PW Institute Of Innovation

Project Report

Introduction to AI

Name – Hitendra Kumar Dewangan

Roll no - 2301010048

Topic – Unsupervised Learning

TABLE OF CONTENT

1.Introduction

1.1 Problem Statement

In the modern banking sector, understanding customer behavior is crucial for personalized marketing and customer retention. A leading bank wants to develop a **customer segmentation report** to offer targeted promotional campaigns to its customers.

To achieve this, the bank has collected a dataset summarizing user activities over the past few months, primarily focusing on **credit card usage patterns**. However, with a large and diverse customer base, manually segmenting customers is inefficient and impractical.

Thus, the challenge is to **identify distinct customer segments** based on spending behavior and usage patterns using **unsupervised machine learning techniques**. By clustering customers effectively, the bank can design **customized marketing strategies** that improve engagement and business performance.

1.2 Project Objectives

This project aims to leverage **clustering algorithms** to segment bank customers based on their transaction patterns.

2.Data Ingestion and Initial Checks

```
df = pd.read_csv('./Dataset/bank_marketing.csv')
```

- Columns and Data Types:
  - There are **7 columns** in total.
  - Each column has **210 non-null entries**, meaning there are no missing values in any of the columns.
  - All columns have the data type float64, which indicates that they contain floating-point numbers.
- Columns Overview:
  1. spending
  2. advance_payments
  3. probability_of_full_payment
  4. current_balance
  5. credit_limit
  6. min_payment_amt
  7. max_spent_in_single_shopping
- Implications:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   spending                      210 non-null    float64
 1   advance_payments              210 non-null    float64
 2   probability_of_full_payment   210 non-null    float64
 3   current_balance               210 non-null    float64
 4   credit_limit                  210 non-null    float64
 5   min_payment_amt               210 non-null    float64
 6   max_spent_in_single_shopping  210 non-null    float64
dtypes: float64(7)
memory usage: 11.6 KB
```

```
df.isnull().sum()
```

```
spending                        0
advance_payments                0
probability_of_full_payment     0
current_balance                 0
credit_limit                    0
min_payment_amt                 0
max_spent_in_single_shopping    0
dtype: int64
```

no null values

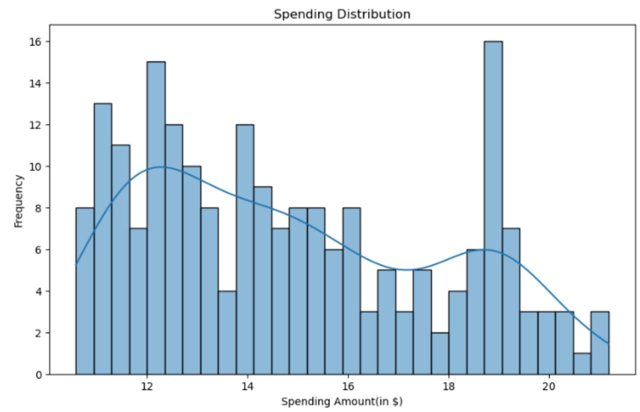
- Since all columns are non-null, you do not need to worry about handling missing data at this stage.
- The uniform data type (float64) suggests that all columns are suitable for numerical operations, which is beneficial for clustering analysis.

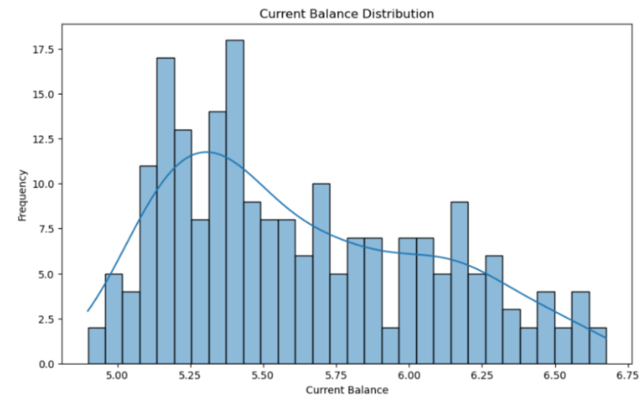
- Decision:
    - we will proceed with exploratory data analysis (EDA) to understand the distributions and relationships within these features.
    - We will scale them in order to use distance-based clustering algorithms like K-means, as they are sensitive to the scale of the data.

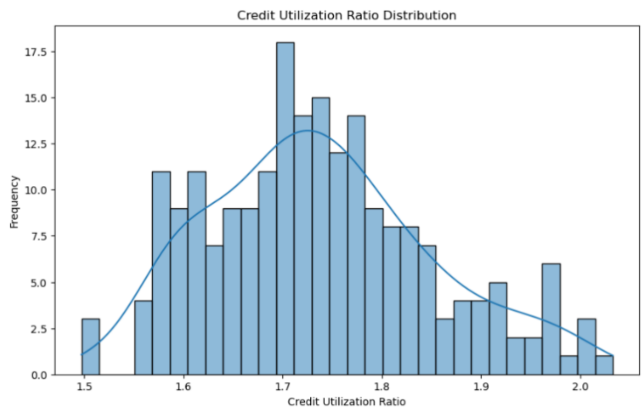
Reasons for Scaling:

- **Varying Scales:**
    - **Observe the ranges of the columns:**
        - **current_balance** and **credit_limit** have significantly larger values compared to probability_of_full_payment.
        - **min_payment_amt** and **max_spent_in_single_shopping** also have a wide range of values.
    - Clustering algorithms that rely on distance calculations (like Euclidean distance in K-means) will be heavily influenced by features with larger scales. Features with smaller scales will have a negligible impact on the distance calculations.
- **Preventing Bias:**
    - Without scaling, features with larger scales will dominate the clustering process, potentially leading to biased results.
- **Improved Convergence:**
    - Scaling can often improve the convergence speed of clustering algorithms.

3.Exploratory Data Analysis (EDA)







Analysis of Variable Distributions

1. Spending Distribution

- **Shape:** The spending distribution appears to be multi-modal, with at least two distinct peaks. This suggests that there may be two or more groups of customers with different spending behaviors.
- **Possible Insights:**
  - One group of customers spends relatively less, clustering around the 12-14 range.
  - Another group of customers spends more, clustering around the 18-20 range.
  - This could indicate different spending habits based on income, lifestyle, or credit card usage patterns.

2. Current Balance Distribution

- **Shape:** The current balance distribution seems to be approximately normal with a slight positive skew.
- **Possible Insights:**
    - Most customers have a current balance in the 5.0 to 6.0 range.
    - The positive skew suggests that there are some customers with significantly higher balances.

3. Credit Utilization Ratio Distribution

- **Shape:** The credit utilization ratio appears to be approximately normally distributed.
- **Possible Insights:**
    - The majority of customers have a credit utilization ratio between 1.6 and 1.8.
    - There are a few customers with much lower or higher ratios.

4.Data Preprocessing

1. Scaling Data

Given the multi-modal distribution of "spending" and approximate normal distribution of "credit utilization ratio", scaling is crucial before applying clustering algorithms.

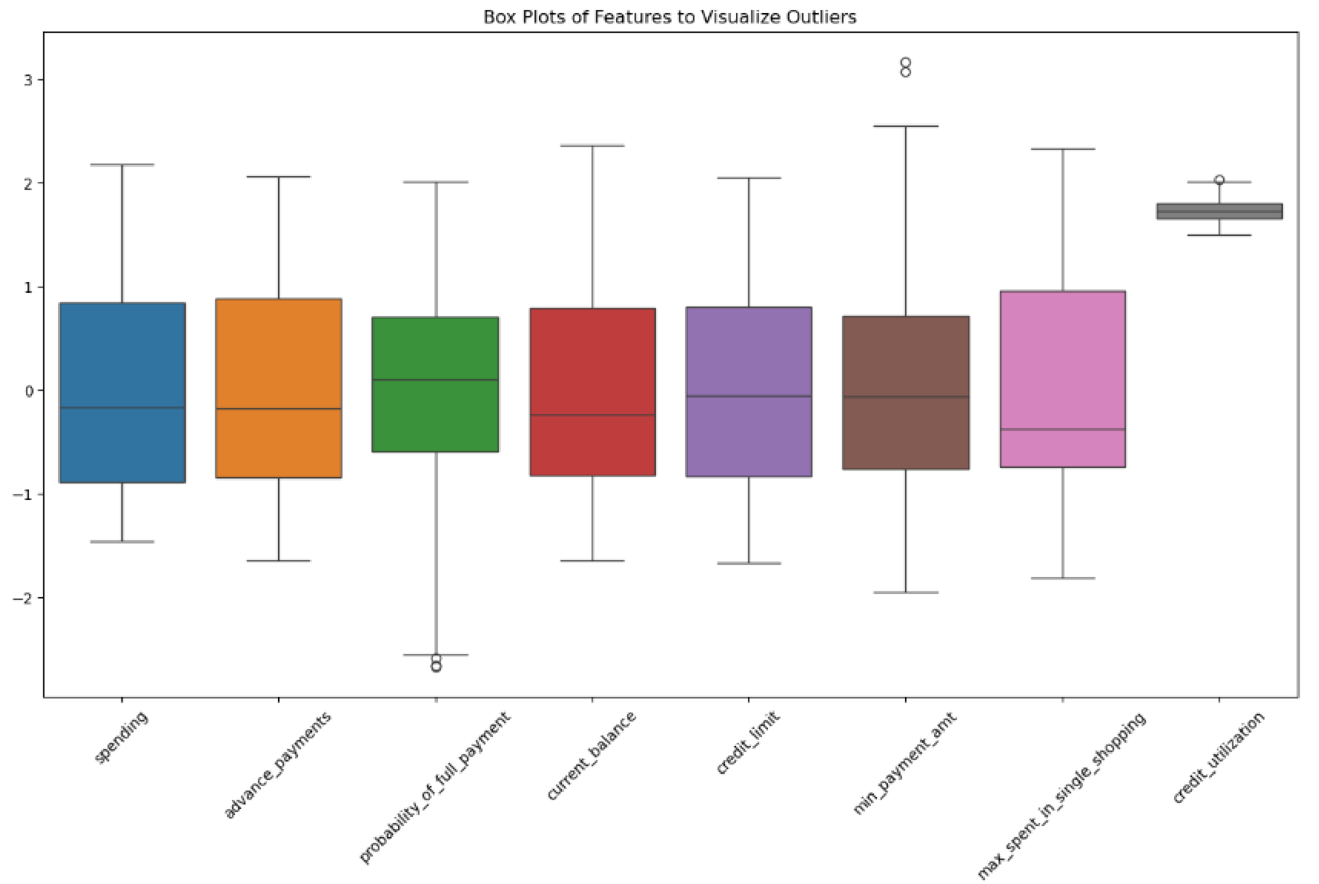
```
]: # Initialize the StandardScaler
scaler = StandardScaler()

# List of columns to scale
columns_to_scale = ['spending', 'advance_payments', 'probability_of_full_payment',
                    'current_balance', 'credit_limit', 'min_payment_amt',
                    'max_spent_in_single_shopping']

# Fit and transform the selected columns
df[columns_to_scale] = scaler.fit_transform(df[columns_to_scale])

# Display the first few rows of the scaled DataFrame
print(df.head())
```

2. Outlier Check and Treatment



We can see some outliers in columns:

- probability_of_full_payment
- min_payment_amt

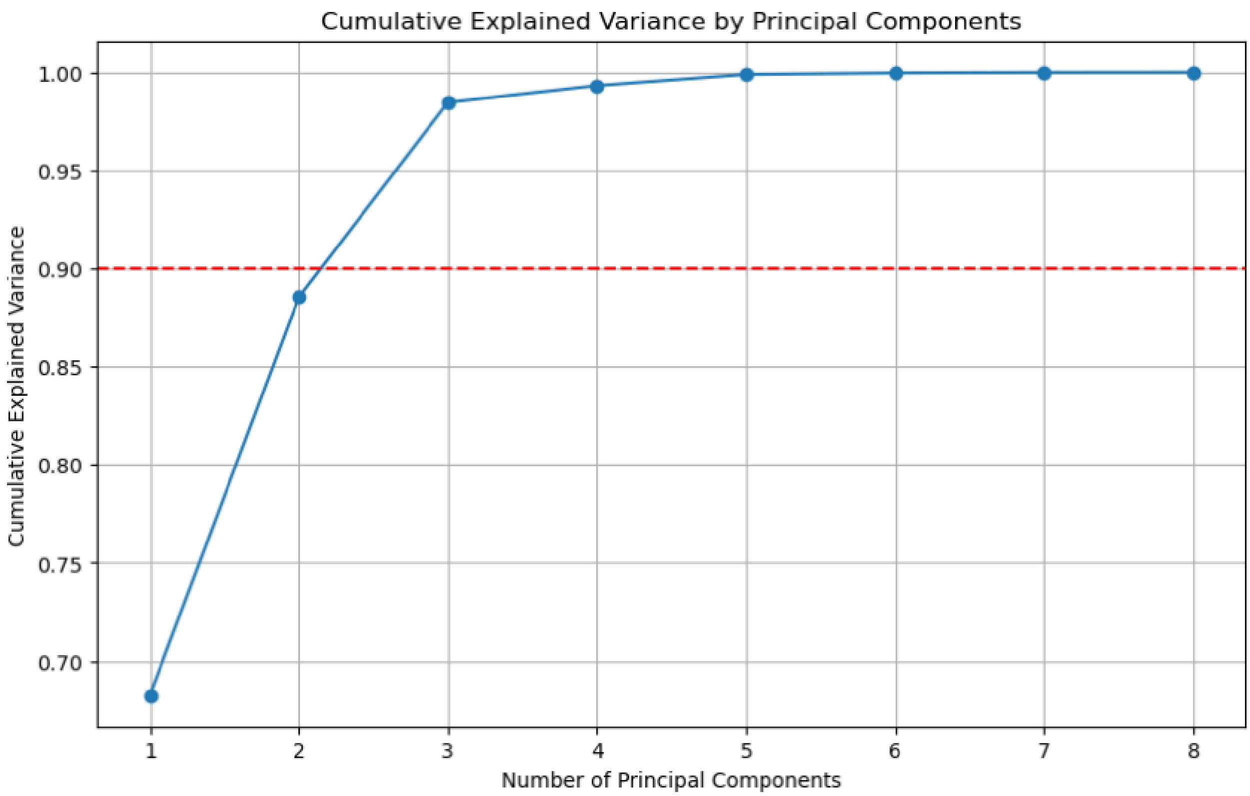
So we will treat the outlier.

```
cols_have_outlier = ['probability_of_full_payment', 'min_payment_amt']

def remove_outliers_iqr(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    df = df[(df[column] >= lower_bound) & (df[column] <= upper_bound)]
    return df

# Example: Removing outliers from 'spending' column
for col in cols_have_outlier:
    df = remove_outliers_iqr(df, col)
```

5.Dimensionality Reduction

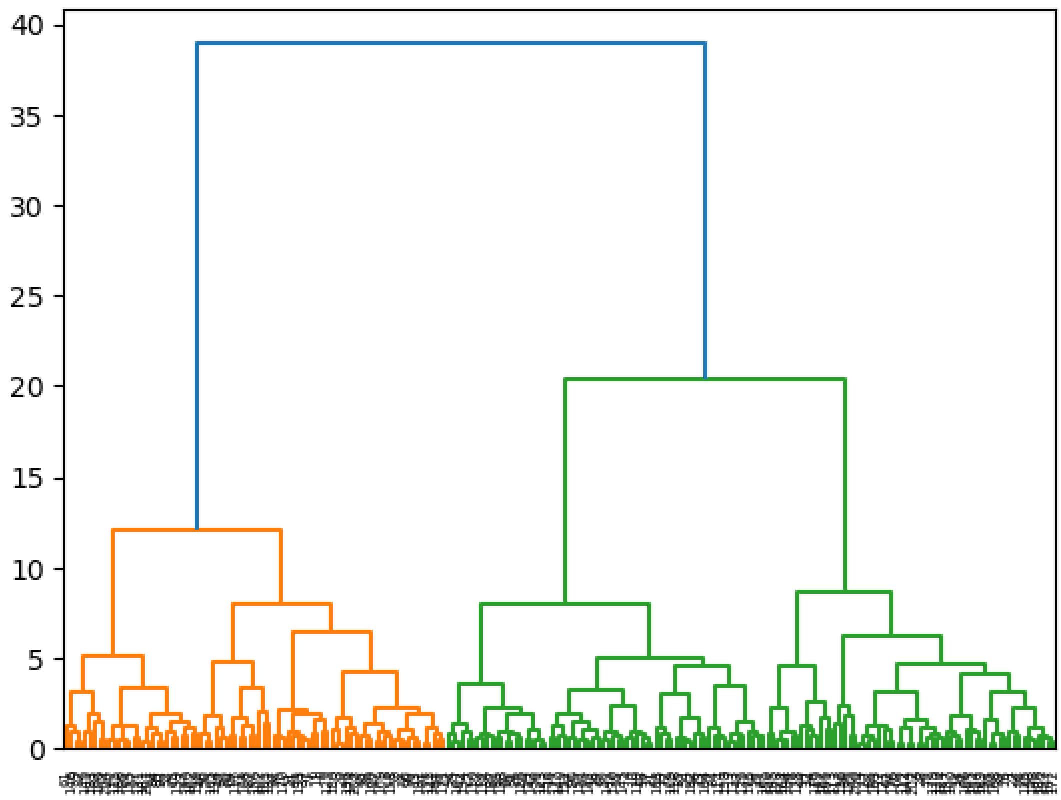


- It generates a plot showing the cumulative explained variance as a function of the number of principal components.
- The plot helps determine the number of components needed to capture a desired amount of variance (e.g., 90%).

Since at no. of cluster = 2, we can get 90% of the variance so we will be taking 2 clusters.

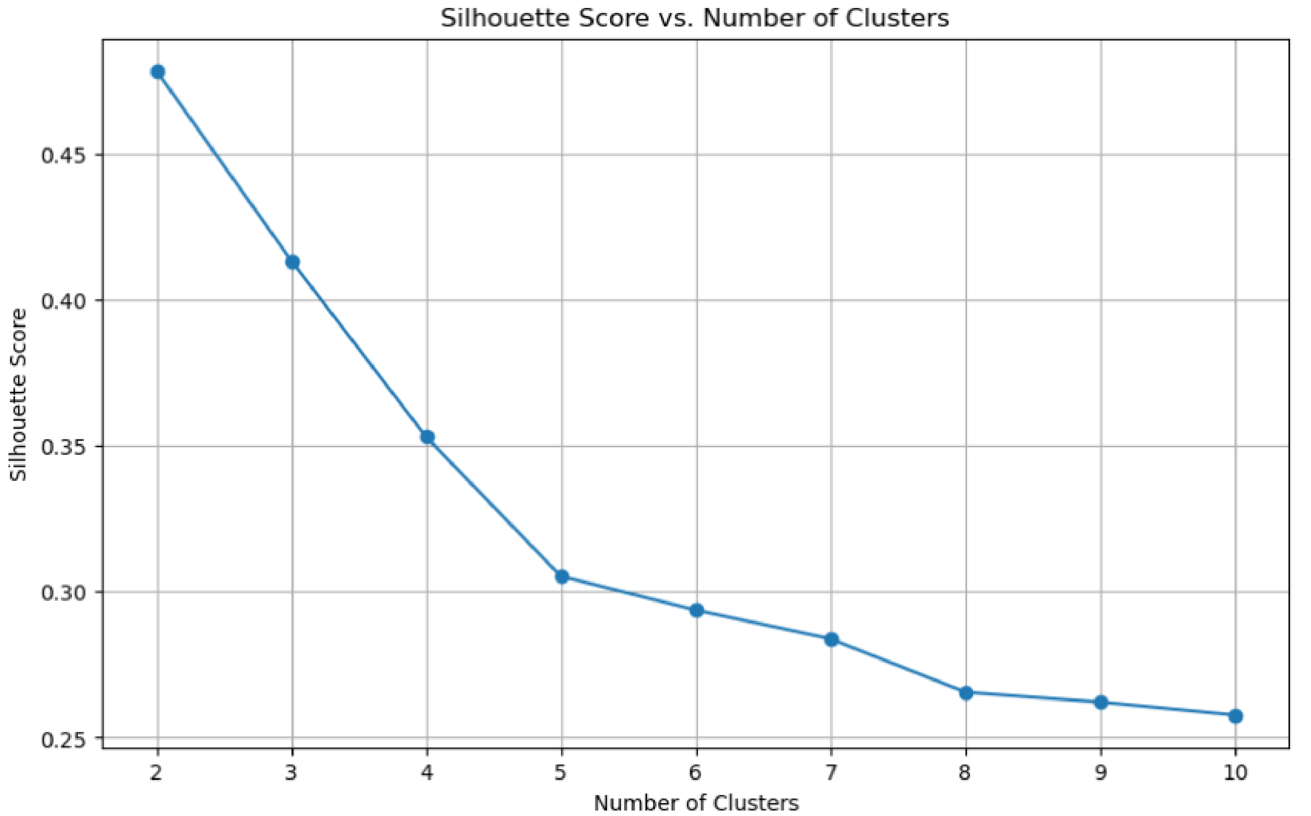
6.Clustering Implementation

- Hierarchical Clustering



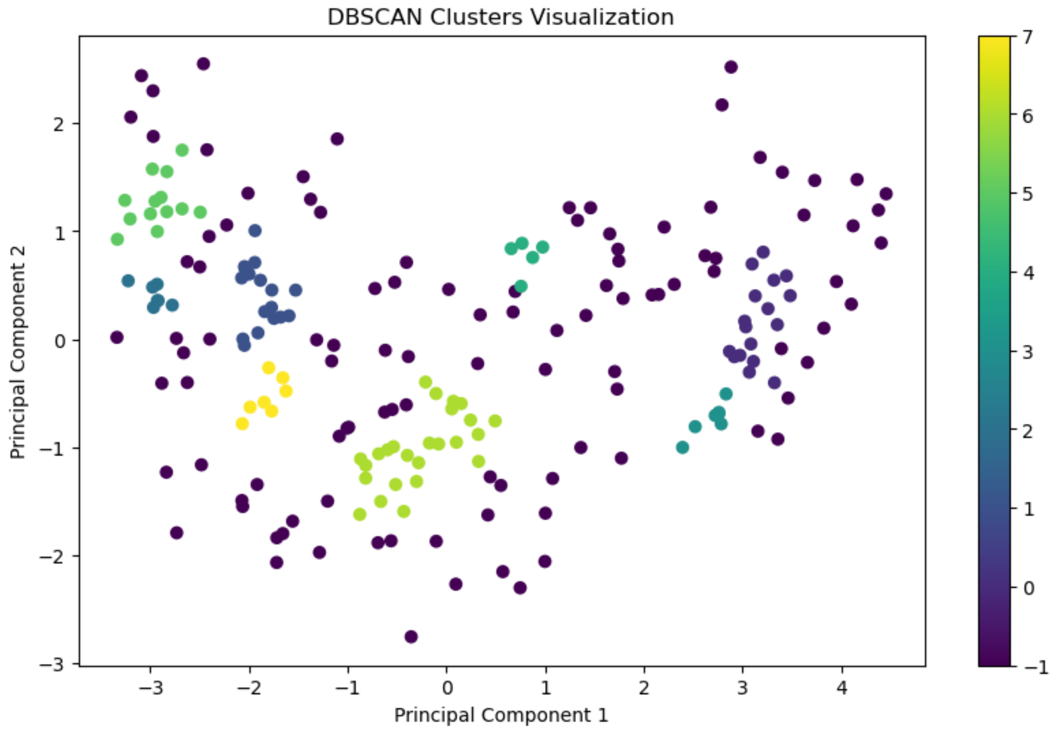
	PC1	PC2
0	4.118919	1.049561
1	0.551561	-1.352736
2	3.127651	0.400743
3	2.391835	-1.000901
4	-2.040791	0.658883

- K-Means Clustering



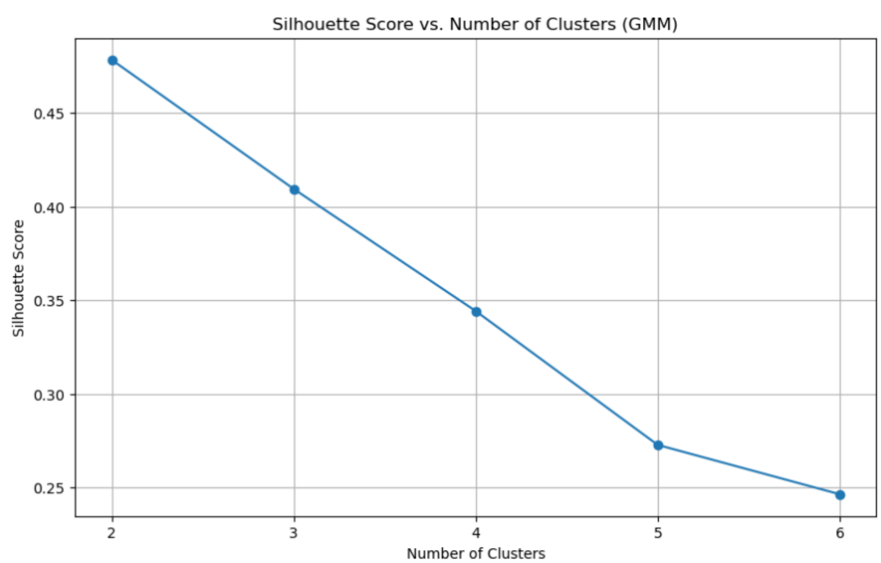
Optimal number of clusters: 2

- DBSCAN

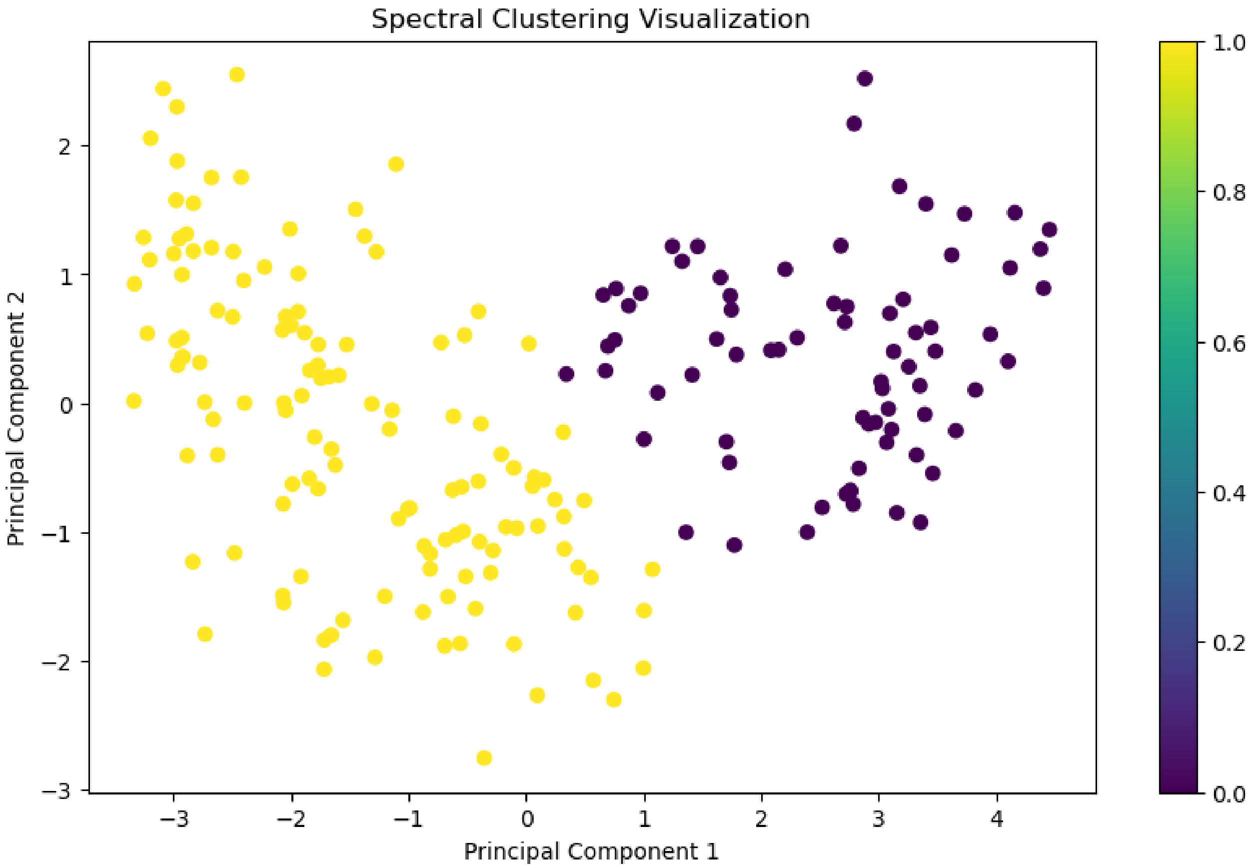


<u>DBSCAN is not performing well as we can see the clusterings are not good.</u>

- Gaussian Mixture Model



- Spectral Clustering



7.Cluster Analysis and Profiling

```
Silhouette Scores:
K-Means: 0.4638505436516679
Agglomerative: 0.4457724430785156
DBSCAN: -0.20993944209206925
GMM: 0.44034299935844456
Spectral: 0.46417904743633365

Best Performing Model: Spectral with Silhouette Score: 0.46417904743633365
```

```
cluster_profiles = df.groupby('cluster')[columns_to_scale].mean()
print(cluster_profiles)
```

```
         spending  advance_payments  probability_of_full_payment  \
cluster
0        1.179843          1.191724                     0.513648
1       -0.616923         -0.626983                    -0.234863

         current_balance  credit_limit  min_payment_amt  \
cluster
0               1.181133      1.090864        -0.067724
1              -0.625648     -0.564961        -0.026462

         max_spent_in_single_shopping
cluster
0                            1.229831
1                           -0.660521
```

So, we can see that the Spectral Clustering is the best model with a Silhouette Score of 0.464

8.Business Development Strategy

Cluster 0: "Value-Driven Customers"

Characteristics:

- These customers tend to have lower balances
- These customers have a moderate credit utilization ratio, indicating responsible usage of their credit lines.

Business Development Strategies:

1. **Personalized Rewards**:
    - Strategy:
      Offer personalized rewards tailored to customer preferences. The data suggests an openness to engagement, so tailor personalized experiences to them
2. **Customer Education**:
    - Strategy:
      Provide insights on financial management and responsibility and suggest how they can achieve higher risk ratings, and better scores to get better offers and credit limit increases.

Cluster 1: "High Credit User"

Characteristics:

- These customers tend to have High credit utilization
- Tend to have low max spending

Business Development Strategies:

1. **Credit Limit Increase**:
    - Strategy:
      Given their responsible credit behavior, consider offering modest credit limit increases and tailor promotions that will benefit the credit score and allow for better performance.
2. **Credit Card Upgrade Programs**:

- Strategy:
  In many banking environments, clients are eligible for premium cards and benefits that may not be available in their current plan. Incentivizing these opportunities may provide a better experience.
- Strategy:
  Encourage the use of services offered for greater client management

Conclusion

By implementing these targeted business development strategies, the bank can effectively cater to the unique needs and preferences of each customer segment. This data-driven approach will not only enhance customer satisfaction and loyalty but also drive sustainable growth and profitability for the organization.