**Big Data and Predictive Analysis**

Lab 1

Assignment 2

**Cluster Analysis**

Submitted by

1. **Bank Direct Marketing Cluster Analysis**

Given the inputs, do clusters of customers exist in the bank direct marketing data set? This exercise explores the bank direct marketing data and tries to profile the resulting clusters.

Take screenshots at every stage, you might want to recheck them or paste them for several questions:

1. Create a new diagram in your project. Name the diagram **Bank Clustering**.
2. Use the **bank\_direct\_marketing** data as a data source for this clustering and profiling exercise.
3. Determine whether the model roles and measurement levels assigned to the variables are appropriate.
4. Examine the distribution of the values of these variables:

* **balance**
* **day**
* **previous**
* **duration**
* **age**
* **campaign**

The three most heavily skewed distributions are for **balance**, **campaign**, and **previous**. Although not optimal, we could reduce the skewness of the distributions by taking the log of the variable.

1. Drag a **Transform Variables** node onto the diagram and connect it to the **Input Data** source.
2. Apply a log transformation to the following variables:

* **balance**
* **previous**
* **campaign**

1. Connect a **Cluster** node to the **Transform Variables** node.
2. Change **Maximum Number of Clusters** to **6**.
3. Change **Use** for all the variables to **No**, except for these:

* **balance**
* **previous**
* **duration**
* **Age**
* **campaign**

d. Examine the distribution of the values of these variables:

* **balance**
* **day**
* **previous**
* **duration**
* **age**
* **campaign**

The three most heavily skewed distributions are for **balance**, **campaign**, and **previous**. Although not optimal, we could reduce the skewness of the distributions by taking the log of the variable.

Paste your screenshot here along with your student number

1. Drag a **Transform Variables** node onto the diagram and connect it to the **Input Data** source.

Paste your screenshot here along with your student number

1. Apply a log transformation to the following variables:

* **balance**
* **previous**
* **campaign**

Paste your screenshot here along with your student number

1. Connect a **Cluster** node to the **Transform Variables** node.

Paste your screenshot here along with your student number

1. Change **Maximum Number of Clusters** to **6**.

Paste your screenshot here along with your student number

1. Change **Use** for all the variables to **No**, except for these:

* **balance**
* **previous**
* **duration**
* **Age**
* **campaign**

Paste your screenshot here along with your student number

1. Run the Cluster node and view the plot of cluster proximities (**View** ⇨ **Cluster Distance** ⇨ **Plot**).

Type your answer here:

How many distinct clusters do you observe?

You can provide your reason here:

1. Are there clusters that can be combined to form fewer clusters?

Type your answer here:

1. Change **Maximum Number of Clusters** to **3** and rerun the **Cluster** node.

Paste your screenshot here along with your student number

1. Drag a **Segment Profile** node onto the diagram workspace and connect it to the **Cluster** node.

Paste your screenshot here along with your student number

1. Run the **Segment Profile** node.
2. Inspect the profile results.

Paste your screenshot here along with your student number

Paste your screenshot here along with your student number

1. How would you label Clusters 1 and 2?

Cluster 1:

Cluster 2:

1. What should you do about Cluster 3 containing only one observation?

Cluster 3:

You can explain your reason and can paste any screenshot to substantiate:

Paste your screenshot here along with your student number

1. **Conducting Cluster Analysis**

The **DUNGAREE** data set gives the number of pairs of four different types of dungarees sold at stores over a specific time period. Each row represents an individual store. There are six columns in the data set. One column is the store identification number, and the remaining columns contain the number of pairs of each type of jeans sold.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Model Role** | **Measurement Level** | **Description** |
| **STOREID** | ID | Nominal | Identification number of the store |
| **FASHION** | Input | Interval | Number of pairs of fashion jeans sold at the store |
| **LEISURE** | Input | Interval | Number of pairs of leisure jeans sold at the store |
| **STRETCH** | Input | Interval | Number of pairs of stretch jeans sold at the store |
| **ORIGINAL** | Input | Interval | Number of pairs of original jeans sold at the store |
| **SALESTOT** | Rejected | Interval | Total number of pairs of jeans sold (the sum of **FASHION**, **LEISURE**, **STRETCH**, and **ORIGINAL**) |

**Answers:**

c. Examine the distribution of the variables.

Paste your screenshot here along with your student number

* Are there any unusual data values?

Type your answer here:

* Are there missing values that should be replaced?

Type your answer here:

* 1. Assign the variable **STOREID** the model role **ID** and the variable **SALESTOT** the model role **Rejected**. Make sure that the remaining variables have the **Input** model role and the **Interval** measurement level. Why should the variable **SALESTOT** be rejected?

Type your answer here:

* 1. Drag the **DUNGAREE** data source to the diagram workspace.

Paste your screenshot here along with your student number

* 1. Add a **Cluster** node to the diagram workspace and connect the **Input Data** node to it.

Paste your screenshot here along with your student number

* 1. What would happen if you did not standardize your inputs?

Type your answer here:

* 1. Run the diagram from the Cluster node and examine the results. Does the number of clusters that are created seem reasonable?

Type your answer here:

Specify a maximum of six clusters and rerun the Cluster node.

Paste your screenshot here along with your student number

* 1. How does the number and quality of clusters compare to that obtained in part e ?

Type your answer here:

* 1. Use the Segment Profile node to summarize the nature of the clusters.

Type your answer here:

Paste your screenshot here along with your student number

* 1. Based on this analysis what can you conclude about the purchasing pattern of the customers?

Paste your screenshot here along with your student number

Paste your screenshot here along with your student number

Type your answer here: