Aston Business School
Birmingham

PORTFOLIO TASK 3: CLUSTERING ANALYSIS

Introduction

A study is conducted on behalf of a business analytics company that has signed contract with a UK bank. The bank's product development team desires to undertake segmentation analysis to identify trends and patterns in a sample of records collected from a number of their customers. This data has to be analyzed to segment the customers into homogenous group to extract meaningful trends and patterns. The IBM SPSS Statistics software will be used to carry out the analysis, hence, the data should be re-arranged accordingly. As we are required to perform segmentation, clustering analysis will be used.

Variables

Based on the variables, the appropriate method for segmentation will be used. The variables involved in this dataset are Current Account, Savings Account, Months Customer, Months employed, Gender, Marital Status, Age, Housing Job and Credit Risk. From this information, clusters can be made based on credit risk to check how many of their customers fall under the category of having high credit risk. This can be an important factor to segment as credit risk value will impact on loan eligibility. People will high credit risk are less likely to get loans. Hierarchical clustering ought to be used as we cannot pre-determine the number of clusters (k) in this case. It is vital to arrange the data accordingly. In the given data, Gender, Marital status, Job and Credit Risk fall under categorical variables; so, they should be converted to numerical for the ease of analysis in SPSS.

Clustering analysis

Clustering analysis is a method that groups similar objects based on different characteristics. These clusters are internally homogenous and externally heterogenous. In Hierarchical clustering a series of partitions take place, which may run from a single cluster to n cluster or vice-versa. Based on this, they can be classified as Agglomerative Hierarchical clustering (where every object starts in its cluster and are narrowed down to single cluster) and Divisive Hierarchical clustering (where all objects start from a single cluster and end up with each object in a single cluster). In every case, after observing the dendrogram, the number of clusters can be estimated.

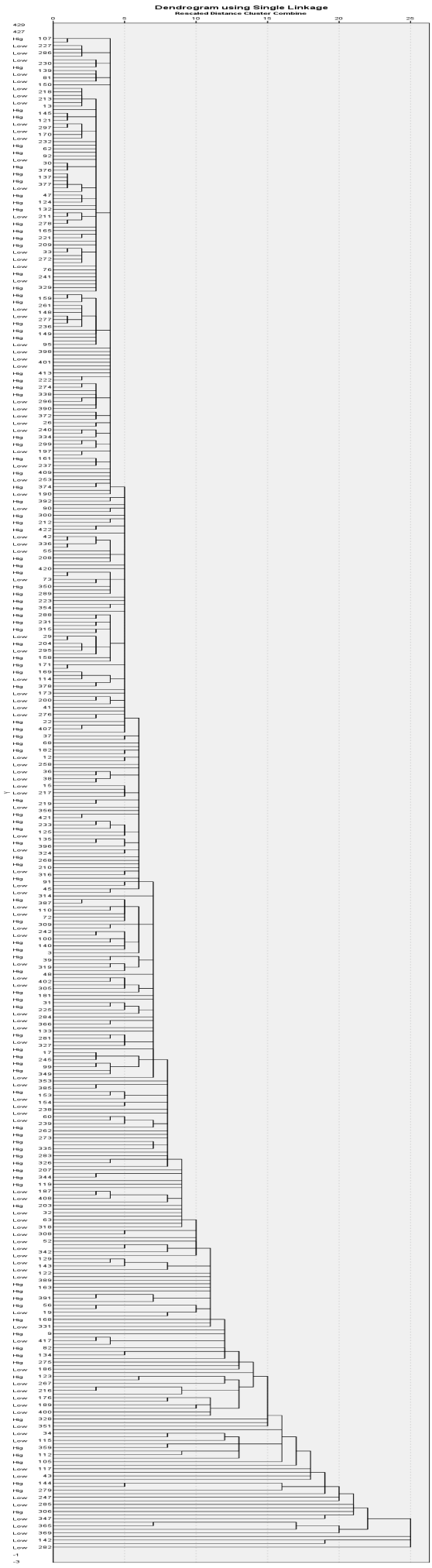
On analysis the following results are obtained.

1. **Cluster Model 1:**

- This cluster is formed by 'Credit Risk' as the label case. Here under method **single-linkage clustering** (Nearest neighbour in SPSS), under interval measure **Euclidean distance** was chosen standardized in the **range 0 to 1**. The dendrogram for this case of cluster analysis is shown below. From this image, it can be seen that the ideal number of clusters may range from 3 to 5.

  The following image is of the Dendrogram obtained using Ward's method.

Aston Business School
Birmingham



Considering number of Clusters as 5, the following distribution can be obtained.

Single Linkage

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	417	98.1	98.1	98.1
	2	4	.9	.9	99.1
	3	1	.2	.2	99.3
	4	2	.5	.5	99.8
	5	1	.2	.2	100.0
	Total	425	100.0	100.0	

- The non-uniform distribution is clear from the above table. Hence, it can be concluded that this is not the ideal cluster.

2. Cluster Model 2:

- This cluster is formed by 'Credit Risk' as the label case. Here '**Ward method'** is used with the same interval measure (**Euclidean distance**) standardized in the **range 0 to 1**. The dendrogram for this case is shown below. It can be seen that the ideal number of clusters can be 4 or 5.
- The distribution is obtained by changing the number of clusters. From the tables below it is clear that number of clusters set to 5 displays good distribution.

Number of clusters =4

Ward Method

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	215	50.6	50.6	50.6
	2	70	16.5	16.5	67.1
	3	91	21.4	21.4	88.5
	4	49	11.5	11.5	100.0
	Total	425	100.0	100.0	

Number of clusters =5

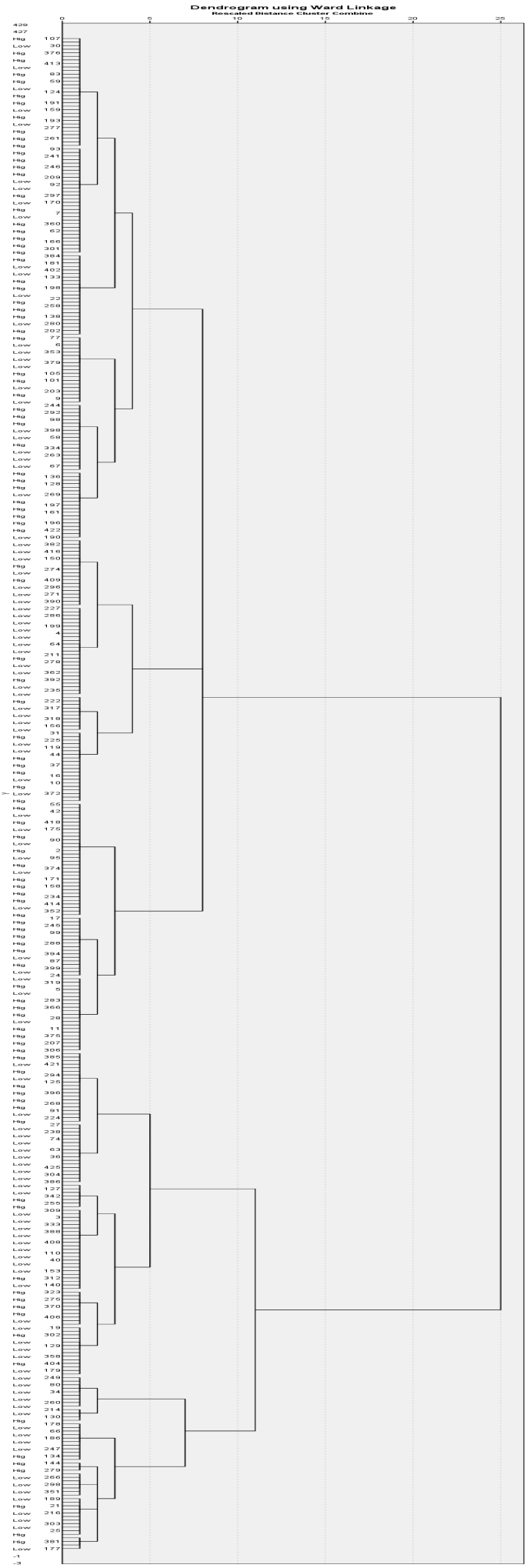
Ward Method

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	141	33.2	33.2	33.2
	2	70	16.5	16.5	49.6
	3	91	21.4	21.4	71.1
	4	74	17.4	17.4	88.5
	5	49	11.5	11.5	100.0
	Total	425	100.0	100.0	

On the basis of these clusters, we can determine the variation of credit risk among various categories.

The following image is of the Dendrogram obtained using Ward's method.

Aston Business School
Birmingham



Cluster Analysis

Out of the two clusters the second cluster, the one formed by using 'Ward's method' is clearer to distinguish. Cluster Model 1 which is formed by using Single linkage uses nearest neighbours hence it results in tight clusters which are a bit difficult to read. Therefore, considering cluster model 2, the following observations were noted.

Clusters	Current account max value	Savings account Max value	People with High Credit Risk
1	9783	8357	54.9%
2	11072	4754	63.4%
3	2641	3613	47.3%
4	3329	3529	26%
5	19812	19811	32.6%

Every cluster was separated so that within each cluster the maximum value of the current account and savings accounts values could be found.

	CurrentAccou	SavingsAccou	MonthsCusto	MonthsEmpl	Age	GenderValue	MaritalStatu	HousingValu	JobValue	CreditRisk	CLU5_1		CreditRisk
2	0	1230	25	0	32	0	0	1	1	Hig	2		Hig
3	963	4754	40	45	31	0	1	2	1	Low	2		Low
4	0	989	49	0	32	0	1	2	2	Hig	2		Hig
5	652	732	49	4	25	1	0	1	1	Hig	2		Hig
6	0	485	37	23	27	1	0	1	2	Hig	2		Hig
7	2484	0	49	46	34	0	1	0	1	Low	2		Low
8	237	236	37	24	23	0	1	2	1	Low	2		Low
9	0	150	49	46	36	1	0	2	1	Hig	2		Hig
10	0	323	49	42	33	0	2	1	1	Hig	2		Hig
11	218	0	49	0	39	0	1	0	2	Low	2		Low
12	0	109	25	26	34	0	1	1	0	Low	2		Low
13	0	724	25	8	30	0	1	2	1	Hig	2		Hig
14	870	917	28	6	35	0	1	1	1	Hig	2		Hig
15	0	789	25	28	37	0	1	1	2	Low	2		Low
16	674	2886	49	32	29	0	1	1	1	Low	2		Low
17	713	784	61	17	41	0	1	0	1	Hig	2		Hig
18	0	680	25	3	34	1	0	1	1	Hig	2		Hig
19	0	104	37	25	23	0	1	1	1	Hig	2		Hig
20	0	706	31	14	31	0	0	1	1	Low	2		Low
21	0	710	25	1	37	1	0	1	1	Low	2		Low
22	0	192	46	13	22	0	1	0	1	Hig	2		Hig
23	514	405	49	13	21	1	0	1	1	Hig	2		Hig
24	0	116	49	45	45	0	1	0	1	Hig	2		Hig
25	509	241	25	14	35	0	1	1	0	Hig	2		Hig
26	0	609	37	6	31	0	1	0	2	Low	2		Low
27	0	609	31	3	33	0	0	1	0	Hig	2		Hig
28	0	270	25	25	34	0	1	1	1	Low	2		Low
29	0	922	37	9	24	1	0	1	2	Hig	2		Hig
30	0	309	49	37	25	0	1	1	1	Low	2		Low
31	216	262	37	2	32	0	1	2	0	Hig	2		Hig
32	109	540	37	1	27	0	2	2	2	Hig	2		Hig
33	0	772	25	19	32	0	0	1	1	Low	2		Low
34	0	750	37	2	27	0	0	1	1	Hig	2		Hig

Sheet1 | Sheet2 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | +

Fig: Image showing data of cluster 2 in excel sheet

In the second cluster model, five clusters were formed. The maximum values of current savings value, savings account value and high credit risk percentage is shown in the table above. In cluster 5, there are few people with high credit risk as they have more current account value and more savings. In cluster 4, the people have less of both and very few have high credit risk, it maybe because most of them do not prefer or require loans. Clusters 1 and 2 are more prone to credit risks given less amount in their accounts. Cluster 3 has less than 50% account holders prone to credit risks.

Conclusion

Many clusters can be formed using various methods to extract meaning overall information through Cluster analysis. In this report, the dataset used 'Hierarchical Method' of clustering. Two cases were considered Ward's method and single linkage clustering, in which, ward's method showed better results. Through clustering, a rough idea of people prone to credit risk can be estimated.