Research Project for Machine Learning: Unknown Groups of Bank Clients

Sarthak Saxena	(7717478)
Jonathan Teja	(7717052)
Munkhsaruul Tumendemberel	(7212775)
Lukas Barbutev	(7716854)
Teruhito Nakajima	(7721971)

1. Introduction

In this assignment, we use clustering techniques intending to help the bank mitigate or manage its risk by using the clustering algorithms that help identify unknown patterns amongst the bank clients by partitioning the dataset into distinct, non-overlapping subsets or clusters.

1.1. Dataset Description

The dataset that will be used contains historical payment behavior and some demographic features for each customer, namely credit, gender, education level, marital status, age, repayment status, account balance, and previous repayment in the corresponding month. The following table below shows the structure of the dataset.

Table 1. Dataset Overview

Variable Name	Description	Type of Data	Sample Data
ID	Customer ID	Categorical	1,2,3, . . .
Credit	Amount of given credit	Continuous	120000, 90000, . . .
Sex	Gender	Categorical	1 = male, 2 = female
Education	Education level of the customer	Categorical	1 = graduate school, 2 = university, 3 = high school, 4 = others
Status	Marital status	Categorical	1 = married, 2 = single, 3 = others
Age	Age in years	Continuous	23, 32, 46, . . .
Repayment 1 - 6	Repayment status in January to June (Repayment 1 for January, Repayment 2 for February, . . .)	Categorical	-2 = no credit consumption, -1 = repayed duly, 0 = use of revolving credit, 1 = payment delay of 1 month, 2 = payment delay of 2 months, . . . , 8 = payment delay of 8 months, 9 = payment delay of equal or more than 9 months
Amount 1 - 6	Amount of the account balance from January to June (Amount 1 for January, Amount 2 for February, . . .)	Continuous	1725, 14027, 13021, . . .
Previous 1 - 6	Amount of the previous repayment in the corresponding month (Previous 1 for January, Previous 2 for February, . . .)	Continuous	1000, 2000, 1542, . . .

1.2. Clustering Algorithms

The clustering algorithms that will be used in this project are partition-based clustering (K-means) and density-based clustering (DBSCAN). We will compare both results and then pick which one is more representative.

2. Data Cleaning and Data Preprocessing

Data preprocessing is an integral step in as the quality of data and the useful information that can be derived from it directly affects our result. We follow the procedure listed below in order to ensure that our data is clean and allow for our model to perform effectively.

2.1. Data Cleaning

- **Null Values Check:** Upon thorough review, it was confirmed that the dataset contains no null values.
- **Duplicate Check:** No duplicate entries were found in the dataset.
- **Handling Missing Values:** As there were no missing values, no further action was required in this regard.
- **Data Manipulation:** The dataset initially included 345 records with education levels of 5 and 6, and 54 records with a status value of 0. Since the education levels 5 and 6, as well as the status value of 0, are undefined categories, these entries were reclassified into an "Others" category for consistency and clarity.

2.2. Feature Engineering

We created new columns based on payment behavior by using Amount 1 – 6, Previous 1 – 6, and Repayment 1 – 6, namely Average Amount, Average Previous Repayments, and Mode Repayment Status. The following table below shows the structure of new features we newly created

Table 2. Overview of Newly Features Added

Variable Name	Description	Type of Data	Sample Data
Average Amount	The average amount of the account balance from January to June per customers	Continuous	2845.17, 16942.2, 3855.7, . . .
Average Previous Repayments	The average amount of the previous repayments from January to June per customers	Continuous	833.3, 1398, 772.6, . . .
Mode Repayment Status	The frequency status that appears the most in repayment status per customers	Categorical	-2 = no credit consumption, -1 = repayed duly, 0 = use of revolving credit, 1 = payment delay of 1 month, 2 = payment delay of 2 months, . . . , 8 = payment delay of 8 months, 9 = payment delay of equal or more than 9 months

3. Explorative Data Analysis (EDA)

This section will perform some EDA to get more insights about dataset. We divided this section by two subsections, which are EDA for categorical variables and EDA for continuous variables.

3.1. EDA for Categorical Variables

3.1.1 The Composition of Gender, Education, and Marital Status

The chart illustrates that there are more females than males in the dataset. Approximately 18,000 individuals are female, while around 12,000 are male. This indicates a higher representation of females.

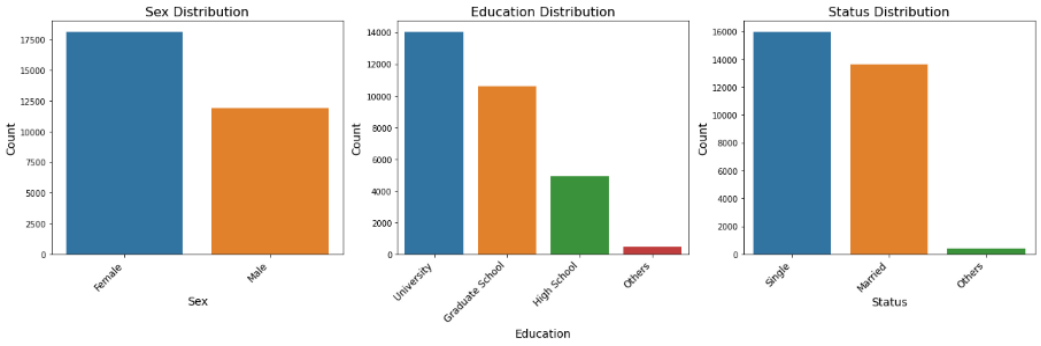


Figure 1. Gender, Education, and Status Composition

The second shows that the majority of individuals have a university education, followed by graduate school. High school graduates form a smaller group, and the "Others" category is minimal. This suggests that the dataset predominantly consists of individuals with higher educational qualifications. The third chart reveals that single individuals are the most prevalent, closely followed by married individuals. The "Others" category is minimal, indicating a nearly balanced representation of single and married individuals.

3.1.2 The Composition of Repayment Status

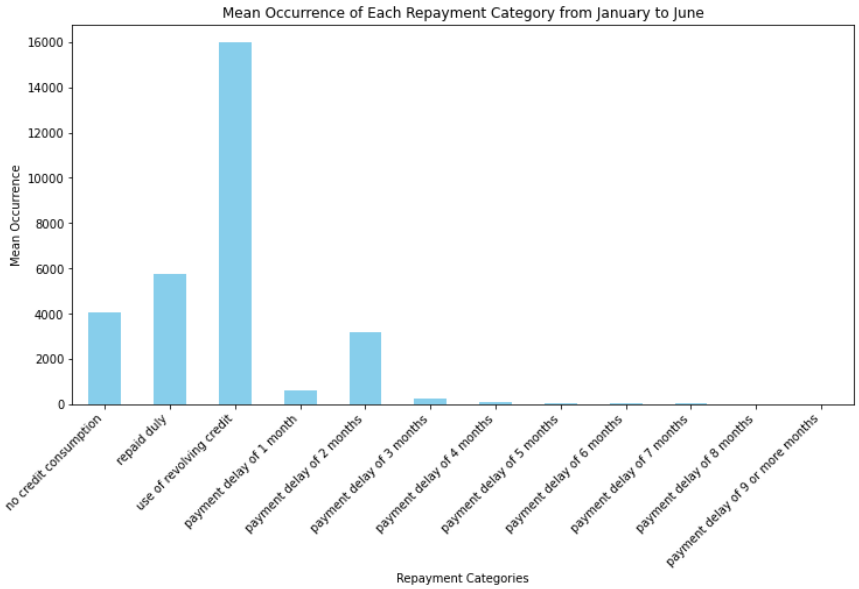


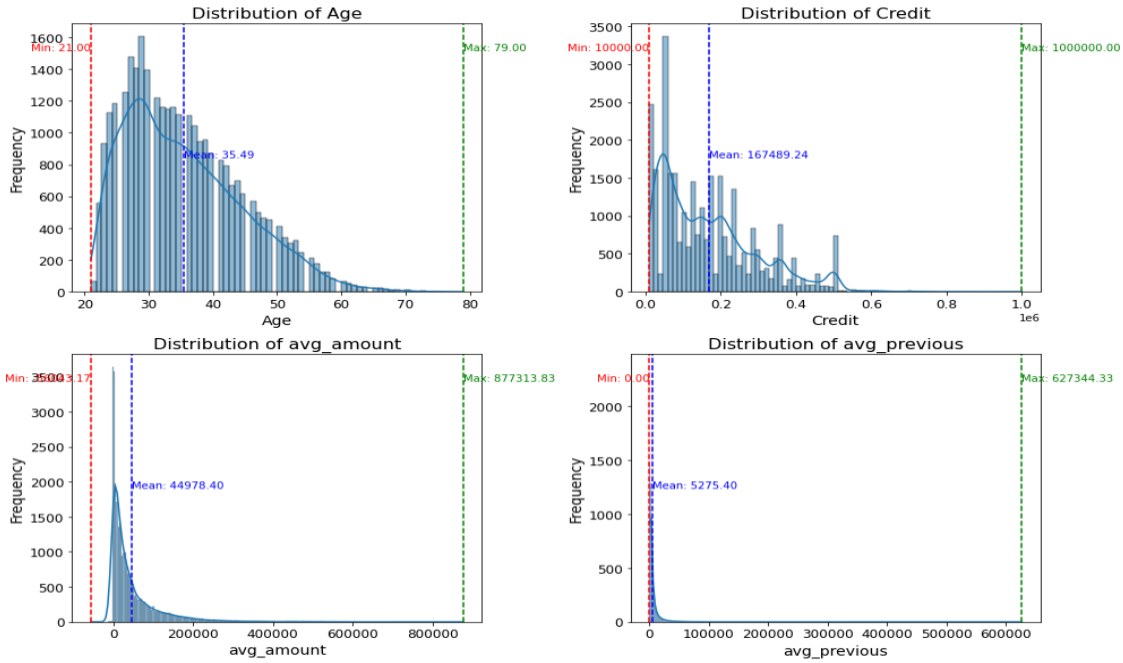
Figure 2. Repayment Status Composition

Since there were no significant differences found in the distribution of repayment statuses across each month, the mean status of repayment from January to June was analyzed to gain deeper insights. Out of approximately 16,000 customers who used credit, 5,800 repaid duly. Of those who experienced delays, totaling 4,200, the largest subgroup consisted of customers with a two-month payment delay, totaling 3,160. The second largest subgroup, with 620 customers, had a one-month delay, followed by 238 customers with a three-month delay. On average roughly 150 customers were delayed for more than three months. Additionally, around 4,070 customers did not use credit at all. Since the average was taken the numbers of customers do not add up to 30000 and are more to give a rough overview.

3.2. EDA for Continuous Variables

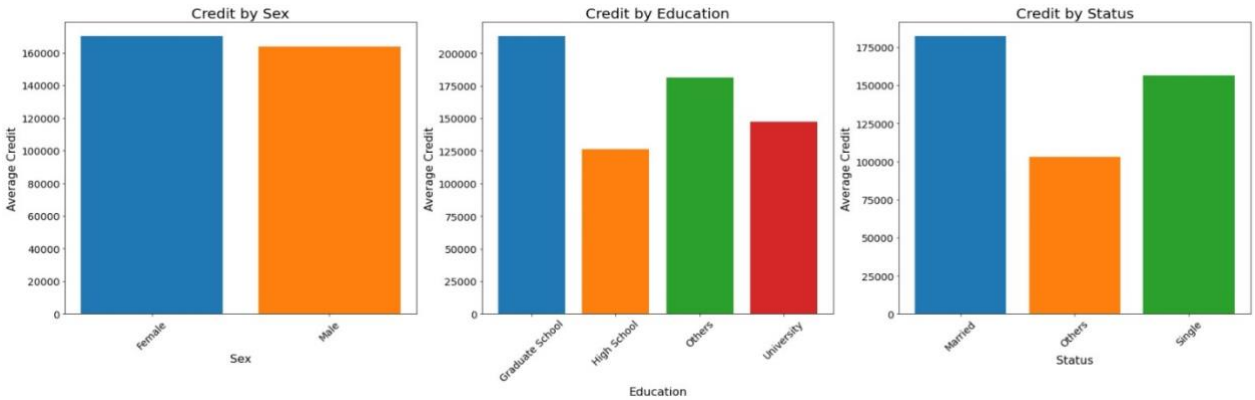
3.2.1 The Distribution of Age, Credit, the Average Amount of the Account Balance, and the Average Amount of the Previous Repayments

The charts below examine the continuous variables: Age, Credit, Average Amount, and Average Previous. The Age variable ranges from 21 to 79 year-old. However, the data shows a higher concentration of younger individuals, with a mean age of 35.49 year-old. The Credit variable also shows a highly skewed distribution towards lower values. The minimum credit amount is 10,000, and the maximum reaches 1,000,000, with a mean of 167,489.24. This indicates that while most individuals have lower credit amounts, there are a few outliers with significantly higher credit.



We created the variables Average Amount and Average Previous by averaging Amount 1-6 and Previous 1-6, respectively. The distributions of these variables are also highly skewed, with most individuals clustered around very low values. The presence of outliers in both variables is evident and the majority of the data points are concentrated in a narrow range.

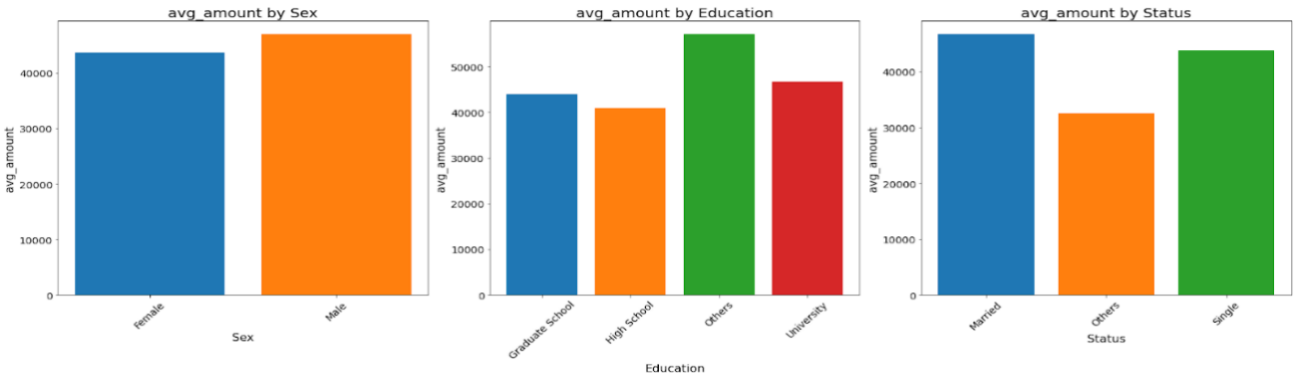
3.2.2 The Distribution of Credit Grouped by Gender, Education, and Marital Status



The average credit for both females and males is very similar, with females slightly leading. In the middle graph, Individuals with a Graduate School education have the highest average credit. Those with a High School education have the lowest average credit. “Others” and “University” categories fall in between, with “Others” having a higher average than “University”. This indicates that higher education levels might correlate with higher average credit. Married individuals have the highest average credit.

Single individuals have a moderately high average credit. The “Others” category has the lowest average credit. Marital status seems to have a noticeable impact on average credit, with married individuals being favored.

3.2.3 The Distribution of the Average Amount of the Account Balance Grouped by Gender, Education, and Marital Status



The average amount is slightly higher for males compared to females. Gender appears to have a minimal effect on the average amount. The “Others” category leads with the highest average amount. Graduate School, High School, and University categories follow, with High School having the lowest average amount. This pattern suggests a varied influence of education level on the average amount, with non-traditional or unspecified education (“Others”) having the most significant impact. Married individuals have the highest average amount, similar to their credit pattern. Single individuals also have a high average amount, but less than married individuals. The “Others” category has the lowest average amount. Marital status again plays a significant role, with married individuals receiving higher average amounts.

4. Model Fitting

4.1. Features Scaling

Since our dataset has huge variability and skew, we need to transform it into a defined range or scale to produce high-quality clusters and boost the precision of clustering algorithms. For this purpose, we use robust scaler to make our dataset robust to outliers.

4.1.1 K-Means

K-means clustering works by classifying a given data set into a number of clusters, defined by the letter "k", which is fixed beforehand. The clusters are then positioned as points and all observations or data points are associated with the nearest cluster, computed, adjusted and then the process starts over using the new adjustments until a desired result is reached.

The standard K-means algorithm is not directly applicable to categorical data as the sample space for categorical data is discrete, and does not have a natural origin, making Euclidean distance function on such a space is not really meaningful. Hence, we only use continuous data when fitting using K-Means. We construct two different clustering modelling for K-Means.

The first one, we select only two features for fitting the K-means, namely Average Amount and Average Previous Repayments. The reason is because Average Amount and Average Previous Repayments directly represent the financial behaviour of customers. Including features like age and credit might introduce noise because these variables can be influenced by various external factors not directly related to spending and repayment behaviours. By focusing on Average Amount and Average Previous Repayments, we reduce the potential for such noise. The second one, we select all continuous variables excluding Average Amount and Average Previous Repayments, and then to reduce the dimensions while preserving as much information as possible, we applied PCA beforehand.

a) Determining the Optimal Number of Clusters

Before implementing K-Means, the first step is to calculate the optimal number of clusters using the Elbow Method and Silhouette Score.

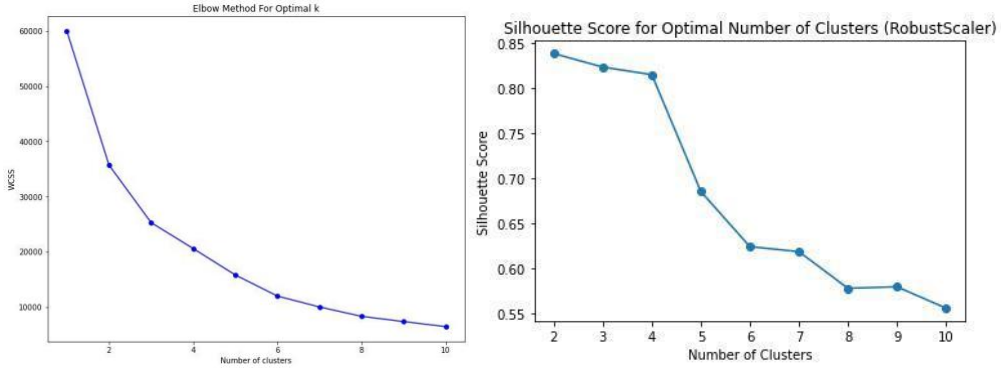


Figure 3. Optimal Clusters

For determining the optimal number of clusters of K-Means using two features, we use the Elbow Method. As shown by the figure above, the optimal cluster when using two features is three. We use Silhouette Score when finding the optimal number of clusters when using PCA. The result shows that the number of optimal clusters is two for PCA.

b) Clustering Result and Quality Assessment

The figure above illustrates cluster tendencies in the scatter plot. Silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1. The higher the score, the better. 1 means clusters are well apart from each other and clearly distinguished. 0 means clusters are indifferent or the distance between clusters is not significant. -1 means clusters are assigned in the wrong way.

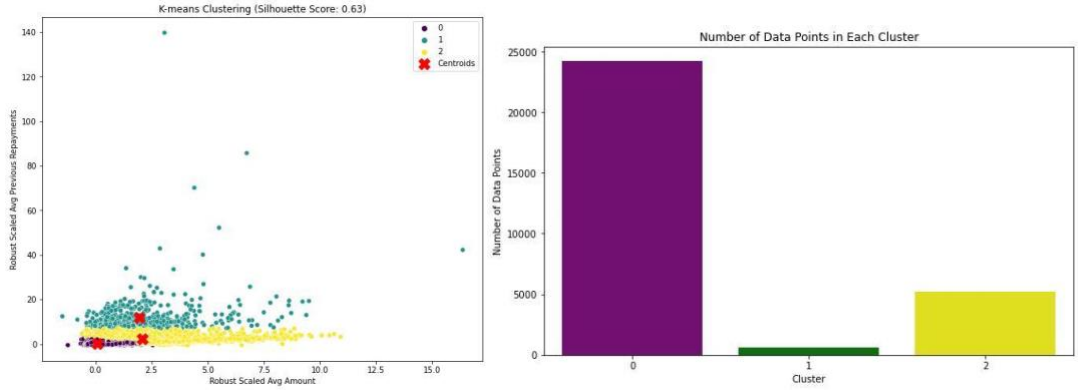


Figure 4. K-Means Clustering Result Using Two Features (Average Amount and Average Previous Repayments) and The Number of Clusters

Based on the silhouette score, it indicates that the clusters are optimal. Cluster 0 apparently has the thickest density and consistency compared to Cluster 1 and Cluster 2.

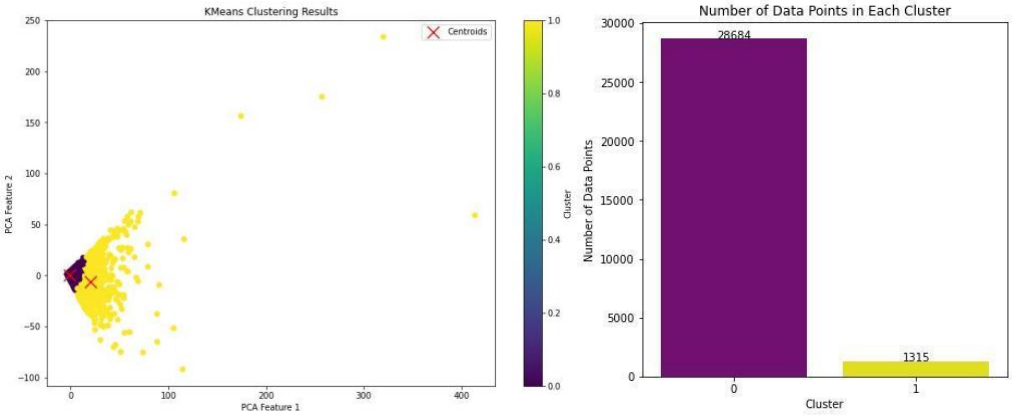


Figure 5. K-Means Clustering Result Using PCA and The Number of Clusters

Applying PCA, the Silhouette Score is boosted to 0.838, indicating the clusters made using PCA are optimal and better than only using two features. Cluster 0 apparently has the thickest density and consistency compared to Cluster 1.

4.1.2 DBSCAN

The DBSCAN algorithm works by dividing regions with a certain density into clusters and effectively characterizing arbitrary shapes of clusters from noisy spatial datasets. To work with DBSCAN, an ϵ-neighborhood, or simply a radius of a data point, has to be specified first. Afterwards, the minimum points, which is the minimum number of core points that fall within epsilon, are set. The clusters will then grow and broadcast their own perimeter until they reach border points. When no more data points can be reached, the cluster is finalized. This process is reiterated until there are no remaining data points left.

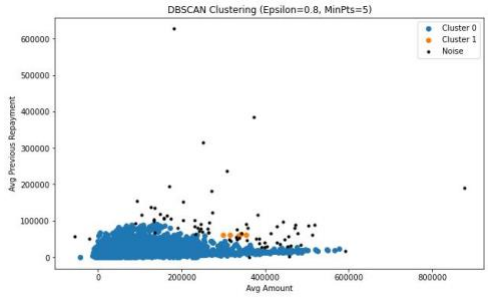


Figure 6. DBSCAN Clustering Using Two Features

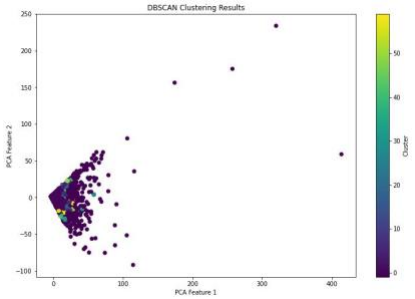


Figure 7. DBSCAN Clustering Using PCA

Based on the results above, DBSCAN is not reliable, as it does not demonstrate the clustering tendency within the data as effectively as the K-means algorithm. The tendency is almost the same across different epsilons and minimum points, providing no clear clusters.

5. Model Diagnostics

In this section, an evaluation of the quality of the clustering results from the algorithm that has been made will be carried out. This section will compare each clustering algorithm's clustering results.

Table 3. Overall Clustering Quality Assessment

Algorithm	Index	Value
K-Means with Two Features	Silhouette Score	0.63
K-Means with PCA	Silhouette Score	0.838
DBSCAN with Two Features	Silhouette Score	0.835
DBSCAN with PCA	Silhouette Score	0.665

K-Means with PCA has the highest Silhouette Score, followed by DBSCAN with two features, DBSCAN with PCA, and lastly K-Means with Two Features. However, the scores are not so much difference across features for every algorithm and DBSCAN does not provide meaningful clustering. Therefore, the K-Means algorithm is used.

6. Plotting and Discussion

6.1. Overall Clustering Visualization

6.1.1. K-means using Average Amount vs Average Previous Repayment

Figure 8. Visualization of All Variables in Each Cluster by using Two Features

Based on the figure above, when we fit the K-Means using two features, it can be seen that there is no clear clustering based on Age as the clusters seem very overlapping. In Credit, there seems better separation. We see Credit versus Average Amount and Credit versus Average Previous could be potential clustering features when we want to keep the algorithm simple.

We can see clear and meaningful clustering in Average Amount versus Average Previous Repayments. It can be concluded that, the more account balance clients have, the more likely they are going to repay. The plots involving the categorical variables (Sex, Education, Status, Mode Repayment Status) are not clear for interpretation since they have many overlapping points.

6.1.2. K-means using PCA

On the figure below, when we fit the K-Means by using PCA, we see a similar tendency compared to the Figure 8. Average Previous Repayment vs Credit, Average Previous Repayment vs Age, and Average Previous Repayment vs Average Amount seems to cluster reasonably.

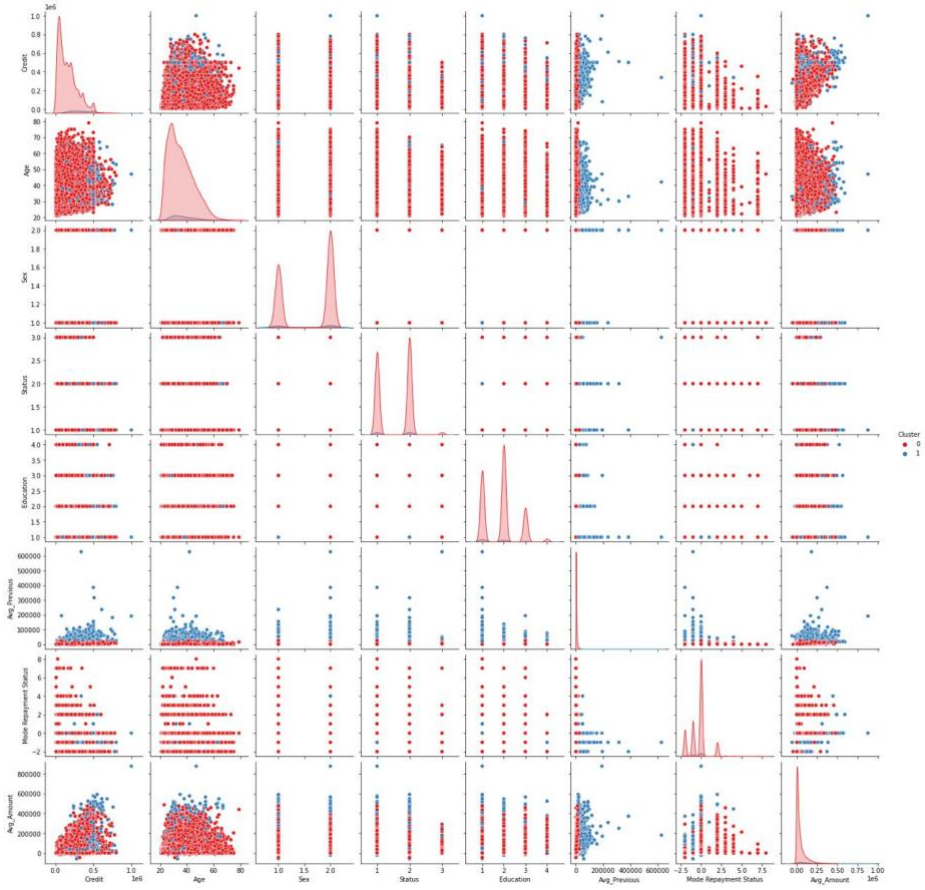


Figure 9. Visualization of All Variables in Each Cluster by using PCA

6.2. Cluster Profiling

6.2.1. K-means using Average Amount vs Average Previous Repayment

The figure below shows that Cluster 1 is more likely to repay than Cluster 2 and Cluster 0. However, Cluster 1 has lower account balance compared to Cluster 2. In Cluster 0, the average previous repayment is very low and the lowest among other clusters. Therefore, we can classify Cluster 0 as high-risk clients, Cluster 1 as low-risk clients, and Cluster 2 as medium-risk clients.

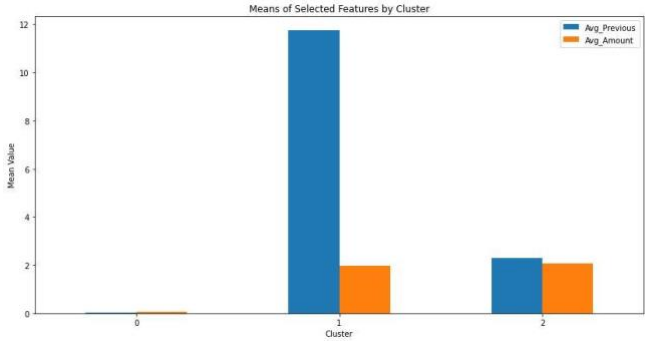


Figure 10. Means of Selected Features (Average Amount and Average Previous Repayments)

The boxplots below show that Cluster 1 has the highest amount of credit given, highest average previous repayment, and older age compared to the other two clusters. For Cluster 0, it has the lowest amount credit given, lowest average previous repayment, lowest account balance, and younger age. For Cluster 2, it has the highest amount of account balance among other clusters.

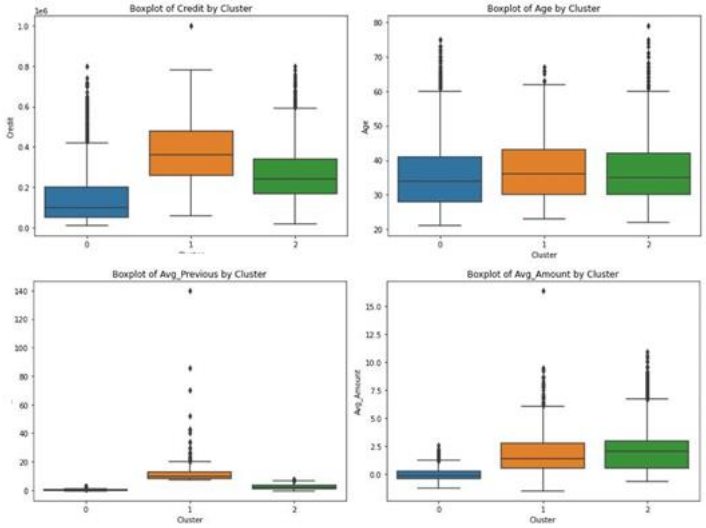


Figure 11. Boxplots for All Variables using Two Features by Cluster

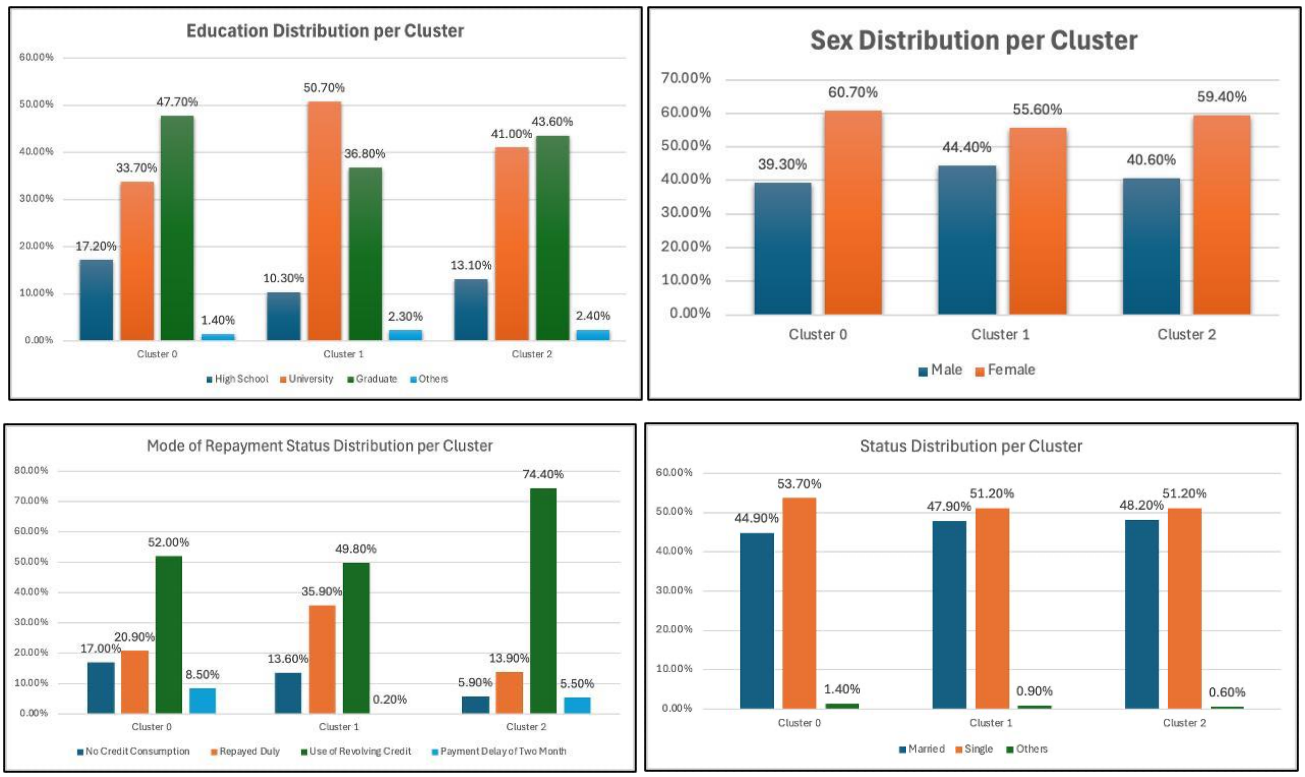


Figure 12. Distribution per Cluster

Based on the clustering analysis using only two features, we observed distinct patterns in the composition of the clusters to gender, education, marital status, and repayment status. These findings are summarized below:

Gender Composition:

- Cluster 0 (High-risk): Females have the highest composition, whereas males have the lowest.
- Cluster 1 (Low-risk): Males have the highest composition, whereas females have the lowest.

Education Composition:

- Cluster 0 (High-risk): Individuals with graduate-level and high school education have the highest composition.
- Cluster 1 (Low-risk): Individuals with university-level education have the highest composition.
- Cluster 2 (Medium-risk): Individuals categorized as 'Other' have the highest composition.

Marital Status Composition:

- Cluster 0 (High-risk): Married individuals have the highest composition.
- Cluster 1 (Low-risk): Married individuals have equal composition as in Cluster 2 (Medium-risk).
- Cluster 2 (Medium-risk): Single individuals have the highest composition, whereas 'Others' have the lowest.

Repayment Status Composition:

- Cluster 0 (High-risk): Individuals with no credit consumption and those with a payment delay of two months have the highest composition.
- Cluster 1 (Low-risk): Individuals who repaid duly have the highest composition.
- Cluster 2 (Medium-risk): Individuals using revolving credit have the highest composition.

6.2.2. K-means using PCA

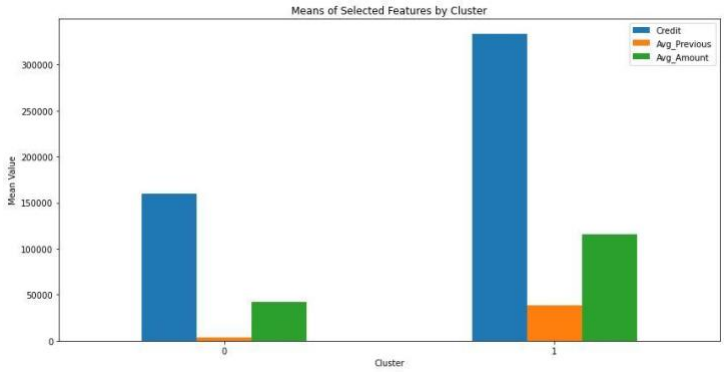


Figure 13. Means of Selected Features by PCA

The figure below shows that Cluster 1 is more likely to repay than Cluster 0 as the mean of Average Previous Repayment for Cluster 1 is higher than Cluster 0. Additionally, Cluster 1 also has a higher amount of account balance and amount of credit given. Therefore, we can classify Cluster 0 as a high-risk client and Cluster 1 as low-risk client.

The boxplots below show that Cluster 1 has the higher amount of credit given, higher average amount of account balance, higher amount of average previous repayments, and older age compared to Cluster 2.

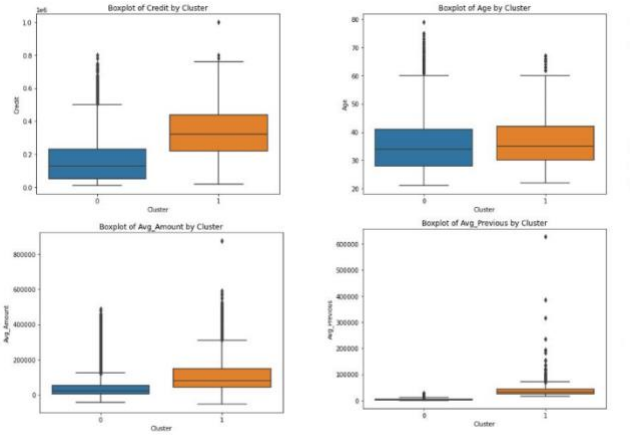


Figure 14. Boxplots for All Variables using PCA by Cluster

From the below graphs, we can conclude the following:

Gender Composition:

- Cluster 0 (High-risk): Females have a higher composition compared to males.
- Cluster 1 (Low-risk): Males have a higher composition compared to females.

Education Composition:

- Cluster 0 (High-risk): Higher composition of individuals with a graduate-level education and high school education.
- Cluster 1 (Low-risk): Higher composition of individuals with a university-level education and those categorized as 'Other.'

Figure 12. Distribution per Cluster

Marital Status Composition:

- Cluster 0 (High-risk): Single individuals have a higher composition. Interestingly, those classified as 'Others' have a higher composition in Cluster 0 (Low-Risk).
- Cluster 1 (Low risk): Married individuals have a higher composition, though the percentage difference is not significant.

This pattern differs from our earlier two-variable analysis, where marital status showed a different distribution.

Repayment Status Composition:

- Cluster 0 (High-risk): Individuals with no credit consumption have a higher composition.
- Cluster 1 (Low risk): Those who repaid their credit duly have a higher composition.

In summary, the PCA analysis provides a nuanced understanding of the composition of high-risk and low-risk clusters, particularly in terms of gender and education, which aligns with our initial findings with the two variables. However, marital status shows a slight deviation from the previous finding.

7. Conclusion

7.1.1 K-means using Average Amount vs Average Previous Repayment

Cluster 0 (High-Risk): Younger clients with the lowest account balances, credit amounts, and previous repayments. A higher proportion of females, high school education, and married individuals. Poor repayment behavior, including frequent payment delays.

Cluster 1 (Low-Risk): Older clients with lower account balances compared to Cluster 2 but highest credit amounts and previous repayments. A higher proportion of males, a university education, and good repayment behavior, including the highest rate of timely repayments and no payment delays.

Cluster 2 (Medium-Risk): Clients with the highest account balances and moderate financial metrics. More diverse education backgrounds, predominantly single individuals. Mixed repayment behaviors, including the highest use of revolving credit and intermediate repayment likelihood.

7.1.2 K-means using PCA

Cluster 0 (High-Risk): Clients with lower average previous repayments, lower account balances, and lower amounts of credit given. A higher proportion of females, individuals with high school education and graduates, and single individuals. This aligns with previous findings.

Cluster 1 (Low-Risk): Clients with higher average previous repayments, higher account balances, and higher amounts of credit given. A higher proportion of males, individuals with university and other types of education, and slightly more married individuals. Consistent with earlier analysis.