Clustering for Bank Client Risk Management: An Interpretive Approach

Introduction

The primary objective of this project is to identify patterns and groups among bank clients to aid in risk management. Our focus is on creating interpretable clusters, even if it means sacrificing some performance and data points. We will evaluate the clusters using various metrics such as the Silhouette Score, Davies-Bouldin Index, and visual analyses through PCA, t-SNE, and pairwise plots. Additionally, we will examine the median and mean values of features within the clusters to gain an intuitive understanding. Importantly, our approach relies on machine learning without human intelligence influencing the decisions.

0. Packages and Importing

We begin by importing the necessary packages and datasets required for our analysis.

1. Explorative Data Analysis and Data Cleaning

We start with exploratory data analysis (EDA) by plotting all variables to understand their distributions and determine the appropriate methods to apply. We inspect the data types and search for NaN values. Column names are stripped of any blank spaces, and the ID column is dropped to prevent it from influencing the clustering process.

For the 'Sex' column, we rename it to 'Female' and code it as 1 for female and 0 for male, creating an intuitive dummy variable. The 'Education' column is simplified by dropping category 4, which is labeled 'Other,' due to its low representation and interpretability issues. We handle ambiguous values (0, 5, 6) similarly by excluding them, reducing dimensionality for better interpretability.

The 'Status' column is renamed to 'Marital Status' and coded as 1 for married and 0 for not married, with category 3 ('Other') being dropped for the same reasons as above. The 'Age' column is retained but checked for outliers and negative values. For the columns 'Previous 1-6', 'Repayment 1-6', and 'Amount 1-6', we use histograms to identify distributions, noting that mostly no normal distribution exists. No data cleaning is performed here, as we aim to identify risk groups, and removing extreme values could be counterproductive.

2. Preprocessing

Next, we create new variables to reduce dimensionality and increase interpretability. For 'Amount 1-6', we calculate their average and create 'Average_Account_Balance'. We consider discounting more recent values but decide against it to avoid imposing assumptions. Similarly, we create 'Median_Repayment_Status' using the median due to the ordinal scale, but also consider the mean for robustness. 'Average_Repayment_Amount' is generated with the mean. Although adding standard deviation could be beneficial, it was found to decrease model quality, so we exclude it.

We introduce 'total_repayment_late', summing up months of late repayments, providing a comprehensive view of repayment behavior. We checked for the opposite as a robustness check with Total_repayment_ontime. Categorical values are converted into dummy variables using One-Hot Encoding. The original columns of 'Previous', 'Repayment', and 'Amount' are dropped to reduce dimensionality further, despite seeing better silhouette scores with their inclusion. Consistent scaling is applied using the RobustScaler due to skewed and non-normal distributions. The elbow method helps determine the optimal number of clusters, which is three in our case, though we explore 2-8 clusters for robustness.

3. Running the Model

We implement K-means clustering with three clusters, using random initialization and a maximum of 600 iterations. The centroid method proves reliable, and switching to k-means++ shows no significant change. The Elkan algorithm is used for its efficiency, despite higher memory usage instead of Lloyd.

Visual analyses through PCA and t-SNE provide insights into cluster separations. Additionally, we apply DBSCAN with grid search to find the best specifications. DBSCAN identifies more but smaller clusters and detects noise, enhancing our understanding. More Robustness checks will be applied by comparing the models.

4. Plotting the Results

Pairwise plots for both K-means and DBSCAN (final specification pending) reveal clear and interpretable clusters. The visual separation of clusters underscores our focus on interpretability.

5. Model Diagnostics

We evaluate model diagnostics by examining scores and robustness. Changing specifications from median to mean has minimal impact, indicating model stability. However, results are sensitive to scaling methods, with StandardScaler significantly degrading performance. Median and mean values of unscaled clusters highlight distinct risk groups: average, low risk, and high risk. The next step involves industry experts interpreting these clusters for practical applications.

Conclusion

This project effectively identifies interpretable clusters among bank clients, providing valuable insights for risk management. While performance trade-offs exist, the focus on interpretability ensures practical utility. Future steps involve expert validation and potential refinement based on industry feedback.