to segment the data based on these demographic variables. In the *Cluster* node I tried the *Maximum Number of Clusters with* 5/6/7 cluster size to get the best cluster size for segmentation. I choose seven clusters for our study as all clusters had unique combination of variables. Also, these clusters map well to the Roy Morgan segments.

Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster ▼	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster
0.318479	0.01027	.	6	8517	0.327226	2.898886	7	1.753329
0.318479	0.01027	.	7	7616	0.286808	2.530734	4	1.518442
0.318479	0.01027	.	4	4447	0.319351	2.830218	7	1.518442
0.318479	0.01027	.	3	3322	0.36644	2.937067	4	2.25367
0.318479	0.01027	.	5	2759	0.283855	3.142851	7	1.794035
0.318479	0.01027	.	2	2660	0.301593	2.789254	4	1.900863
0.318479	0.01027	.	1	1156	0.404859	3.55505	4	2.292188

Figure 1 Cluster Node results

Another factor which helped me decide the number of clusters was the '*Frequency of Cluster'* column in the *Mean Statistics panel* in the *Results* window of *Cluster* node. The Frequency of clusters show that all cluster had a significant number of observations. Thus, it was worth exploring seven clusters to find the differences in terms of demographics amongst them.

In the *Segment Profile* node Results window, we can see four windows presenting information on seven segments. We are interested to check *Profile* window for detailed information on segments. We can see the segments visual shows pie char for nominal variables (*Job, Education* and *Marital_Status*) and bar plot for continuous variable (*Age*). I set the *Minimum Worth* in *Segment Profile* node to 0.00010 to get all variables in the *Profile* panel in Results. This

will help giving more insights into the differences of variables amongst all segments. I have discussed the key demographic segments which stand apart in whole dataset in terms of the attributes they reflect.

2.1 Segments

a. Segment 1:

Segment one consists of 1156 observations (3.79% of the whole dataset) in. The pie chart for *Job* variable shows that ~94% of the people are *retired* and ~4% of the people in this segment are *housemaid.* This shows that most people are elderly and are not very well educated in this segment. We can verify this information from the bar chart of *Age* variable. We can see that ~31% of the people in this segment are aged 55, ~26% are aged 62, ~20% are aged 70, and ~23% are aged above 77. In the pie chart for *Education,* we can see that ~44% of the people have *Primary_Education* in this segment. ~23% people have *Secondary_Education*, ~18% have *Tertiaty_Education* and only ~15% have *Profesional_Education.* This explains the fact that why some people are working as *housemaid*, as they might not be rich and might still need to work to meet their expanses. The pie chart for *Marital_Status shows ~75%* of the people in this segment are *married*. Approximately 20% are *divorced* and ~5% are *single.* In this segment, we can see that most of the people are retired and are in older age. Most are married, while some are divorced which relates to strong family ties whether living together with their partner in marriage or living with kids and grandkids after a divorce. We can map this segment to *'Basic Need'* and *'Traditional Family Life'* in Roy Morgan values segments, as both these segments illustrates; *basic needs, simple life, live independent, stay healthy* and *enjoy retirement* attributes.

b. Segment 5:

Segment five consists of 2759 observations (9.05 percent of the whole population). We can see from the pie chart for *Education,* that all people in this segment have *Professional* education. This means all people in this segment are professionals who are paid well. In the pie chart for *Job,* we can see that ~69% of people are *technician,* ~15% of people are in *other* jobs, ~9% are in *blue-collar* jobs, and ~6% are in *admin* jobs. This is aligned to the information we got from *Education* pie chart. In the pie chart for *Marital_Status,* we can see that the ~84% of the people in this category are *married* and ~16% are *single.* This tells that most of the people are family oriented in this segment. The bar chart for *Age* shows most people in this segment are in their 30's and are younger than other segments. We can map this segment to *'Young Optimistic'* in Roy Morgan values segments, as this segment illustrates; *innovative and interested in technology, planning careers, and thinking about the future* attributes.

c. Segment 7:

Segment seven consists of 7616 observations (24.99% of the whole population). We can see in the pie chart for *Marital_Status* that all people are *married* in this segment. In the bar chart for *Age,* we can see that most people are in 30s in this segment. In the pie chart for *Job,* we can see that ~35% are in *admin* jobs, ~19% are in *blue-collar* jobs, ~12.5% are in *services*, ~12% are in *other* jobs, ~11.5% are in *technician* jobs, ~9.5% are in *management* jobs. This shows most of the people are in stable jobs. In the pie chart for *Education,* we can see that ~57% people have *Secondary_Education* and ~44% people have *Tertiary_Education*. This shows most of the people are well educated in this segment. We can map this segment to *'Conventional Family life'* in Roy Morgan values segments, as this segment illustrates; *to express love and affection to all family members, to satisfy household needs and have good food, to help their kids be safe, smart, and successful, people seeking greater financial security, struggling to improve their basic living standards and give their families better opportunities* attributes.

Even though I have discussed only a few Segments, for our study I have mapped all the segments discovered in SAS to Roy Morgan Segments based on their attributes as follows:

- Segment 1 -> Basic Need/Traditional Family life
- Segment 2 -> A Fair Dealer
- Segment 3 -> Something better
- Segment 4 -> Visible Achievement/Socially aware/Real Conservatism
- Segment 5 -> Young Optimistic

- Segment 6 -> Look at me
- Segment 7 -> Conventional Family life

2.2 Important variables in each segment

All segments have their unique set of combinations of attributes which set them apart from each other. Each segment is unique due to a certain attribute having a major difference from the whole data and this level of uniqueness is important for that segment. In the *Results* window of *Segment Profile* node, we can check these variables in *Variable Worth* panel, in the order of their worth in setting the segment apart from others. The segments list below are order (decreasing) in how uniquely they can be identified in the whole dataset and show the worth of variables:

Segment 6: Marital_Status (0.348) -> Age (0.107) -> Job (0.033) -> Education (0.018)

Segment 7: Marital_Status (0.092) -> Age (0.070) -> Education (0.047) -> Job (0.014)

Segment 4: Age (0.093) -> Marital_Status (0.023) -> Education (0.014) -> Job (0.011)

Segment 3: Marital_Status (0.180) -> Age (0.007) -> Job (0.001) -> Education (0.0004)

Segment 5: Education (0.096) -> Job (0.029) -> Marital_Status (0.005) -> Age (0.003)

Segment 2: Education (0.108) -> Job (0.028) -> Marital_Status (0.007) -> Age (0.003)

Segment 1: Job (0.062) -> Age (0.043) -> Education (0.0027) -> Marital_Status (0.000940.108

As we can see that the segment six is most unique as it has the most distinct attributes and segment one is the least distinctive in the whole data set.

To find more about the relation of variable worth in our segments, we can map these finding to the segment descriptions we found in the past. As we have determined segment one is linked to the *'Basic Need'* and *'Traditional Family Life'* in Roy Morgan values segments. In terms of the importance of variables we can see *Job* is the most important valuable for segment one. This aligns with the '*Basic Needs'* as the value that sets them apart from other segments is that they are mostly *retired.* So, this correct and confirms our previous findings. We can compare the segment 5 *'Young Optimistic'* in Roy Morgan values segments to variable worth *Education* that is on top. This is because most youngsters are focused on *Education* and are finding *Job* which sets them apart from other segments. So, variable worth tells us which variables are responsible for reflecting the demographic values of that segment, for instance whether they are young, and jobs oriented, or they are old, and mostly retired.

2.3 Long-term deposit: subscribers v/s non-subscribers

To split the data into Subscribers and non-subscribers for long-term deposits. I filtered the data using the *Filter* node into subscribed =yes and subscribed =no. Under the *Class Variables* in properties of *Filter* node, I set the *Default Filtering Method* to *None, Normalize Values* to *No,* and under *Interval Variables* I set the *Default Filtering Method* to *None*. I applied this setting in *Filter* nodes used in the behavioural based segmentation as well. I applied clustering and segment profiling on both subsets using same setting we used for the whole dataset (Number of clusters = 7) except Minimum Worth which I set as 0.00001 to get more detailed information in *Profile* panel. In segments discovered for both Subscribed as Yes and No data set, we can notice there are major differences. All segments are different, but some segments partially overlap with each other. These segments show key differences in one or two deciding variables on what makes them subscribe to long-term deposits or not. We are interested in knowing about these key differences in the partially overlapping segments. I have discussed key segments which partially overlap and have clear differences between them.

a. Segment 1 of *Subscribed* as *No* V/S segment 2 of *Subscribed* as *Yes*

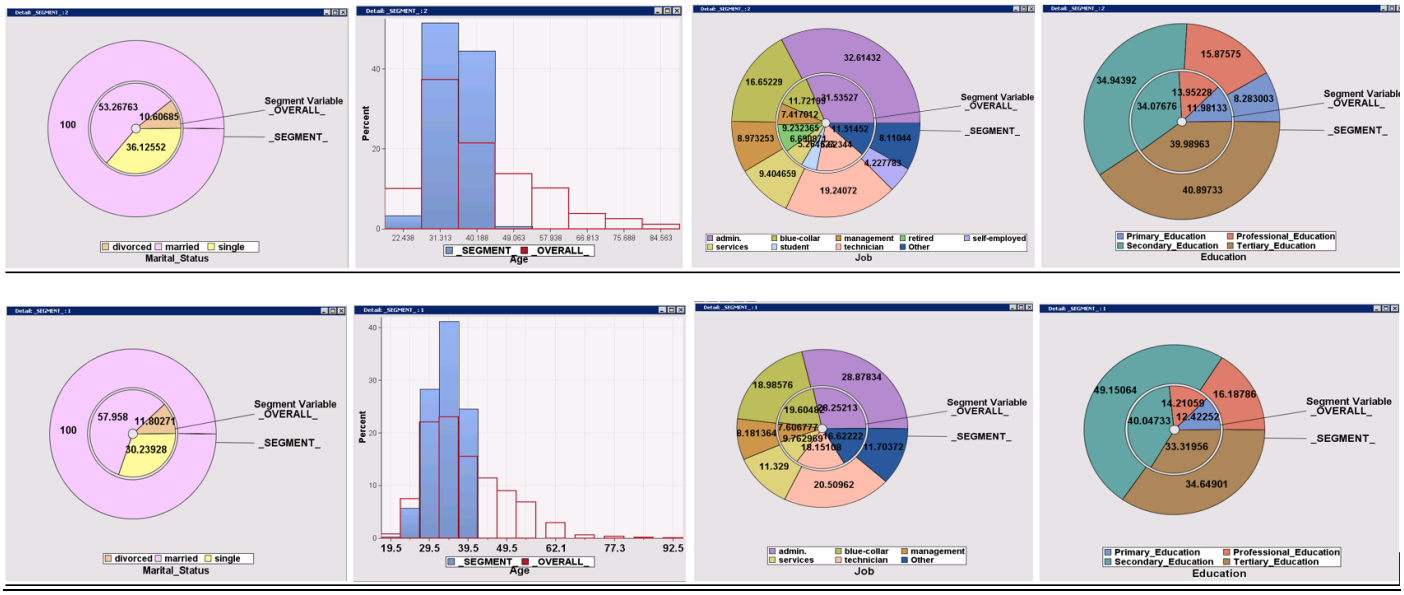

Figure 2 Segment 1 of Subscribed as No (bottom) V/S segment 2 of Subscribed as Yes (top)

We can see in the *Profile* Panel, that Segment 1 of *Subscribed* as *No* and segment 2 of *Subscribed* as *Yes* share similarity in *Marital_Status (married)* and *Age (23-40)*. But the key differences are in *Education* and *Job* attributes. We can see that in segment 2 (*Subscribed* as *Yes)* ~41% people have *Tertiary_Education,* ~35% have *Secondary_Education* and ~16% have *Professional_Education.* However, in segement 1 (*Subscribed* as *No*) ~49% people have *Secondary_Eduaction*, ~35% have *Tertiary_Education,* ~16% have *Professional_Education.* This difference tells us about the people in Age group of 27-40 who are married and have higher education are more likely to subscribe for long-terms deposits than who are less educated. We can verify our findings from the pie chart for *Job* in segment 2 (*Subscribed* as *Yes)* and segment 1 (*Subscribed* as *No*). We can see that in segment 2, ~32.5% people are in *admin*, ~19% are in *technician*, ~16.5% are in *blue-collar*, ~9.5% are in *services*, ~9% are in *management,* ~8% are in *other,* and ~4% are in *self-employed* jobs. However, segment 1 shows ~29% people are in *admin*, ~20.5% are in *technician*, ~19% are in *blue-collar*, ~11% are in *services*, ~8% are in *management, and* ~11.5% are in *other* jobs. It shows that segment 2 with *Subscribed* as *Yes* have more people working as *admin and management* jobs as compared to segment 1. This observation confirms that fact that people in this age group who are married, more educated and are into better paying jobs are more likely to subscribe to long-term deposits.

b. <u>Segment 4 of *Subscribed* as *No* and segment 4 of *Subscribed* as *Yes*</u>

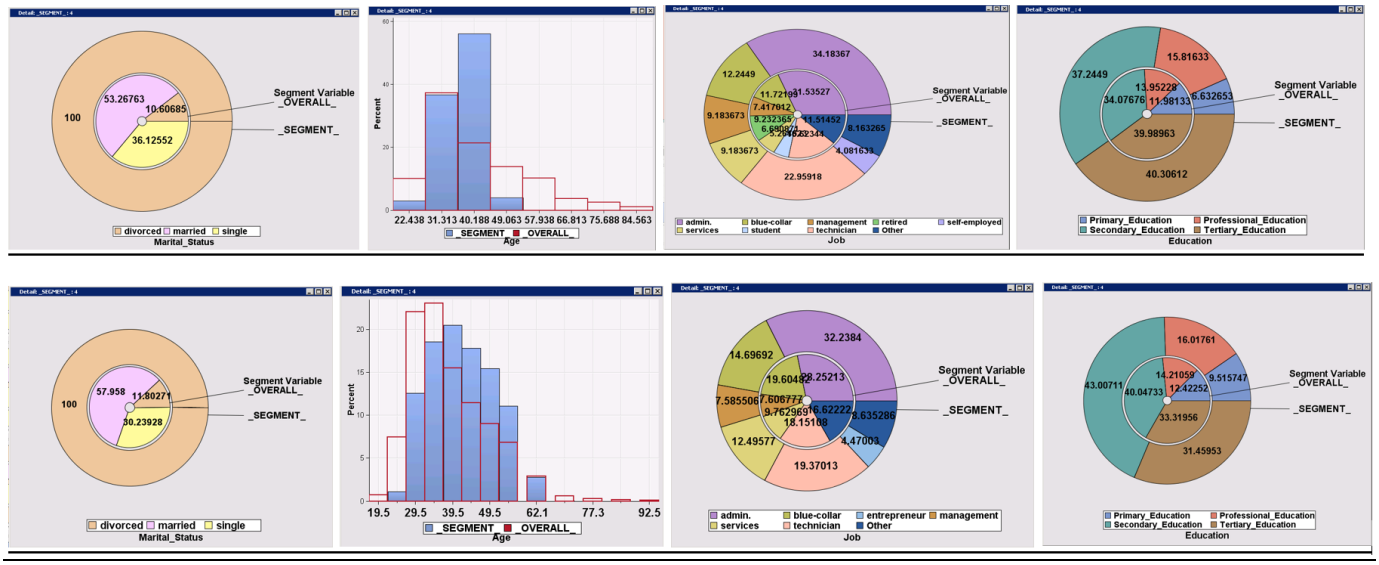

Figure 3 Segment 4 of Subscribed as No (bottom) and segment 4 of Subscribed as Yes (top)

We can see in the *Profile* Panel, that Segment 4 of *Subscribed* as *No* and segment 4 of *Subscribed* as *Yes* share similarity in *Marital_Status (married)* and *Age (23-63)*. But the key differences are in *Education* and *Job* attributes. We can see in segment 2 (*Subscribed* as *Yes)* ~34% are in *admin* jobs, ~23% are in *technician*, ~12% are in *blue-collar* jobs, ~9% are in *management,* ~9% are in *services,* ~8% are in *other* jobs, and ~4% are *self-employed.* However, in segment 2 (*Subscribed* as *No*), we can see ~32% are in *admin* jobs, ~19% are in *technician* jobs, ~14.5% are in *blue-collar* jobs, ~12.5% are in *services,* ~8.5% are in *other* jobs, ~7.5% are in *management* jobs, and ~4.5 % are in *entrepreneur* jobs. This tells that most of the people in segment 2 (*Subscribed = Yes*) are into *admin* and *technician* jobs, which means they have more money than people in segment 2 (*Subscribed = No*) as more people are into blue collar jobs. So, people in segment 2 (*Subscribed = Yes*) are more likely to subscribe to long-term deposit. We can confirm this information from the *Education* pie chart in both segments. We can see that in segment 2 (*Subscribed = Yes*) more people are into Tertiary_Education (~40%), whereas in segment 2 (*Subscribed = No*) more people have Secondary_Education (~43%). This points to higher literacy rate in segment which choose to subscribe.

3. Customer segmentation based on behavioural data

In this Task, I used *'Default_Credit'*, *'Housing_Loan',* and *'Personal_Loan'* and set all other variable to *Use* as *No (*in both *Cluster* and *Segment Profile* nodes as we are only trying to segment the data based on these behavioural variables. I tried five clusters for our study. I noticed that there are significant number of observations in four clusters, except one cluster which has only three observations. I changed the *Maximum Number of Clusters* to four and still got significant number of observations in three clusters, except the fourth one which had three observations. This shows that cluster with three observations is an outlier and has its own attributes. As the number of observations in this cluster are very small, we can ignore this cluster in our study. For the final analysis, I selected five clusters for my study and set the *Minimum Worth* in *Segment Profile* node to 0.00010 to get all variables in the *Profile* panel in Results. This will help giving more insights into the differences of variables amongst all segments.

3.1 Segments

a. Segment 1

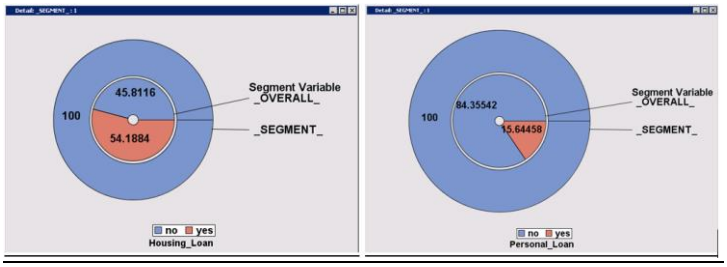

Figure 4 Customer segmentation based on behavioral data: Segment 1

Segment one consists of 12035 observations (39.49% of the whole population). We can see in this segment that people don't have *Housing_Loan* and *Personal_Loan.* This means people in this segment are less financially burdened. They might be in very young age as most youngsters are still studying and don't have many liabilities. They can also belong to the old age group as most aged people have paid off their liabilities.

b. Segment 2

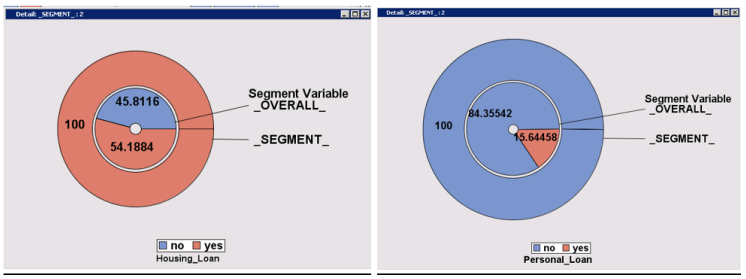

Figure 5 Customer segmentation based on behavioral data: Segment 2

Segment two consists of 13671 observations (44.86% of the whole population). We can see that in this segment people have *Housing_Loan* but they don't have any *Personal_Loan*. This shows that people in this segment are maybe in their 30-50's where they are paying for a house but don't have any personal loans to pay off.

c. Segment 3

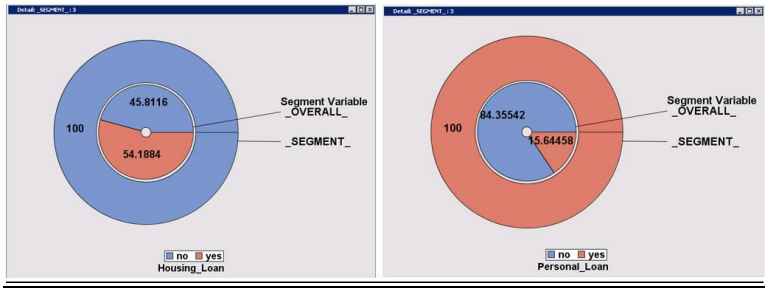

Figure 6 Customer segmentation based on behavioral data: Segment 3

Segment three consists of 1925 observations (6.32% of the whole population). We can see that in this segment that people have *Personal_Loan* and don't have *Housing_Loan.* This shows that people in this segment are either youngsters who might have finished studies and had a Personal_Loan for education or buying their first car or it can be people in their later stages of life where they might have paid off their housing loan but are buying a business or buying expensive items for their house.

d. Segment 4

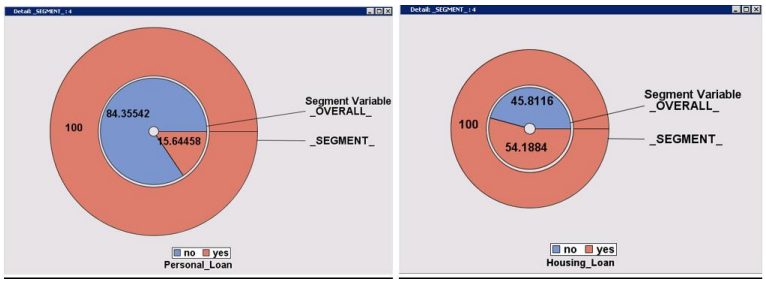

Figure 7 Customer segmentation based on behavioral data: Segment 4

Segment four consists of 2843 observations (9.33% of the whole population). We can see that people in this segment have both *Hosuing _Loan* and *Personal_Loan.* This means people in this group are financially burdend and have more liabilities to pay than ther segements. These are people who have just started their families and are in their early 30's.

e. Segment 5

Even though we are not considering this segment in our study as number of observations are small, I tried to look into what are the attrbiutes of this segment.

```
Variable: _SEGMENT_ Segment: 5 Count: 3
Decision Tree Importance Profiles

  Variable          Worth       Rank

Personal_Loan     3.7141E-9        1
Housing_Loan      3.4924E-9        2
```

Figure 8 Customer segmentation based on behavioral data: Output panel of Segments node

From the *Output* panel of *Segments* node I see that there are 3 observation and *Personal_Loan* is of more worth than *Housing_Loan.* I checked all the data attributes for this segment from Properties-> Exported Data-> TRAIN. I

sorted the *Segments* column in decreasing order to get the segment five on top.

Obs #	Age	Job	Marital_Status	Education	Default_Credit	Housing_Loan	Personal_Loan	Subscribed	ID	Segmen... ▼	Distance	Segment Description
9286	48	technician	married	Professional_Education	ye	yes	no	no	14330	5	1.338014	Cluster5
10969	31	unemployed	married	Secondary_Education	ye	no	no	no	16874	5	0.669007	Cluster5

Figure 9 Customer segmentation based on behavioral data: Properties-> Exported Data-> TRAIN

We can see that this segment has *Default_Credit* as *Yes.* Also this is the only sgement with *Default_Credit* as *Yes*, rest all are *No.* Hence, this explains the fact that why we don't see the *Default_Loan* attribute in other segments in *Profile*. The reason for this is that all have the same value of *Deafult_Loan*, so SAS can't differentiate the segments based on *Default_Loan*, so it don't include it in *Profile* information. Whereas, for cluster five this value is what makes them a separate segment.

3.2 Important variables in each segment

The segments list below are order (decreasing) in how uniquely they can be identified in the whole dataset and show the worth of variables:

Segment 2: Housing_Loan (0.337) -> Personal_Loan (0.075)

Segment 1: Housing_Loan (0.370) -> Personal_Loan (0.056)

Segment 4: Personal_Loan (0.093) -> Housing_Loan (0.014)

Segment 3: Personal_Loan (0.044) -> Housing_Loan (0.009)

Segment 5: Personal_Loan (3.7141E-9) -> Housing_Loan (3.4924E-9)

As we can see that the segment two is most unique as it has the most distinct attributes and segment five is the least distinctive in the whole data set. We are not mapping these segments to Roy Morgan segments so we cannot determine which segment maps to what age, education, marital status, or job. But we still can get an idea of the financial conditions of the people in these segments as it tells how much debt they are in.

3.3 Long-term deposit: subscribers v/s non-subscribers

I applied clustering and segment profiling on both subsets using same setting we used for the whole dataset. In segments discovered for both *Subscribed* as *Yes* and *No* data set, we can notice is no difference in the segments in terms of attributes they share.

Table 1 Matching Subscribed Yes and No segments

Subscribed = *Yes*	Subscribed = *No*
Segment 1: 38.38% (1480)	Segment 1: 39.65% (10555)
Segment 2: 9.05% (349)	Segment 4: 9.37% (2494)
Segment 3: 46.45% (1791)	Segment 2: 44.63% (11880)
Segment 4: 6.12% (236)	Segment 3: 6.34% (1689)
	Segment 5: 0.01% (3)

Although subscribers and non-subscribers have different number of observations, but they have segments with same attributes. This shows that the whole data shares the same behavioural attributes whether subscribed or non-subscribed. The reason for this is because these behavioural attributes are independent of *Subscriber* attribute. This shows that people's decision of whether they subscribe for a long-term deposit or not is not based on *Personal_Loan, Housing_Loan, Default_Credit* attributes.

4. Cross cluster analysis – demographics to behavioural segments

I save the results from Task 1 and 2 using the *Save* node, where I used the whole data set to check behavioral and demographic segments. I used R Studio as suggested to perform the cross-cluster analysis.

	1	2	3	4	5
1	473	523	58	102	0
2	1109	1162	170	219	0
3	1329	1489	217	287	0
4	1708	2050	273	416	0
5	1065	1255	167	270	2
6	3317	3847	536	817	0
7	3034	3345	504	732	1

Figure 10 cross cluster for demographic segments (seven segments) and the behavioral segments (five segments)

In the cross cluster for demographic segments (seven segments) and the behavioral segments (five Segments) we can see that there are thirty-five associations (5*7). These associations display the customers or observations which are common in segments of demographics and behavioral. The segments with significant number of associations (top five) in decreasing order are listed below:

Table 2 Top five associations in decreasing order

Demographical	Behavioral	Number of observations
Segment 6	Segment 2	3847
Segment 7	Segment 2	3345
Segment 6	Segment 1	3317
Segment 7	Segment 1	3034
Segment 4	Segment 2	2050

These associations between segments show which segments have common customers across both demographics and behavior clusters sharing similarity or have associations. We can see segment six associated with segment two has the greatest number of observations (3847), followed by segment seven with segment two (3345), segment six with segment one (3317), segment seven with segment one (3034), and segment four with segment two (2050). I am discussing the two most significant associations.

4.1 Associations of demographics with behavioral

a. Segment 6 of demographics with segment 2 of behavioral

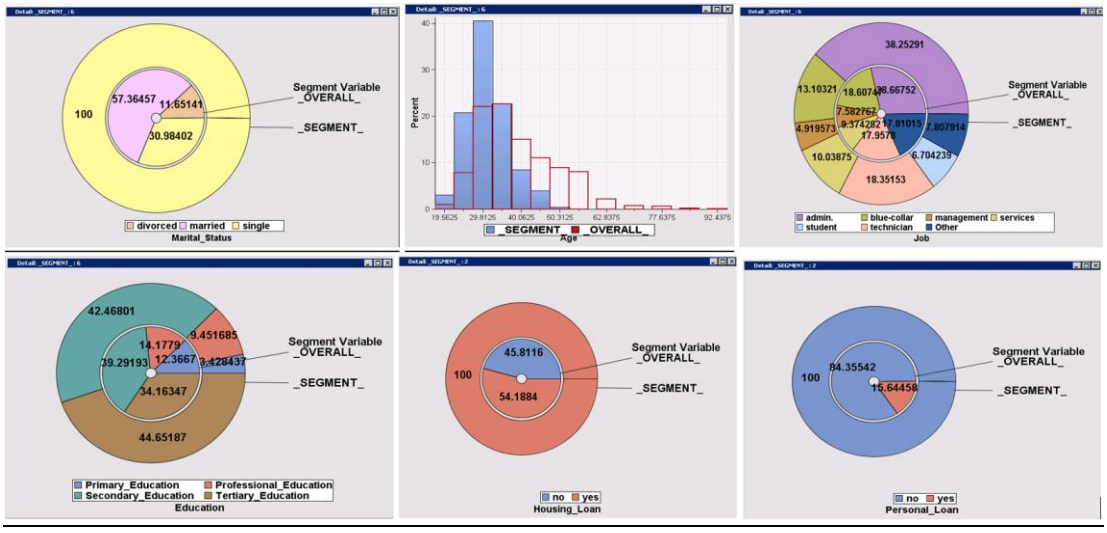

Figure 11 Segment 6 of demographics with segment 2 of behavioral

We can see in this association that all people are single and are mostly aged between 20-50. Most of them have tertiary qualification and are mostly into admin and technician jobs. All of them have a housing loan but people in this group don't have a personal loan. This shows that in our data set most people are single with high qualification, a decent job and have a house loan.

b. Segment 7 of demographics with segment 2 of behavioral

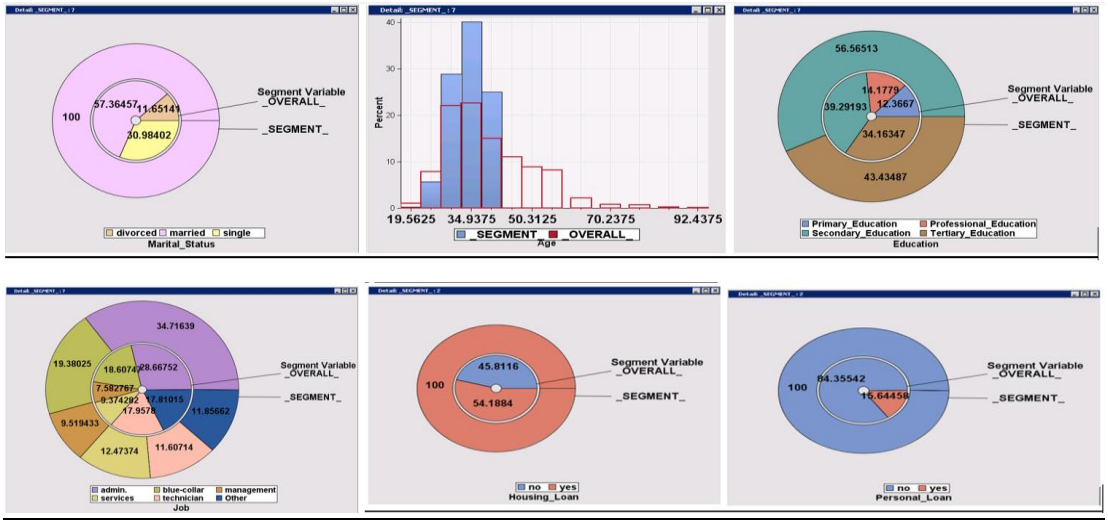

Figure 12 Segment 7 of demographics with segment 2 of behavioral

In this association we can see that all people are married and are age between 20-45. Most of them have secondary qualification and are mostly into admin and blue-collar jobs. All of them have a housing loan but people in this group don't have a personal loan. This shows that in our data set we have a big segment of people who are married, have secondary qualification, have an admin or blue-collar jobs (low paying jobs as compared to technician) and have a house loan.

These associations talk about the attributes of the whole dataset but doesn't tell us about the people who can subscribe for a long-term loan in terms of demographics and behavioral segments.

For this we need to get the cross-clustering associations between demographics and behavioral for the whole data set and cross-cluster associations between demographics and behavioral for the people who subscribed. To calculate the importance of the association we are using *lift* of subscriber as "yes".

4.2 Associations of demographics with behavioral for Subscribed as *yes*

The results of calculated *lift* I got from R studio show some major associations between the outcome (Subscribed) and the combined demographics and behavioural segments. We can see the associations in decreasing order of their *lift* value below:

Table 3 Lift values in decreasing order

Demographical	Behavioral	Lift value (decreasing)
Segment 1	Segment 4	0.372
Segment 1	Segment 3	0.327
Segment 1	Segment 2	0.321
Segment 1	Segment 1	0.317
Segment 6	Segment 2	0.157
Segment 6	Segment 3	0.156
Segment 4	Segment 4	0.151
Segment 6	Segment 1	0.145
Segment 6	Segment 4	0.133
Segment 5	Segment 2	0.132
Segment 4	Segment 2	0.125
Segment 5	Segment 4	0.118
Segment 4	Segment 3	0.117
Segment 4	Segment 1	0.116
Segment 5	Segment 3	0.113
Segment 7	Segment 2	0.109
Segment 3	Segment 1	0.105
Segment 7	Segment 3	0.099
Segment 3	Segment 4	0.097
Segment 5	Segment 1	0.096
Segment 3	Segment 2	0.094
Segment 3	Segment 3	0.092
Segment 7	Segment 4	0.086
Segment 2	Segment 2	0.075
Segment 2	Segment 1	0.073
Segment 2	Segment 4	0.073
Segment 2	Segment 3	0.070
Segment 1-7	Segment 5	0/NAN

From the table above I am discussing the top two associations with the highest *lift* value to check their attributes for better understanding what key factors are responsible for people to subscribe for a long-term deposit.

a. Association of segment 1 with segment 4

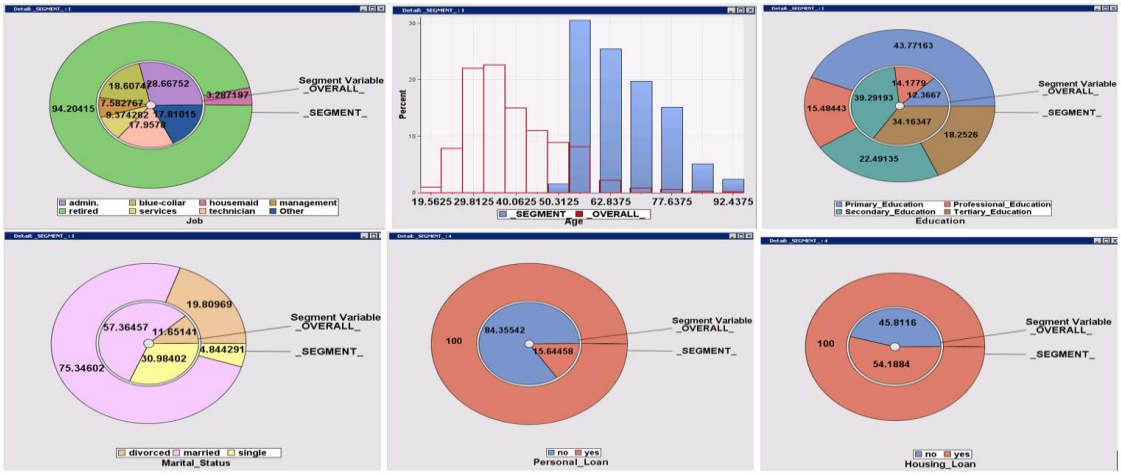

Figure 13 Associations of demographics with behavioral for Subscribed as yes: *segment 1 with segment 4*

As we can see in the *Profile* information for this association, most of the people are retired and are mostly aged between 50-90. Most of them have primary education and are married and divorced. All of them have a housing loan and have a personal loan. This makes sense as people of this age group might have a lot of savings as they have worked for most of their lives and could be interested to opt for long-terms deposit as a backup for any situations like disease or hospitalization. However, we can see that these people still have house and personal loans which is interesting to see.

b. Association of segment 1 with segment 3

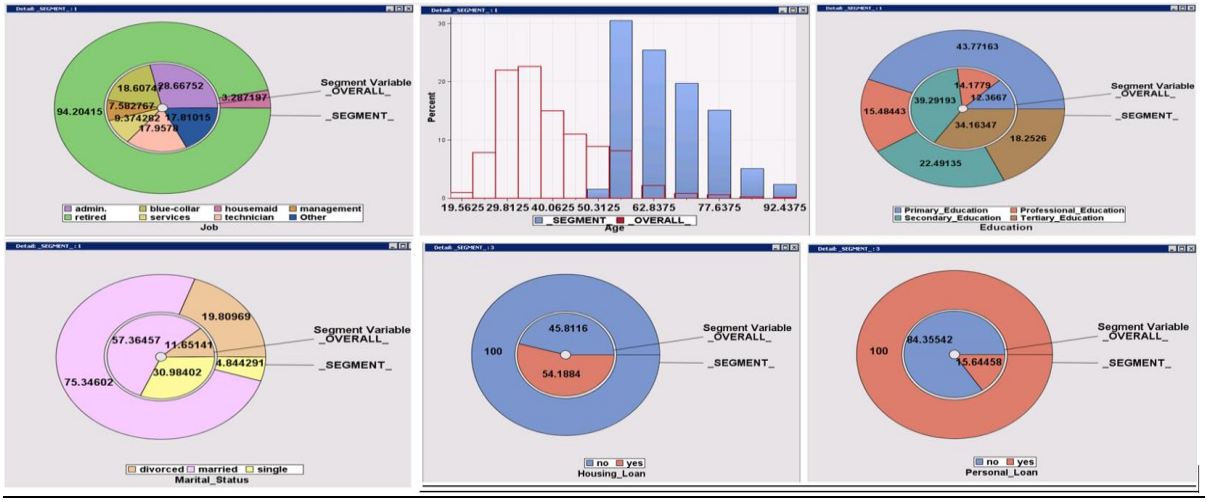

Figure 14 Associations of demographics with behavioral for Subscribed as yes: *segment 1 with segment 3*

We can see in the *Profile* information for this association, all attributes are the same expect people in this association don't have a housing loan.

The relation of all these attributes to long-term deposits tells us a story and gives an idea to the Future Bank about the customers they can target for long-term deposits based on the *lift* values we discovered in our study.

5. Customer segmentation based on combined demographic and behavioural data

In this Task, I set all the demographic and behavioural variables to *Use* as *Yes (*in both *Cluster* and *Segment Profile* nodes*)* as we are trying to segment based on combined demographic and behavioural data. In the *Cluster* node I tried the *Maximum Number of Clusters with* 5/6/7 cluster size to get the best cluster size for segmentation. I

considered cluster size of five as clusters are more spread out and don't overlap with each other as much they do in six or seven clusters. I confirmed this using the *Cluster Proximities* plot. I set the *Minimum Worth* in *Segment Profile* node to 0.000001to get all variables in the *Profile* panel in Results. This will help giving more insights into the differences of variables amongst all segments. I have discussed the key demographic segments which stand apart in whole dataset in terms of the attributes they reflect.

5.1 Segments

a. Segment 2

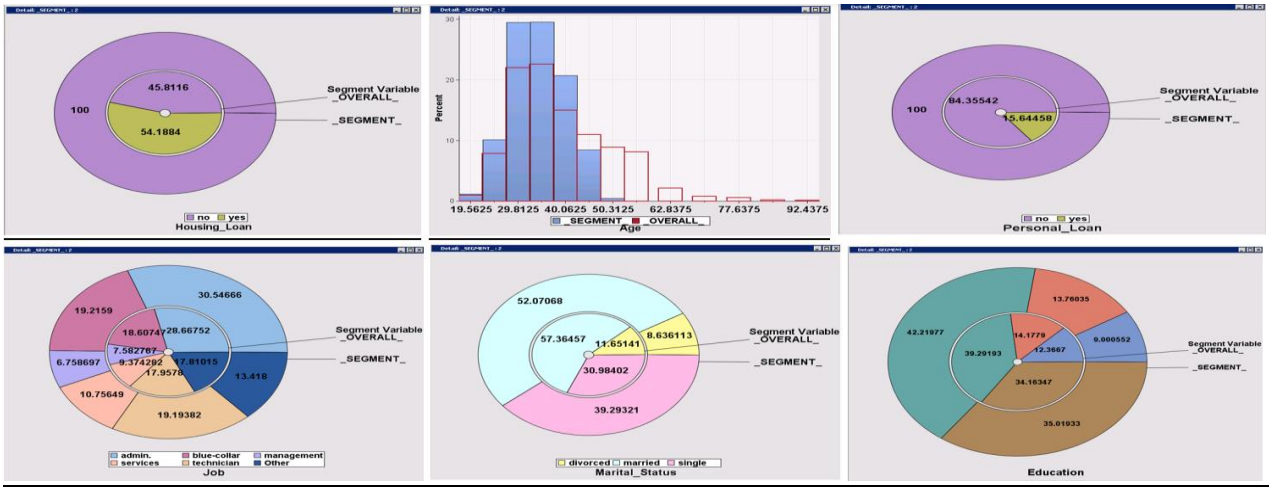

Figure 15 Customer segmentation based on combined demographic and behavioural data: Segment 2

In this segment we can see most people are between the age of 25 to 45. Nearly half of the people are married, and others are mainly single, and a few are divorced. Most people are working in admin jobs, and in blue-collar, technician jobs. Most people have secondary education and don't have a house or personal loan. This shows that most people in this segment are working and have decent jobs with good education.

b. Segment 4

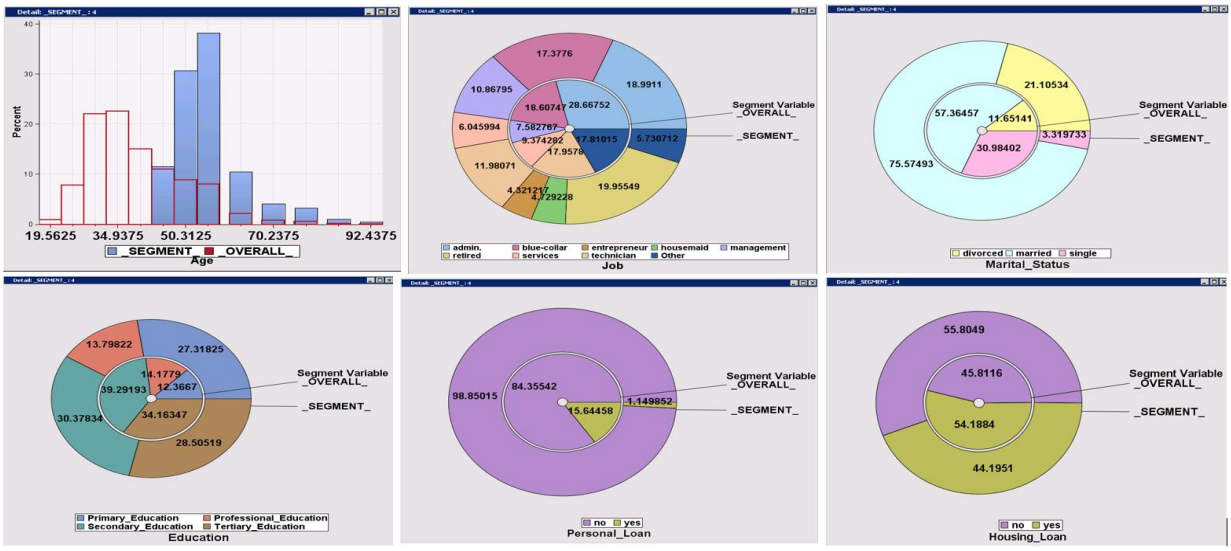

Figure 16 Customer segmentation based on combined demographic and behavioural data: Segment 4

In this segment we can see most people are between the age of 45 to 75. Most of the people are *married* and others are mainly *single*. Most people are working in *retired*, followed by *admin, blue-collar, technician* jobs. Around 30% people have secondary education and ~28.5% have *Tertiary_Education*, followed by people with *Primary_Education* People don't have a *Personal_Loan* but ~44% have *Housing_Loan*. This shows that most people in this segment are in

their later stage of life and are retired or going to retire soon. People have decent jobs with good education. They don't have a personal loan, but some do have a housing loan to payoff.

5.2 Important variables considering the outcome (Subscribed)

I used the *filter* node to get only the *Subscribed* as *Yes* observations in segments. I set all the demographic and behavioural variables to *Use* as *Yes (*in both *Cluster* and *Segment Profile* nodes*).* In the *Cluster* node I choose the *Maximum Number of Clusters with* 5 based on previous clustering results. Following are the clusters with the variable worth:

Segment 4: Martial_Status (0.30800) -> Age (0.12246) -> Job (0.04452) -> Personal_Loan (0.03115) -> Education (0.01394) -> Housing_Loan (0.00007)

Segment 5: Housing_Loan (0.098783) -> Marital_Status (0.062622) -> Age (0.030882) -> Personal_Loan (0.022032) -> Job (0.018694) -> Education (0.003277)

Segment 2: Housing_Loan (0.11893) -> Marital_Status (0.04675) -> Age (0.02587) -> Personal_Loan (0.01706) -> Job (0.01302) -> Education (0.00101)

Segment 1: Personal_Loan (0.23670) -> Age (0.00242) −> Job (0.00146) -> Housing_Loan (0.009) -> Education (0.00021) -> Marital_Status (0.00006)

Segment 3: Age (0.11806) −> Job (0.11458) -> Education (0.03920) -> Marital_Status (0.00969) -> Personal_Loan (0.00096) -> Housing_Loan (0.00001)

The information above shows the segment-wise important variables. We can see that *Marital_Status, Housing_Loan, Personal_Loan* and *Age* are the most important variables considering our outcome in these segments (*Subscribed*).

5.3 Difference in segments and profiles

Associations found in cross-cluster analysis of behavioral and demographic segments are different to what we discovered in last part (*Subscribed* as *yes* and segments for all the variables). I am comparing an instance which overlaps but has some difference in attribute values.

a. Association of segment 1 with segment 3 v/s Segment 4 of combined segmentation

We can see in figure 14 and figure 16, that both have approximately same values in *Marrital_Status*. However, in terms of *Job, Housing_Loan, Age, and Education* they have different values. From his instance we can see that the segments are very different from each other. This is simply because in SAS clustering when we use the whole data set, it will consider all the variable at once, whereas the associations we created are not done in one step. So, the way they come together is different.

6. Conclusion

We can see from the analysis that demographic based segments are different to the behavioral based segments. Also, from Future Bank's marketing strategy point of view some variables and profiles are more worth marketing so that more people can subscribe to long-term deposits. This analysis also tells that cross clustering gives us different results to what we get in clustering the whole dataset.

7. Appendices

First three figures in this section are for the segments discussed in part 2.1 Segments.

1. Segment 1

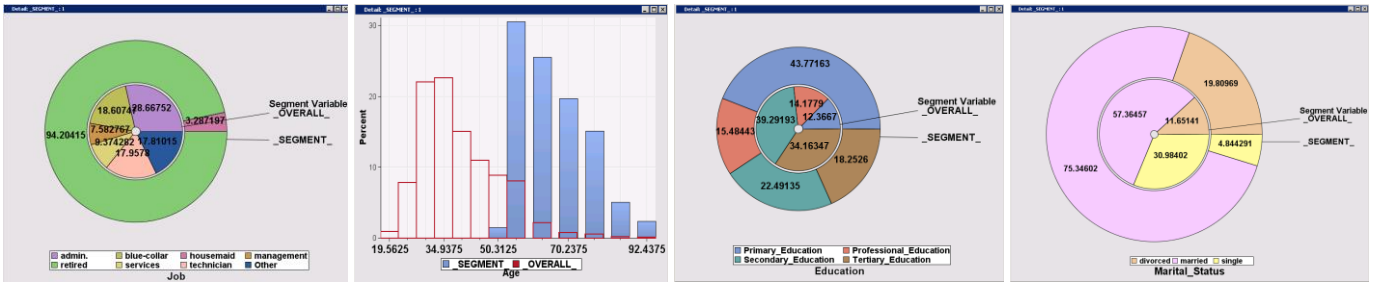

Figure 17 Customer segmentation based on demographics data: Segment 1

2. Segment 5

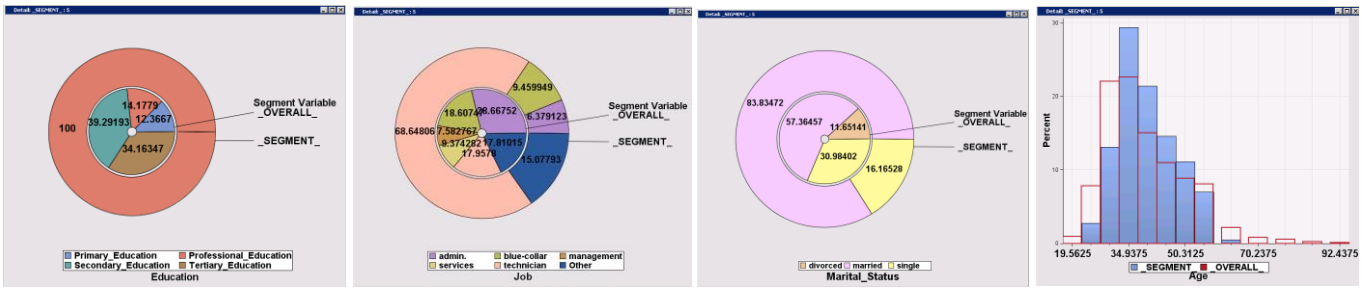

Figure 18 Customer segmentation based on demographics data: Segment 5

3. Segment 7

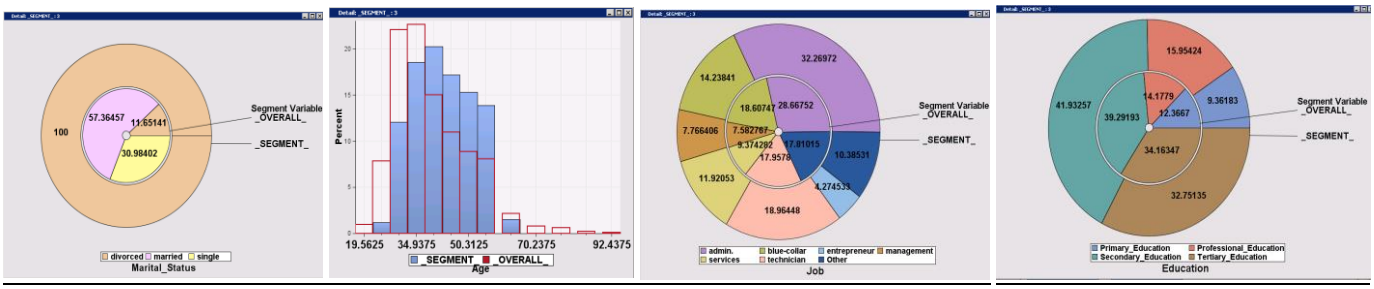

Figure 19 Customer segmentation based on demographics data: Segment 7

4. Cluster proximity plot used in part 5. Customer segmentation based on combined demographic and behavioural data. This is to decide on the number of clusters.

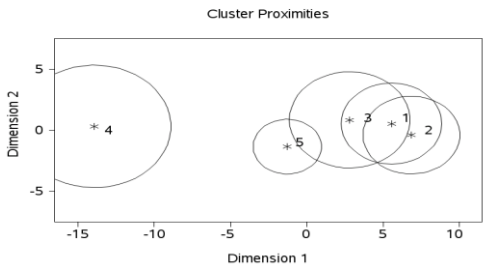

Figure 20 Customer segmentation based on combined demographic and behavioural data: Cluster proximity plot

5. *Lift* calculated in part 4.2 Associations of demographics with behavioral for Subscribed as *yes.*

	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
Segment 1	0.31712474	0.32122371	0.32758621	0.37254902	NaN
Segment 2	0.07394049	0.07573150	0.07058824	0.07305936	NaN
Segment 3	0.10534236	0.09469443	0.09216590	0.09756098	NaN
Segment 4	0.11651054	0.12536585	0.11721612	0.15144231	NaN
Segment 5	0.09671362	0.13227092	0.11377246	0.11851852	0
Segment 6	0.14531203	0.15726540	0.15671642	0.13341493	NaN
Segment 7	0.10678972	0.10941704	0.09920635	0.08606557	0

Figure 21 Associations of demographics with behavioral for Subscribed as yes: Lift values