Machine Learning Project Report

Written by : Aung Kaung Myat

1 Dataset Description

The dataset is about the usage behavior of about 9000 active credit card holders during last 6 months. The dataset is created based on customer level with 17 behavioral variables. The dataset has 8950 observations and 18 features.

2 Goal of the project

The goal of this project is to segment the credit card holders into different groups based on their usage behavior to create a marketing stragegy. For the new marketing strategy, the credit card holders will be grouped into at least 4 groups and not more than 10 groupus in order to be effective. In order to achieve the goal, clustering algorithm will be used to segment the credit card holders into different groups.

3 Data Exploration

The dataset contains 18 features and 8950 observations. There is one categorical feature and 17 numerical features. The categorical feature is CUST_ID. As it does not have useful information about behavior of the credit card holders, it will be dropped. The numerical features are BALANCE, BALANCE_FREQUENCY, PURCHASES, ONEOFF_PURCHASES, INSTALLMENTS_PURCHASES, CASH_ADVANCE, PURCHASES_FREQUENCY, ONEOFF_PURCHASES_FREQUENCY, PURCHASES_INSTALLMENTS_FREQUENCY, CASH_ADVANCE_FREQUENCY, CASH_ADVANCE_TRX, PURCHASES_TRX, CREDIT_LIMIT, PAYMENTS, MINIMUM_PAYMENTS, PRC_FULL_PAYMENT, TENURE.

3.1 Data Visualization

As shown in the figure 1, the data contains outliers and it is squashed to the left corner. Since the data is squashed to the left corner, it is hard to see the distribution of the data.

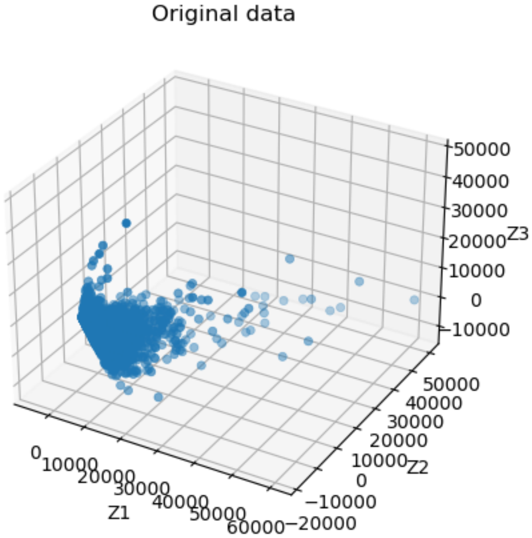

Figure 1: Original Data

3.2 Checking the missing data

It is important to check the missing data. There are missing values in two features: *MINIMUM_PAYMENTS* and *CREDIT_LIMIT* . In *MINIMUM_PAYMENTS* there are 313 missing values and in *CREDIT_LIMIT* there are 1 missing value.

3.3 Checking the outliers

The outliers are checked using the boxplot. As can be seen in the figure 2, there are outliers in most of the features.

3.4 Checking the distribution of data

The distribution of data is checked using the histogram. As can be seen in the figure 3, the data is skewed to the left or right corners.

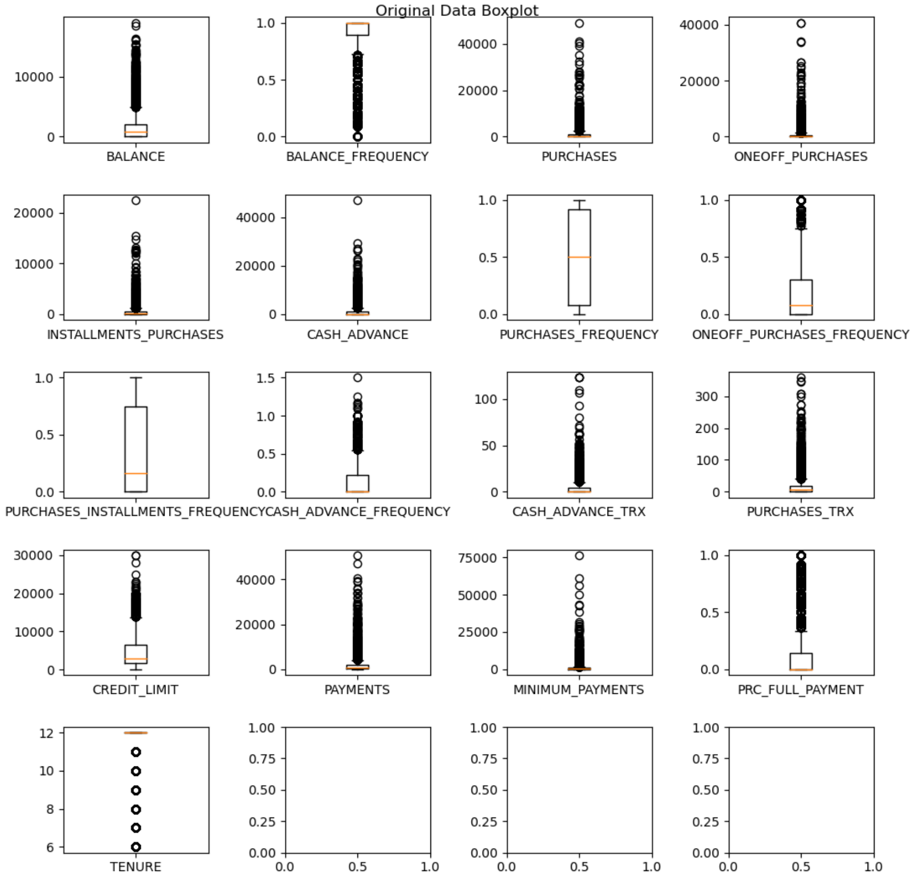

Figure 2: Original Data Boxplot

4 Data Preprocessing

4.1 Handling missing data

The missing values are handled as follows:

- Missing value in *CREDIT_LIMIT* is replaced using the mode of values because its values are not continuous values and there exisit only certain number of credit limit.
- Missing values of *MINIMUM_PAYMENT* is replaced with the mean of the values because its values are continuous.

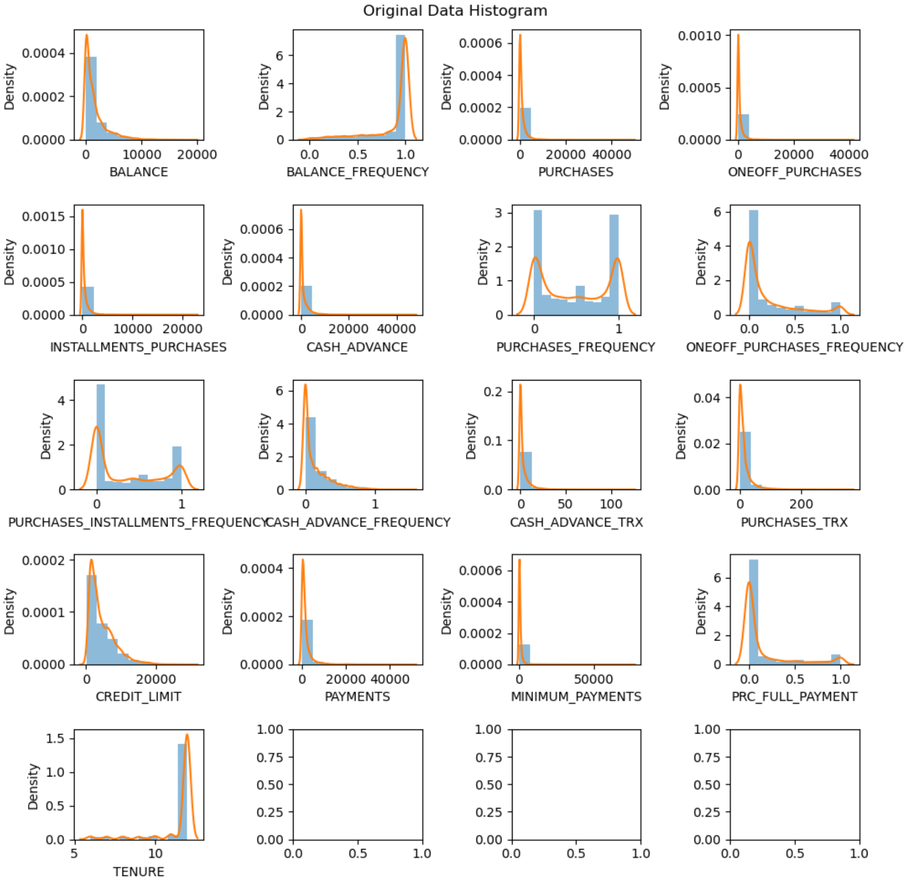

Figure 3: Original Data Histogram

4.2 Handling outliers

The outliers are first detected using the quantile range method. Since the number of outliers are too many, it is not possible to remove all the outliers. Thus, it is better to replace with calculated values. In this project, 3 methods are tried for replacing the outliers values. They are:

- mean: Replacing the outliers with the mean of the values.
- median: Replacing the outliers with the median of the values.
- k-Nearest Neighbors: Replacing the outliers with the values calculated using k-nearest neighbors.

4.3 Handling skewed data

There are many methods to handle the skewed data. Two methods are tried in this project. They are:

- Logarithmic Transformation: Transforming the data using the logarithmic function.
- Square Root Transformation: Transforming the data using the square root function.

4.4 Feature Scaling

Since, the scale of the features have effect on clustering algorithms, in this project normalization method is tried.

4.5 Dimensionality Reduction

Dimension reduction technique can filter out some noise and reduce the risk of high-dimensional datasets being very sparse. As a lot of noises are found in the dataset, dimensionality of the dataset is reduced using the PCA.

The dimensionality of the dataset is reduced to 5 components. This is done by find the number of components that explain 95% of the variance.

4.6 Data Visualization after preprocessing

All the combinations of the methods are tried. Among them the best combination is the one that have done the following:

- outliers removal: k-Nearest Neighbors
- skew removel: Logarithmic Transformation
- scaling: no normalization

It can be seen on the 2nd image from the right in the 3rd row in figure 4. This combination is chosen because the clusters of the data are more clearly visible.

5 Clustering

To group the data into clusters, 4 clustering algorithms are tried. They are:

- K-Means Clustering

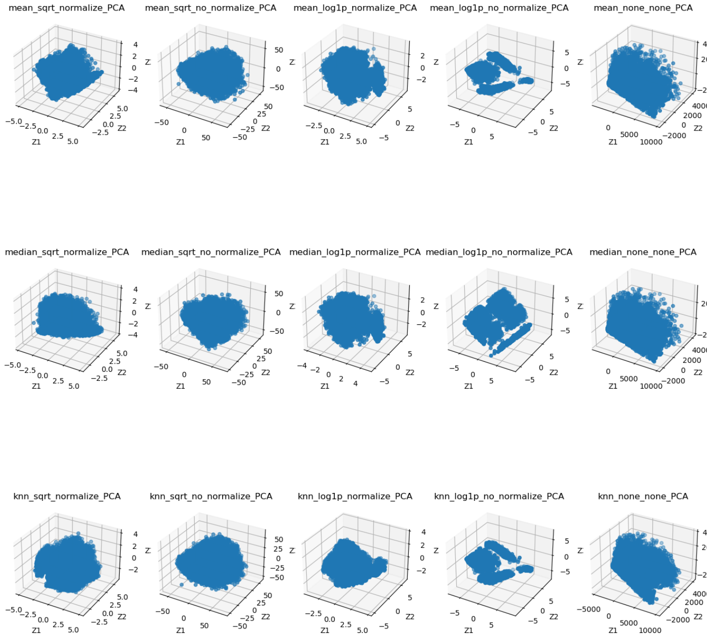

Figure 4: All Combinations of the Methods

- Hierarchical Clustering
- DBSCAN
- HDBSCAN

5.1 Hyperparameters

The hyperparameters of the algorithms are tuned using the grid search method. The best number of cluster is found using the silhouette score.

The hyperparameters of the algorithms are:

Algorithm	Hyperparameter	Value
K-Means Clustering	number of clusters	2 to 21
Hierarchical Clustering	number of clusters	2 to 21
DBSCAN	mininum points	5, 10, 20, 50, 100, 200
HDBSCAN	mininum cluster size	5, 10, 20, 50, 100, 200

Best parameters of the algorithms are:

Algorithm	Number of Clusters Created	Hyperparameter	Value
K-Means Inertia Method	6	number of clusters	6
K-Means Silhouette Method	7	number of clusters	7
Hierarchical Silhouette Method	7	number of clusters	7
DBSCAN Silhouette Method	7	mininum points	100
HDBSCAN Silhouette Method	10	mininum cluster size	20

6 Results

Among the algorithms K-Means Clustering is best suited for this dataset. It achieves the highest silhouette score of 0.5. The silhouette score of the other algorithms are 0.49 for Hierarchical Clustering, 0.48 for HDBSCAN and 0.47 for DBSCAN. DBSCAN performs the worst among the algorithms and many of the data are grouped as outliers.

In terms of data interpretability, K-Means Clustering is also the best algorithm. Especially, the clusters created by K-Means Clustering with inertia method gives results that have good interpretability.

6.1 Interpretation of the Clusters Created by K-Means Clustering with Inertia Method

The following are the clusters created by K-Means Clustering with inertia method.

Cluster 1

This cluster contains the card holders who have the medium credit limit and do frequently intallment purchases.

Feature	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
P_INSTALLMENTS_FREQ mean	0.6645	0.0021	0.6535	0.0029	0.6381	0.0030
CREDIT_LIMIT min CREDIT_LIMIT max	300 23000	50 19000	300 30000	300 20000	300 30000	150 25000

Cluster 2 This cluster contain the card holders who have low creidt limit and take cash advance to do purchases.

Feature	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
C_ADVANCE_FREQ mean	0.0032	0.2716	0.01381	0.2719	0.2767	0.0281
CREDIT_LIMIT min CREDIT_LIMIT max	300 23000	50 19000	300 30000	300 20000	300 30000	150 25000

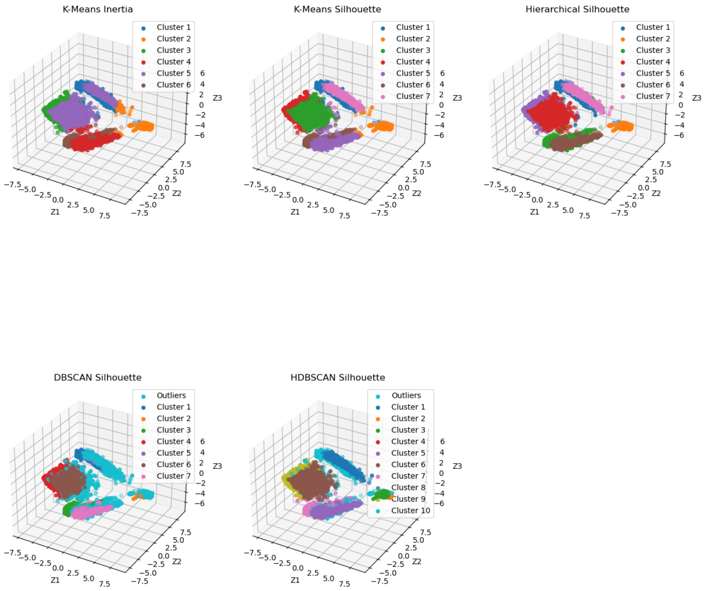

Figure 5: Clusters According to Algorithms

Cluster 3 This cluster contains the card holders who have high credit limit and high spenders but they usually do not take cash advance to do purchases.

Feature	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
PURCHASES mean	546.0464	1.8188	2540.06339	679.3299	1539.6699	854.3325
INSTAL_PURCHASES mean	528.5540	1.8085	987.3085	0.5512	717.5807	0.5593
CREDIT_LIMIT min	300	50	300	300	300	150
CREDIT_LIMIT max	23000	19000	30000	20000	30000	25000

Cluster 4 This cluster contains the card holders who have low to medium credit limit and do all type of purchases.

Feature	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
PURCHASES mean	546.0464	1.8188	2540.06339	679.3299	1539.6699	854.3325
INSTAL_PURCHASES mean	528.5540	1.8085	987.3085	0.5512	717.5807	0.5593
CASH_ADVANCE mean	27.0042	1996.4171	130.586	1818.6383	1984.0701	223.8385

Cluster 5 This cluster contains the card holders who have high credit limit and have high mininum spending.

Feature	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
MINIMUM_PAYMENTS mean	721.1698	1002.9839	666.4005	995.4895	1284.1791	557.6469
CREDIT_LIMIT min	300	50	300	300	300	150
CREDIT_LIMIT max	23000	19000	30000	20000	30000	25000

Cluster 6 This cluster contains the card holders who have low to medium credit limit and have medium spending on purchases.

Feature	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
PAYMENTS mean	769.7390	1666.6717	2423.8807	1814.1463	2633.7040	1288.4316
CREDIT_LIMIT min	300	50	300	300	300	150
CREDIT_LIMIT max	23000	19000	30000	20000	30000	25000

7 Conclusion

The goal of the project is to segment the card holders into different clusters based on their spending habits. To cluster the card holders, K-Means Clustering with inertia method is best suited for this project. The number of clusters created is 6 which is between the expected range of 4 to 10. The new marketing strategy can be implemented based on the clusters created.