Mälardalen University
School of Innovation Design and Engineering
Västerås, Sweden

Applied Machine Learning
Course Code: DVA263

ASSIGNMENT 2: UNSUPERVISED LEARNING INL2

Nicola BALDONI
nicola.baldoni@gmail.com

Vincent NAZZARENO
vincent.nazzareno@edu.devinci.fr

Examiner: Mobyen Uddin Ahmed, Shaibal Barua
Mälardalen University, Västerås

13th December 2024

1. Introduction

In this assignmnent we will analyze the Market Segmentation dataset. Each customer is characterized by different statistics, binded to different behavioral aspects. As this is an unsupervised learning assignment, there is no target variable to predict. Some customers might have the same behavior even though no clear target variables is present. Clustering is the technics that will be used to shed light on those customer groups. The goal of this assignment is to find ways to give recommendations like saving plans, loans, wealth management based on those behavior to target customer groups. The link to the code is in the GitHub.

2. Data Loading

We are using the Market Segmentation dataset, which contains all the information needed to conduct marketing recommendations for customer groups. The dataset includes numerical statistics representing various behaviors of different customers. It contains 17 original features (and 1 ID column) and 8,950 unique *CUST_ID*. We added 8 new dependant features.

To begin, we will perform data cleaning (missing data imputation), data preprocessing (scaling for the different algorithms) and finally data analysis (with metrics).

	CUST_ID	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_P
0	C10001	40.900749	0.818182	95.40	0.00	95.4	0.000000	0.166667	
1	C10002	3202.467416	0.909091	0.00	0.00	0.0	6442.945483	0.000000	
2	C10003	2495.148862	1.000000	773.17	773.17	0.0	0.000000	1.000000	
3	C10004	1666.670542	0.636364	1499.00	1499.00	0.0	205.788017	0.083333	
4	C10005	817.714335	1.000000	16.00	16.00	0.0	0.000000	0.083333	

Figure 1: Market Segmentation dataset (not all columns are displayed here)

2.1. Data Cleaning

After loading the dataset using the *pandas.read_csv()* function, we identify that 314 rows contain at least one missing variable, specifically in *CREDIT_LIMIT* or *MINIMUM_PAYMENTS*.

We have used a clustering-based imputation approach to imput the missing values. By doing K-Means imputation, we leverage the customer behavior to imput a "right" value and not simply the mean of the whole dataset but the mean of the cluster. We use a two step approach to imput first the missing values of *CREDIT_LIMIT*, then do the same for *MAXIMUM_PAYMENTS*.

First, we drop the *CUST_ID*, *CREDIT_LIMIT*, and *MINIMUM_PAYMENTS* columns from the dataset to focus on other features for clustering. We then apply the *KMeans* algorithm with 5 clusters (*n_clusters=5*), performing 10 initializations (*n_init=10*) and a maximum of 300 iterations (*max_iter=300*).

The dataset is assigned cluster labels, and we impute missing values in the *CREDIT_LIMIT* column within each cluster, ensuring that imputation considers patterns specific to each group. After imputing, the cluster labels are removed, and the process is repeated for the *MINIMUM_PAYMENTS* column.

2.2. Data analysis

To analyze the dataset, we first computed the hierarchical correlation clustering matrix (Figure 2), which reveals two prominent clusters of strongly correlated features. This clustering highlights a natural separation in the data based on related customer behaviors.

Next, we conducted a silhouette analysis for KMeans clustering with clusters ranging from 2 to 9 and computed the *silhouette score*. 4 clusters was the lowest average silhouette score at *0.1937*, and the plot can be seen on Figure 3 . The silhouette plot shows that most clusters

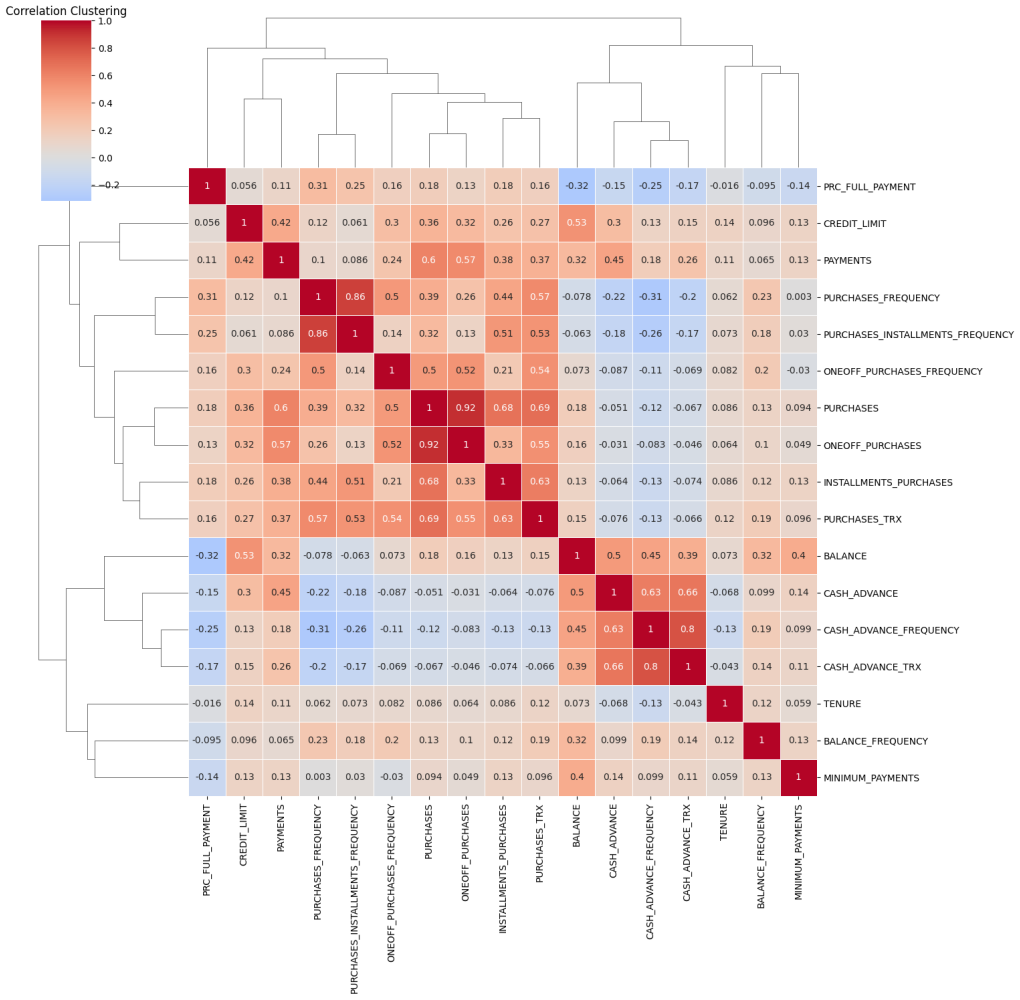

Figure 2: Hierachical Correlation Clustering Matrix of the cleaned dataset

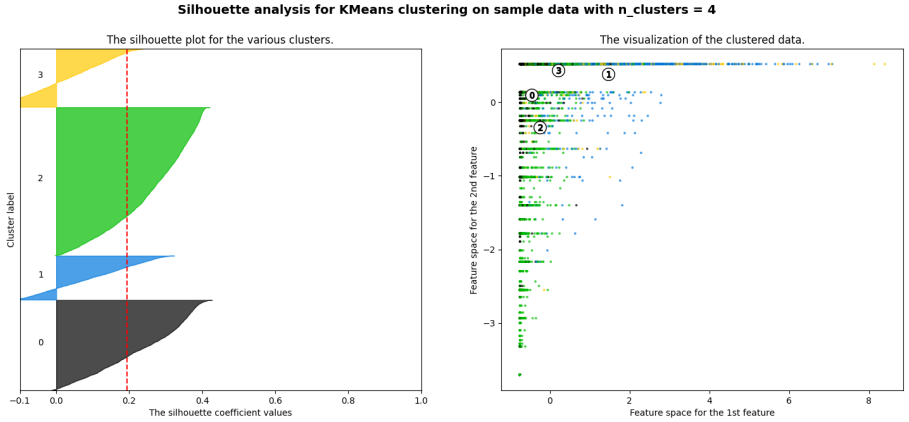

Figure 3: Silhouette analysis for KMeans clustering on sample data with 4 clusters

have positive silhouette coefficients, indicating good separation between clusters. The K-Mean clustering approach appears effective for segmenting customers into meaningful groups based on their behaviors.

The following Figure 4 shows the PCA decomposition of the first two dimension applied to the Z-Score Scaled dataset. The right plot shows the effect of KMeans clustering pretty clearly and how different behaviors can be segmented into distinct groups. A T-SNE plot can be seen in the appendix section under Figure 6.

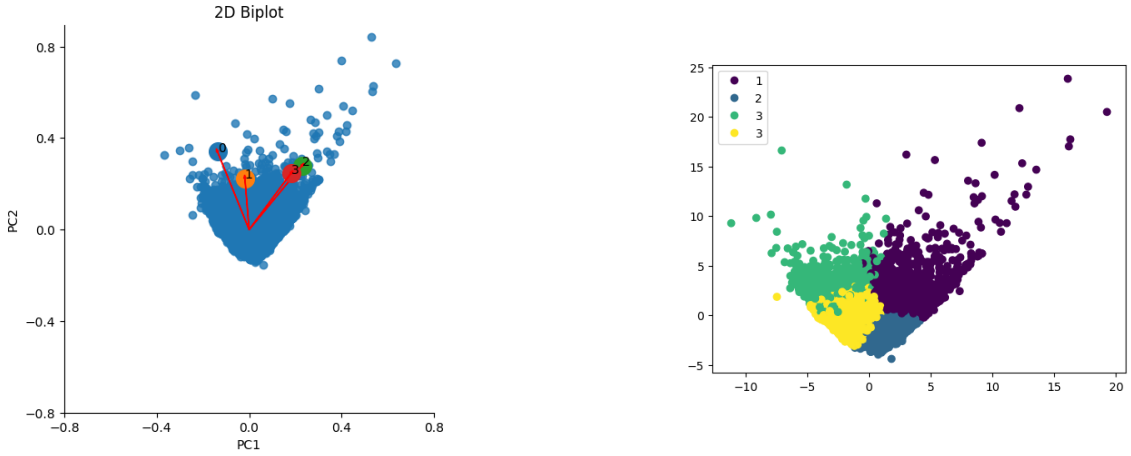

Figure 4: PCA Analysis: Biplot and PCA with KMeans Clustering

3. Results

The heatmap and feature engineering clearly highlighted group characteristics (Figure 5) It allows us to better understand the specific behaviors of each cluster across the various features. Each feature is normalized (using z-score) to highlight the relative importance of the clusters within that feature, making it easier to interpret their tendencies compared to other consumers cluster. From those 4 clusters of consumers, we can assign a specific behavior category: *Balanced Consumers, Highly Engaged Consumers, Outstanding Consumers* and *Risky Consumers.*

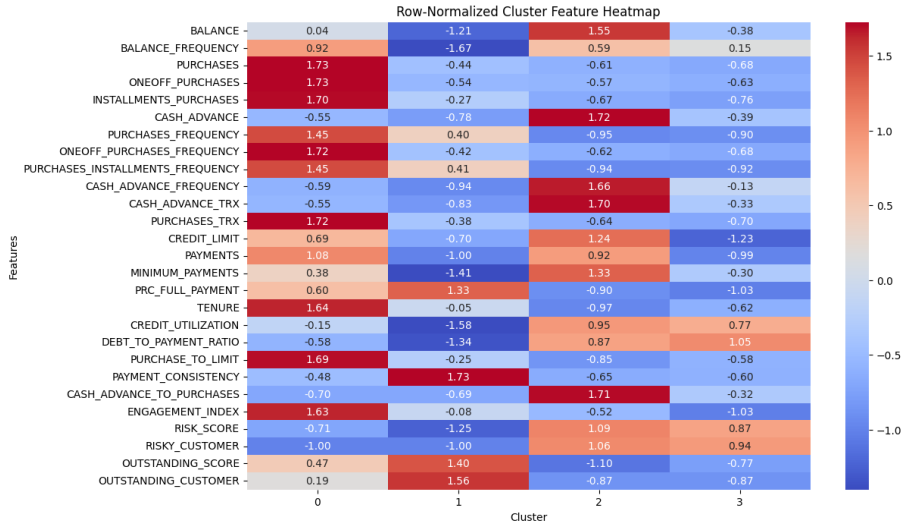

Figure 5: Row-Normalized Cluster Feature Heatmap (Z-Score)

Table 1: Summary of Consumer Types by Cluster

Cluster	Consumer Type	Marketing Recommendations
0	Balanced Consumers	Normal financial behavior with average *BALANCE* and *CREDIT_LIMIT*. Target with simple, basic financial products and some incentives to improve engagement to transition to a Highly Engaged Consumers.
1	Highly Engaged Consumers	High *PURCHASES*, *ENGAGEMENT_INDEX*, and *PURCHASE_TO_LIMIT*. Offer cashback rewards and installment plans while keeping track of repayment behavior. Most money earned from these consummers.
2	Outstanding Consumers	High *PAYMENTS* and *PAYMENT_CONSISTENCY*. Low *CREDIT_UTILIZATION*. Promote premium products like savings plans and some financial investment opportunities.
3	Risky Consumers	High *CREDIT_UTILIZATION*, *DEBT_TO_PAYMENT_RATIO*, and *CASH_ADVANCE_TO_PURCHASES*. Mainly Focus on credit counseling, debt consolidation, and offers to lower financial struggles.

4. Conclusion

Unsupervised clustering techniques gave us great insights into customer behavior and how to segment them without having a target variable to optimize onto. Instead, we saw that a segmentation of customers into balanced, engaged, outstanding, and risky consumers showed the main behaviors that can be used to construct targeted marketing strategies using unsupervised learning. Outstanding customers could cross-sell premium financial products without risk of financial struggle, while risky customers need financial support and financial help or credit management plans.

These insights were found using different visualization plots in the appendix. For example, the bar plot of cluster average behavior (Figure 7) and the radar chart of cluster behaviors (Figure 8) were used to understand the behavior. The row-normalized heatmap and the feature engineering made it easier to see all the characteristics at once, and understand the difference between each groups.

4.1. Ethics of Using These Technologies

However, using unsupervised algorithms in marketing brings some ethical considerations. Customers should be informed if their data is being analyzed, and care must be taken to avoid reinforcing biases in the data, such as disparities based on income or demographics (look at Bayesian statistics). Recommendations should be fair: for example vulnerable customers should be helped and not exposed to more financial complex products. Another relevant issue is data privacy: data on customers should be made anonymous to comply with the general data protection regulation (GDPR) in Europe for example, as well as other legal frameworks in place where the company has a market.

Appendix

A T-SNE Visualization with Clusters

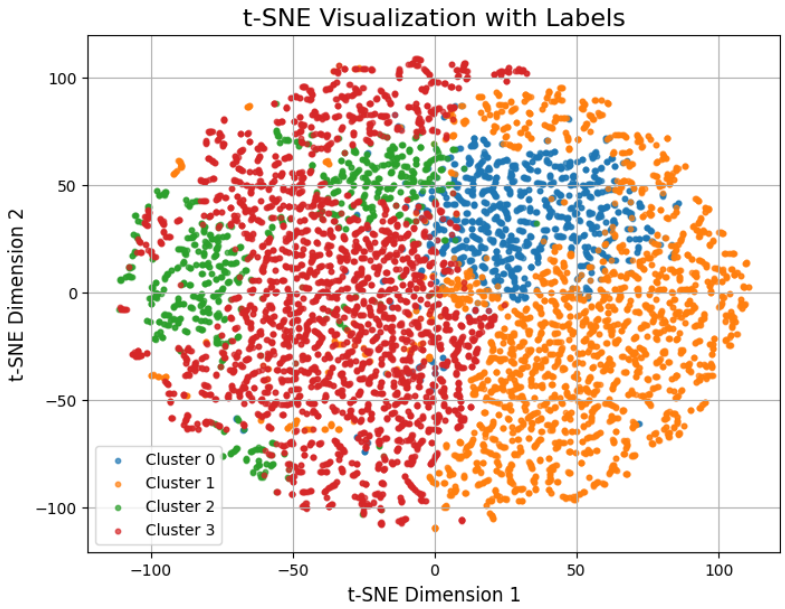

Figure 6: t-SNE Visualization with Clustering Results.

Figure 6 shows the t-SNE (t-Distributed Stochastic Neighbor Embedding) projection of the dataset into a 2D-space of the four clusters foundwith Kmeans. The clusters, while still distinctively separated remains close to each other with reflects similarities between groups of consummers.

B Cluster Average Behavior

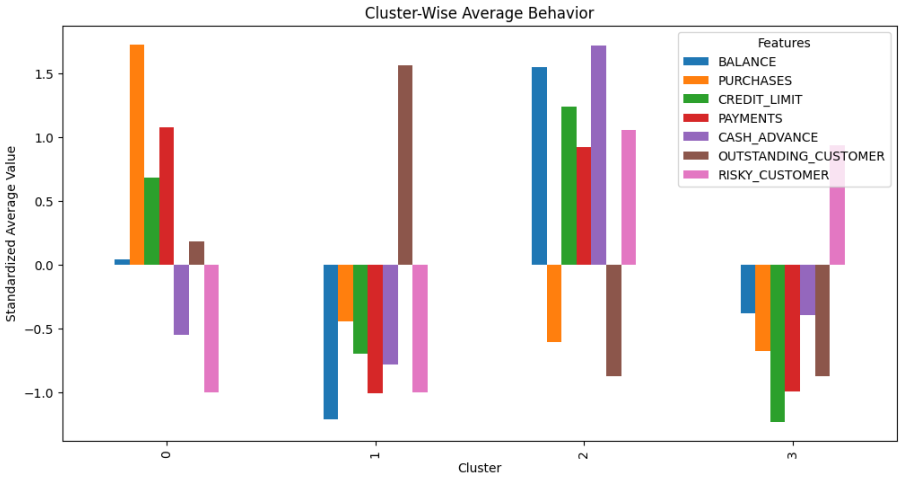

Figure 7: Cluster-Wise Average Behavior.

Figure 7 shows the Z-Score scaled average values of features for each cluster. Each bar corresponds to a feature, such as *Balance*, *Purchases*, *Payments*, etc. Clusters are segmented by how their average behaviors is different these metrics.

C Cluster Behavior Radar Chart

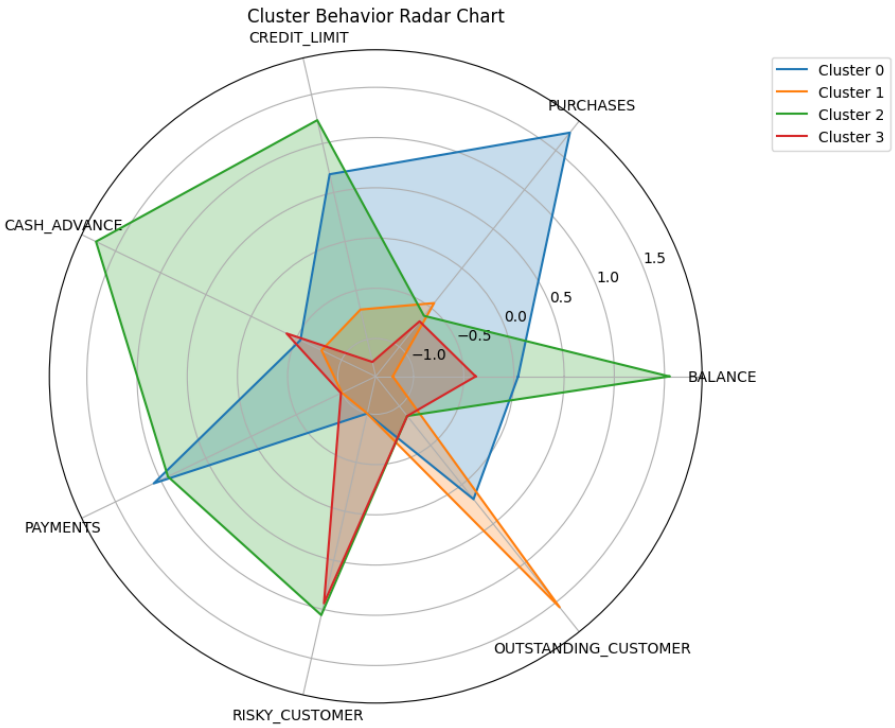

Figure 8: Cluster Behavior Radar Chart.

Figure 8 provides a radar chart representation of cluster behaviors. It clearly visualizes the importance of each feature for each cluster. For example, the chart shows how clusters are characterized by features like *Risky Customer*, *Outstanding Customer*, and others, which makes comparaison possible and derive segmentation of behavior for each clusters.