Customer Segmentation of Credit Card Users Using PCA and Clustering Techniques for Behavioral Risk Profiling

Chris Lawrence De Vera
College of Computing and Information Technologies
National University
Manila, Philippines
devaraca@students.national-u.edu.ph

Emil John Llanes
College of Computing and Information Technologies
National University
Manila, Philippines
llaneses@national-u.edu.ph

***Abstract*—This study proposes an unsupervised learning framework for behavioral segmentation of credit card users using Principal Component Analysis (PCA) and clustering techniques. A dataset of 8,950 customers obtained from Kaggle was analyzed to identify latent patterns in credit utilization, repayment behavior, spending structure, and liquidity pressure. Following systematic preprocessing—including median imputation, logarithmic transformation, multicollinearity reduction, and robust scaling—PCA reduced the dataset to four principal components explaining 89.40% of total variance.**

Multiple clustering algorithms were evaluated using internal validation metrics. K-Means with six clusters achieved the best overall performance (Silhouette $\approx$ 0.377; Davies–Bouldin $\approx$ 0.964; Calinski–Harabasz $\approx$ 6289). The resulting segments revealed distinct behavioral risk profiles, ranging from high-risk revolvers to disciplined full payers. The findings demonstrate that combining dimensionality reduction with clustering provides an interpretable and scalable framework for data-driven credit risk profiling.

***Index Terms*—Credit card analytics, customer segmentation, unsupervised learning, principal component analysis (PCA), K-means clustering, behavioral risk profiling, dimensionality reduction**

I. Introduction

A credit card is a type of financial technology that can be used as both a revolving personal credit facility and a cashless payment method [1]. Credit cards allow customers to conduct electronic transactions while delaying payment for a predetermined amount of time through approved credit lines provided by financial institutions. The credit card serves as an interface to extensive backend infrastructures in charge of authorization, settlement, and recordkeeping in addition to its obvious role at the point of sale [2]. Every transaction creates durable digital records that are methodically maintained and processed by issuing banks and payment networks. These records include structured financial data such as transaction amount, timestamp, merchant, category, and location.

The historical and technological advancement of credit card systems is based on the accumulation of transactional data, which is not incidental. From early shop charge systems in the late nineteenth century to modern digital payment networks, the development of credit cards is inextricably linked to the growth of transactional data capture and customer surveillance, as detailed by Lauer [2]. Modern credit card infrastructures were specifically intended to enable datafication, allowing institutions to analyze user behavior for objectives including risk assessment, fraud detection, marketing optimization, and behavioral prediction. As a result, behavioral and technological viewpoints are becoming more and more integrated into empirical research on credit card adoption. For instance, in the study of Trinh et al. employ the Technology Acceptance Model (TAM) and the Theory of Perceived Risk (TPR) to investigate adoption behavior and illustrate that user intention is greatly influenced by perceived usefulness, convenience of use, and financial, privacy, and security concerns [1]. The study's results demonstrate that concerns related to trust, security, and data consumption are fundamental properties of the technology rather than extraneous considerations.

National settings, institutional growth, regulatory frameworks, and competitive payment technologies all influence the adoption trajectory of credit cards. In contrast to Western economies, credit cards were introduced relatively late in the Philippines. The Philippine Commercial Credit Card Company, which subsequently charged its name to Bankard Inc., was the first domestic issuer of credit cards in 1982 [3]. The Philippine credit card ecosystem had fast institutional expansion in spite of its delayed launch, with the introduction of deferred payment scheme in the mid-1980s, international network integration in the early 1990s, and early digital payment services by the late 1990s [3]. Simultaneously, improvements in fraud resistance have been brought about by technological developments in card security, most notably the switch from magnetic stripe cards to EMV-based microprocessor chips [4]. However, these developments have also created new vulnerabilities, especially in contactless NFC-enabled transactions. More recently, the Philippine payment landscape has been further altered by the quick spread of alternative payment methods (APMs), such as mobile wallets and QR-based systems, which have altered the adoption of cashless transactions while posing new issues with data

governance, cybersecurity, and consumer protection [5]. Data-driven analysis of credit card usage, especially methods that use unsupervised learning to find latent user categories based on transactional and risk-related patterns, is made possible by this changing institutional and technical context.

II. REVIEW OF RELATED LITERATURE

The global proliferation of credit cards reflects a significant shift in consumer finance, with the number of consumers holding a credit card account reaching approximately 166 million worldwide by the end of 2022, as reported by TransUnion's Q4 2022 Credit Industry Insights Report [6]. In the Philippines, the card payments market has also been on a strong growth trajectory; total card payment value is forecast to grow substantially in the coming years, with projected increases driven by financial inclusion efforts and expanding digital infrastructure [7]. According to GlobalData's 2025 forecast, the Philippine card payments market is expected to grow to PHP 4.2 trillion in 2025, with credit and charge cards accounting for 64.9% of total transaction value, significantly surpassing debit cards due to compelling rewards and value-added propositions [7], [8], [9], [10], [11]. The COVID-19 pandemic accelerated this adoption, catalyzing a preference for online transactions and installment payment plans. [12]. As online purchasing proliferated during lockdowns, consumers demonstrated a high propensity to transact using installment schemes [12]. This behavioral shift is reflected in evolving credit card eligibility frameworks across Philippine financial institutions. Access to credit cards is governed by specific criteria that vary by bank but share common requirements. Applicants typically must be Filipino citizens or foreign residents with at least two years of permanent residency, aged between 21 and 70 years depending on the issuing bank. All banks require demonstrated employment stability—typically six months to one year for regular employees or two to three years of profitable operation for self-employed applicants—alongside valid government-issued identification and substantiated proof of income [13], [14], [15].

The financial industry has historically relied on traditional machine learning techniques for customer profiling and segmentation, with unsupervised clustering methods such as K-Means and Gaussian Mixture Models commonly applied to transactional and demographic data [16], [17]. K-Means, which minimizes within-cluster variance under Euclidean distance, assumes approximately spherical cluster structures, limiting its ability to model non-convex or irregular patterns in high-dimensional data. Gaussian Mixture Models offer probabilistic cluster assignments but require iterative parameter estimation, which can increase computational complexity in large-scale datasets. Supervised approaches such as decision trees have also been employed; however, without careful validation and regularization, they are prone to overfitting and may exhibit reduced generalization performance in noisy financial data [18]. These challenges have motivated increasing interest in more flexible representation learning approaches, including deep learning frameworks [19].

Dimensionality reduction techniques such as Principal Component Analysis (PCA) have been widely applied in credit card analytics. Agarwal et al. (2020) conducted a comparative study evaluating the impact of PCA as a preprocessing step on classification performance [20]. Using a credit card dataset, they compared Logistic Regression, Decision Tree, K-Nearest Neighbor, and Naive Bayesian classifiers under two conditions: training on the original feature set and training on the dataset after PCA transformation. Their results demonstrated that applying PCA prior to model training improved performance metrics, with Logistic Regression emerging as the most efficient classifier for their specific dataset [20]. This empirical finding supports the utility of PCA as a preprocessing strategy in credit card classification tasks.

In addition, Asmath [21] applied a log1p transformation to transaction-related features to mitigate the effects of right-skewed monetary distributions. Logarithmic transformations are widely recognized for reducing positive skewness and stabilizing variance in non-negative financial data, thereby improving compatibility with statistical and machine learning models [22], [23]. By compressing extreme values, log transformations reduce the dominance of large transactions in optimization and distance-based computations, which is particularly beneficial in anomaly and fraud detection contexts where model sensitivity to outliers can distort the representation of normal behavior.

Beyond consumer-facing digital payment methods, digital payment frameworks have also been explored within institutional contexts. In a study on Philippine higher education institutions, Cendana and Palaoag [24] proposed a digital payment framework utilizing smart card technology to address persistent challenges in tuition fee collection and payment management. The framework was designed to convert traditional student identification cards into smart cards with embedded EMV chips, functioning as restricted payment cards that integrate with existing banking infrastructures through Electronic Data Capture (EDC) machines [24].

The study identified several advantages of this approach, including enhanced transaction traceability for administrative monitoring, improved financial accountability among students, and the ability to restrict fund usage to institution-specific transactions—thereby addressing behavioral issues where students diverted tuition funds to other expenditures [24]. With 91.5% of surveyed students expressing positive adoption intentions and 95 % demonstrating willingness to embrace the smart ID system, the framework demonstrated strong potential acceptability within the Philippine higher education context [24]. This institutional application of digital payment technology complements broader trends in the Philippine card payments market, where the convergence of banking infrastructure, digital literacy, and evolving payment preferences continues to reshape financial transaction ecosystems across consumer and institutional domains.

III. Methodology

The study aims to carry out customer segmentation on credit card holders with the use of unsupervised clustering algorithms to behavioral usage data. Without depending on labeled datasets, the objective of the study is to find distinct patterns in credit utilization, payment behavior, and spending.

A. Data Collection

The dataset was uploaded and updated by Arjun Bhasin in 2014 at Kaggle, a platform for data science competitions, where data scientists, students, and machine learning engineers can compete to create the best models for solving specific problems or analyzing certain datasets [25]. The used dataset consists of approximately 9,000 active credit card customers, summarizing their transaction behavior over a period of six months. The data is structured at the customer level with 18 features.

The dataset contains exactly 8,950 rows and 18 behavioral variables that capture key dimensions of credit card usage. These can be categorized into balance and credit information, purchasing behavior, cash advance behavior, and payment behavior. Table I provides the complete data dictionary.

B. Data Exploratory

During the data exploratory phase, data analysis and data cleaning is mandatory; this is done to ensure that all duplicates, missing data are eliminated before proceeding to data preprocessing to guarantee consistency and quality.

```
counts = X['CUST_ID'].value_counts()
repeating_counts = counts[counts > 1]
print(repeating_counts)

Series([], Name: count, dtype: int64)
```

Fig. 1. Check CUST_ID Uniqueness

An initial inspection confirmed that each CUST_ID was unique and no duplicate records were present, shown in Figure 1. Additionally, the dataset contains a total of 314 missing values. 1 to CREDIT_LIMIT and 313 to MINIMUM_PAYMENT respectively.

To address the missing values of the dataset, A binary indicator variable, MIN_PAY_MISSING, was created to capture potential informational value in the absence of minimum payment data. MINIMUM_PAYMENTS and CREDIT_LIMIT were imputed using their respective median values to preserve robustness against skewness and extreme values.

Feature Distributions and Skewness

In Figure 2 is a histogram analysis revealed pronounced right-skewness across monetary variables, including:

- BALANCE
- PURCHASES
- CASH_ADVANCE
- PAYMENTS
- MINIMUM_PAYMENTS
- CREDIT_LIMIT

Additionally, several transaction count variables displayed high concentration at zero, particularly those related to cash advances.

Low-Variance Feature Assessment

The variable TENURE exhibited limited variability:

- 84.7% of the observations corresponded to 12 months.
- Variance = 1.79
- Seven unique values (range: 6–12 months)

Given its high concentration at a single value and minimal discriminatory power, TENURE was excluded from further analysis.

TABLE I
Credit Card Dataset

Variable Name	Description
CUST_ID	Unique identifier for the credit card holder.
BALANCE	Balance amount remaining in the account.
BALANCE_FREQUENCY	Frequency of balance updates (0 to 1, where 1 is frequent).
PURCHASES	Total amount of purchases made.
ONEOFF_PURCHASES	Maximum purchase amount made in a single transaction.
INSTALLMENTS_PURCHASES	Amount of purchases made using installment plans.
CASH_ADVANCE	Total amount of cash advances taken.
PURCHASES_FREQUENCY	Frequency of purchases (0 to 1, where 1 is frequent).
ONEOFF_PURCHASES_FREQUENCY	Frequency of one-off purchases (0 to 1).
PURCHASES_INSTALLMENT_FREQUENCY	Balance amount remaining in the account.
CASH_ADVANCE_FREQUENCY	Frequency of cash advance transactions.
CASH_ADVANCE_TRX	Number of cash advance transactions.
PURCHASES_TRX	Total number of purchase transactions
CREDIT_LIMIT	Credit limit assigned to the cardholder.
PAYMENTS	Total amount paid by the user.
MINIMUM_PAYMENTS	Minimum amount of payments made.
PRC_FULL_PAYMENT	Percentage of months where the full balance was paid.
TENURE	Tenure of the credit card service (in months or years).

TABLE II
MISSING VALUES OF CREDIT CARD DATASET

Variable Name	Missing Values
CUST_ID	0.
BALANCE	0
BALANCE_FREQUENCY	0
PURCHASES	0
ONEOFF_PURCHASES	0
INSTALLMENTS_PURCHASES	0
CASH_ADVANCE	0
PURCHASES_FREQUENCY	0
ONEOFF_PURCHASES_FREQUENCY	0
PURCHASES_INSTALLMENT_FREQUENCY	0
CASH_ADVANCE_FREQUENCY	0.
CASH_ADVANCE_TRX	0
PURCHASES_TRX	0
CREDIT_LIMIT	1
PAYMENTS	0
MINIMUM_PAYMENTS	313
PRC_FULL_PAYMENT	0
TENURE	0

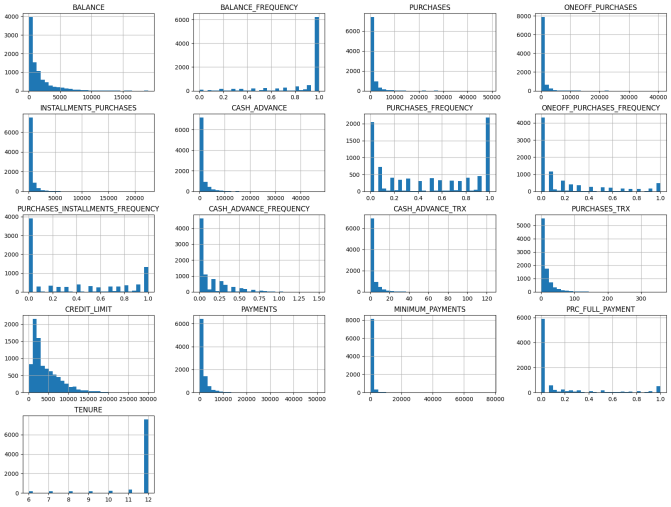

Fig. 2. Histogram of Monetary Variables

Correlation Structure

Pearson correlation analysis identified strong positive associations among several variables:

- ONEOFF_PURCHASES ↔ PURCHASES (r ≈ 0.92)
- PURCHASES_INSTALLMENT_FREQUENCY ↔ PURCHASES_FREQUENCY (r ≈ 0.86)
- CASH_ADVANCE_FREQUENCY ↔ CASH_ADVANCE_TRX (r ≈ 0.80)

These high correlations indicate redundancy among certain behavioral indicators, suggesting overlapping measurement of similar underlying constructs (e.g., spending intensity or borrowing activity).

The presence of multicollinearity has implications for distance-based modeling and dimensionality reduction.

After analyzing the state of our data, we proceeded with the following to ensure data consistency:

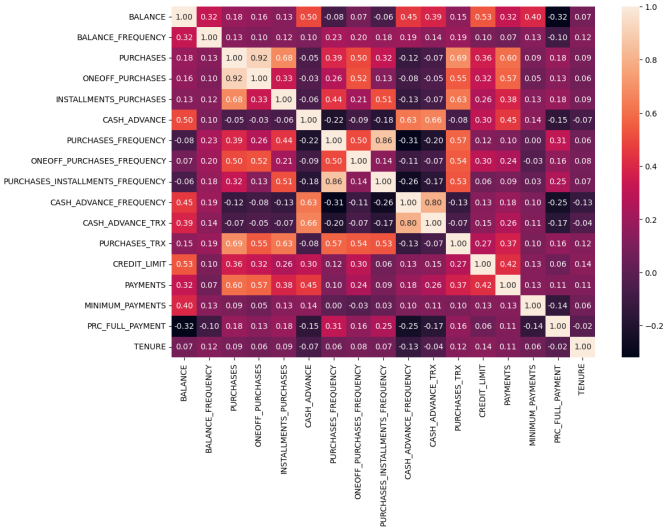

Fig. 3. Heatmap Correlation of Features

C. Data Pre-Processing

Treatment of Missing Values

As stated above, there are two variables that exhibited a missing values. 1 at CREDIT_LIMIT, and 313 missing values for MINIMUM_PAYMENTS. In a total of 314 missing values. To address these, we apply median impute for CREDIT_LIMIT. Whereas for MINIMUM_PAYMENTS, a binary indicator variable *MIN_PAY_MISSING* was introduced to preserve potential behavioral signal for missing values. Additionally, the same technique, median impute is also applied to MINIMUM_PAYMENTS

To address missing data, CREDIT_LIMIT was imputed using the median value. For MINIMUM_PAYMENTS, we first introduced a binary indicator variable (MIN_PAY_MISSING) to preserve any potential behavioral signal associated with missingness. Subsequently, missing values in MINIMUM_PAYMENTS were imputed using the median. Median imputation was selected due to the skewed distribution of monetary variables and its robustness to extreme values. Following these preprocessing steps, the dataset contained no remaining missing values.

Low-Variance Feature Removal

The variable TENURE exhibited limited variability, with a variance of approximately 1.79 and 84.7% of observations concentrated at 12 months. Given its low dispersion and minimal discriminatory power, the variable was removed from further analysis to avoid introducing noise and redundancy into the model.

Skewness Reduction via Log Transformation

Many monetary and transaction-based variables exhibited substantial right-skewness. To reduce tail dominance and stabilize variance, a logarithmic transformation (log1p) was applied to the following features:

- BALANCE
- PAYMENTS
- MINIMUM_PAYMENTS
- CREDIT_LIMIT
- PURCHASES
- PURCHASES_TRX
- ONEOFF_PURCHASES
- INSTALLMENTS_PURCHASES
- CASH_ADVANCE
- CASH_ADVANCE_TRX

The *log1p function* was chosen to handle zero values appropriately while preserving the ordinal structure of the data. This transformation compresses extreme values and improves suitability for variance-based techniques such as PCA.

Multicollinearity Mitigation

Correlation analysis revealed strong linear associations among several variables, indicating redundancy:

- ONEOFF_PURCHASES - PURCHASES (r = 0.92)
- PURCHASES_INSTALLMENT_FREQUENCY - PURCHASES_FREQUENCY (r = 0.86)
- CASH_ADVANCE_FREQUENCY - CASH_ADVANCE_TRX (r = 0.80)

To reduce multicollinearity and prevent disproportionate weighting of highly correlated dimensions, several proxy variables were removed from the analysis, including PURCHASES_TRX, CASH_ADVANCE_FREQUENCY, INSTALLMENTS_PURCHASES, and ONEOFF_PURCHASES_FREQUENCY. Eliminating these overlapping indicators improves interpretability by reducing redundancy among features and contributes to more stable and reliable principal component extraction.

Feature Scaling

Prior to PCA, all remaining features were scaled using RobustScaler, which centers variables using the median and scales them according to the interquartile range (IQR).

Robust scaling was selected to mitigate the influence of residual outliers while preserving the relative structure of the data. Scaling ensures that features contribute comparably to variance-based dimensionality reduction and distance-based clustering.

Pre-Processing Summary

The final preprocessing workflow consisted of several systematic steps to ensure data quality and analytical robustness. First, the non-informative identifier (CUST_ID) was removed. Missing values were addressed through median imputation, with the retention of a missingness indicator to preserve potential behavioral signals. The low-variance feature (TENURE) was eliminated due to its limited discriminatory power. Heavily skewed monetary variables were subjected to logarithmic transformation to reduce skewness and stabilize variance. Highly correlated proxy variables were removed to mitigate multicollinearity, and all retained features were subsequently standardized using robust scaling. The resulting dataset served as the standardized input for Principal Component Analysis and subsequent clustering procedures.

D. Principal Component Analysis (PCA) Results

Principal Component Analysis (PCA) was conducted on the fully preprocessed and robust-scaled dataset to examine latent structure and reduce dimensional complexity prior to clustering. Two configurations were evaluated: a 4-component solution and a 6-component solution.

Four-Component Solution

The first four principal components collectively explain 89.40% of the total variance in the dataset. Notably, the first two components alone account for 72.52% of the total variance, indicating that repayment-related behavior constitutes the dominant underlying structure within the data.

Six -Component Solution

Extending the model to six principal components increases the cumulative explained variance to 94.88%. However, the fifth and sixth components together contribute only an additional 5.48% of variance, indicating diminishing marginal returns beyond the fourth component and suggesting that most of the meaningful structure in the data is already captured within the first four components.

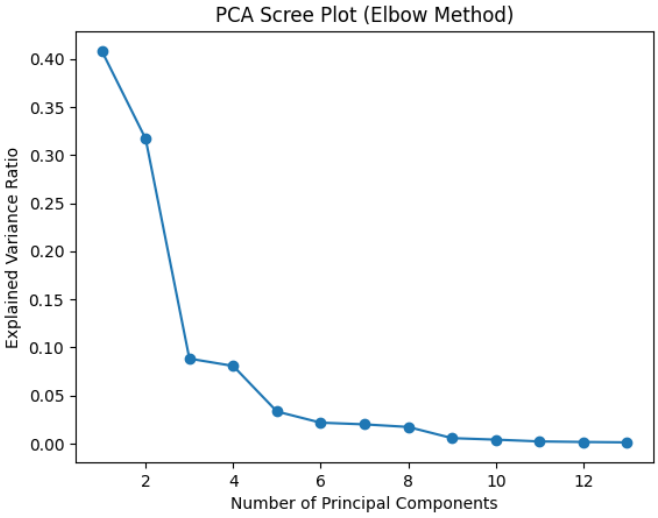

Fig. 4. PCA Explained Variance Ratio Scree Plot

Component	Explained Variance	Percentage
PC1	0.4082	40.82%
PC2	0.3170	31.70%
PC3	0.0882	8.82%
PC4	0.0806	8.06%
PC5	0.0332	3.32%
PC6	0.0216	2.16%

Interpretation of Principal Components

Loadings were examined to interpret the behavioral meaning of each component. Only dominant loadings are discussed for clarity. The first principal component (PC1), labeled Revolving Credit Utilization, explains 40.82% of the total variance and represents the dominant behavioral axis in the dataset. It exhibits strong positive loadings on BALANCE_FREQUENCY, BALANCE, and MINIMUM_PAYMENTS, alongside a strong

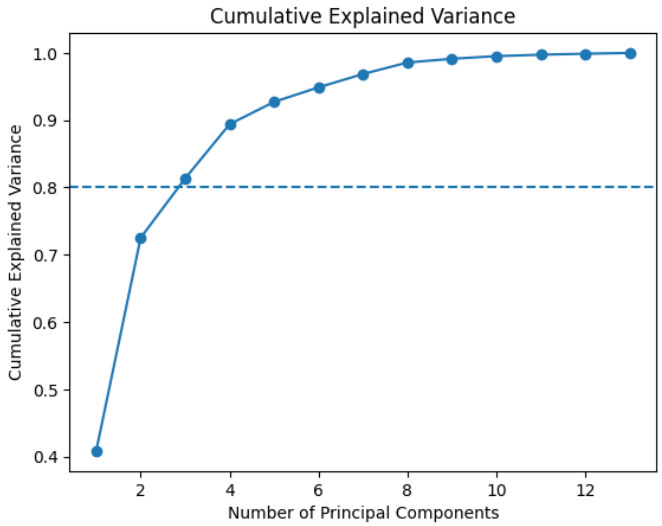

Fig. 5. PCA Cumulative Explained Variance Scree Plot

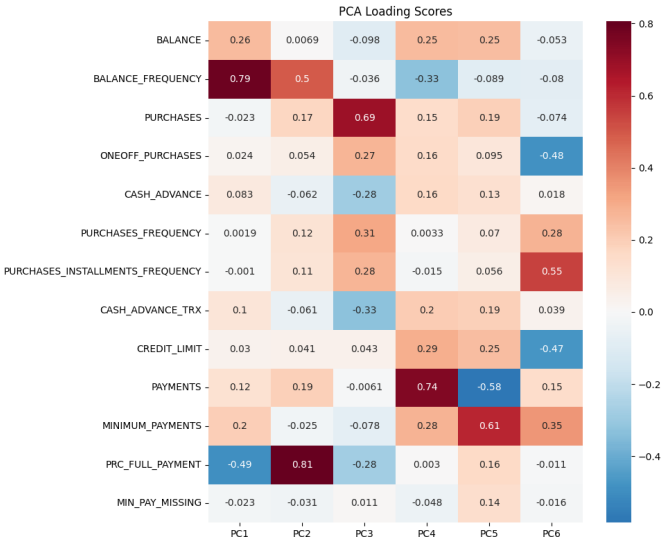

Fig. 6. PCA Loading Score Heatmap

negative loading on PRC_FULL_PAYMENT. This structure indicates that higher PC1 scores correspond to customers who frequently carry balances and rely on minimum payments, whereas lower scores reflect disciplined users who consistently settle their balances in full.

The second principal component (PC2), termed Payment Discipline and Repayment Consistency, accounts for 31.70% of the total variance. It shows strong positive loadings on PRC_FULL_PAYMENT and BALANCE_FREQUENCY. While conceptually related to PC1, this component isolates repayment reliability as a distinct behavioral dimension independent of overall spending magnitude. Customers with high PC2 scores demonstrate consistent full-payment behavior and structured balance management patterns.

The third principal component (PC3), labeled Transaction and Spending Activity, explains 8.82% of the variance. It loads strongly on PURCHASES, ONEOFF_PURCHASES, PURCHASES_FREQUENCY, and PURCHASES_INSTALLMENTS_FREQUENCY. This component differentiates customers according to transaction intensity and purchasing structure. Higher scores indicate more active credit card usage, characterized by frequent transactions and higher purchase volumes, capturing spending behavior largely independent of repayment discipline.

The fourth principal component (PC4), referred to as Financial Pressure and Liquidity Stress, explains 8.06% of the variance. Key loadings include PAYMENTS, CREDIT_LIMIT, CASH_ADVANCE, and MINIMUM_PAYMENTS. This dimension reflects short-term financial burden and liquidity dynamics, where higher scores suggest heavier payment activity, larger credit limits, and greater reliance on cash advances—potential indicators of financial pressure.

Although the analysis was extended to a six-component solution, the additional components contributed limited incremental explanatory power. The fifth component explains 3.32% of the variance and primarily refines distinctions between installment-based and one-off purchase behavior already captured in PC3. The sixth component explains 2.16% of the variance and captures minor residual variation largely driven by minimum payment-related features. Collectively, these additional components provide marginal structural nuance but limited additional interpretive value beyond the first four principal components.

Justification for Retaining Four Components

The four-component solution was selected for downstream modeling based on its strong balance between explanatory power and interpretability. This solution explains 89.40% of the total variance, with the first two components alone capturing over 72% of the dataset's structural variation. The third and fourth components introduce meaningful additional behavioral dimensions, specifically spending structure and financial pressure, thereby enriching the representation of customer credit behavior. In contrast, the inclusion of further components yields only minimal gains in explained variance and offers limited additional interpretive clarity. Overall, the four-component model achieves an effective balance between variance coverage, behavioral insight, and parsimony, providing a compact yet behaviorally rich representation of customer credit activity suitable for clustering and segmentation tasks.

E. Model Selection and Clustering Results

Evaluation Framework

Clustering was performed on the PCA-reduced feature space consisting of four principal components, which together explain approximately 89.40% of total variance. Using PCA ensures dimensional compactness while retaining dominant behavioral structure.

The following clustering algorithms were evaluated:

- DBSCAN (density-based)
- Gaussian Mixture Models (GMM)

- Hierarchical Agglomerative Clustering (Ward, Complete, Average linkage)
- K-Means

Model performance was assessed using three standard internal validation metrics:

- Silhouette Score (higher is better) — measures cohesion and separation.
- Davies–Bouldin Index (DBI) (lower is better) — evaluates cluster compactness and separation.
- (higher is better) — ratio of between-cluster dispersion to within-cluster dispersion.

DBSCAN (Density-Based Clustering)

DBSCAN was evaluated through systematic parameter tuning. The min_samples parameter was set to twice the number of features (min_samples = 2 × n_features = 8), while the eps value was initially selected using the k-distance elbow method and further refined through manual trials around approximately 0.5 and 0.3. Despite these efforts, the algorithm demonstrated high sensitivity to the choice of eps. Small changes in this parameter significantly altered the clustering structure, either collapsing most observations into a single dominant cluster or labeling a large proportion of data points as noise. The best observed performance metrics were weak, with a Silhouette score of approximately 0.0123, a Davies–Bouldin index of about 1.644, and a Calinski–Harabasz score of 153.62. These values indicate poor cluster separation and limited structural stability. Based on these results, DBSCAN was rejected due to its inadequate metric performance and instability across parameter configurations.

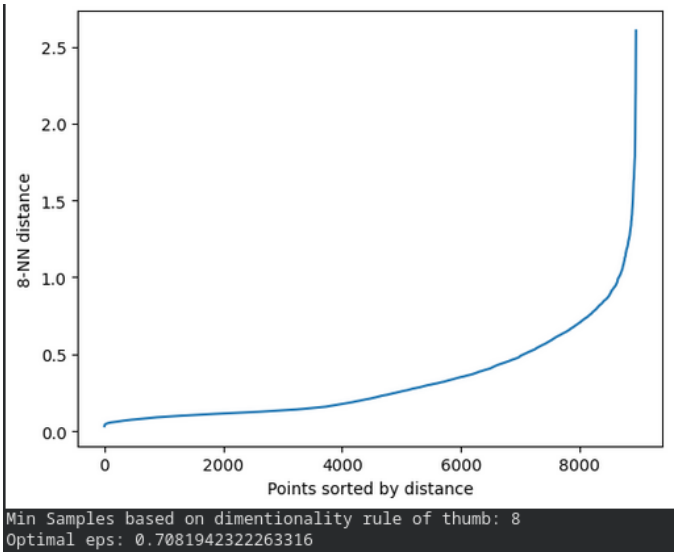

Fig. 7. 8-NN Distance Elbow Graph

Gaussian Mixture Models (GMM)

Gaussian Mixture Models (GMMs) were assessed over a range of k = 1–29 components using the Bayesian Information Criterion (BIC), with the minimum BIC observed near k ≈ 5. The final configuration was set to n_components = 5, covariance_type = full, and random_state = 42. At k = 5, the

Fig. 8. DBSCAN Labeled 3D Scatter Plot 1

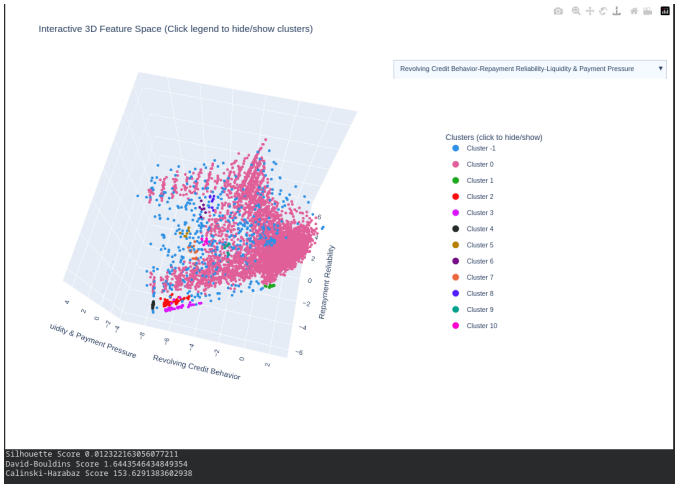

Fig. 9. DBSCAN Labeled 3D Scatter Plot 2

model achieved a Silhouette score of approximately 0.1770, a Davies–Bouldin index of 1.568, and a Calinski–Harabasz score of 2224.48. Although GMMs provide probabilistic (soft) cluster assignments, the resulting cluster separation was moderate and notably weaker than that achieved by hierarchical and K-Means methods. Consequently, GMM was retained as a secondary option but was not selected as the final clustering model.

Hierarchical Agglomerative Clustering

Hierarchical clustering was evaluated using three linkage strategies, with cluster counts determined through dendrogram inspection and consideration of structural interpretability. Ward linkage with k = 6 produced well-balanced clusters, achieving a Silhouette score of approximately 0.3490, a Davies–Bouldin index of 1.090, and a Calinski–Harabasz score of 5555.81, reflecting strong compactness and separation. Complete linkage with k = 8 underperformed relative to Ward and Average linkage, with a Silhouette score of 0.2097, Davies–Bouldin index of 1.543, and Calinski–Harabasz score

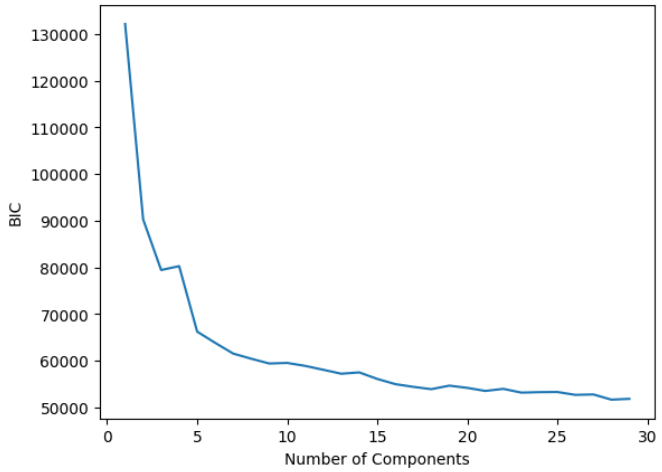

Fig. 10. BIC Elbow Graph

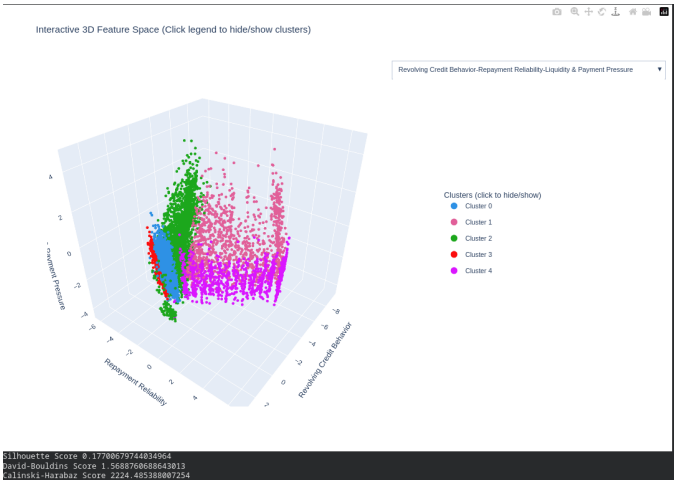

Fig. 11. GMM Labeled 3D Scatter Plot 1

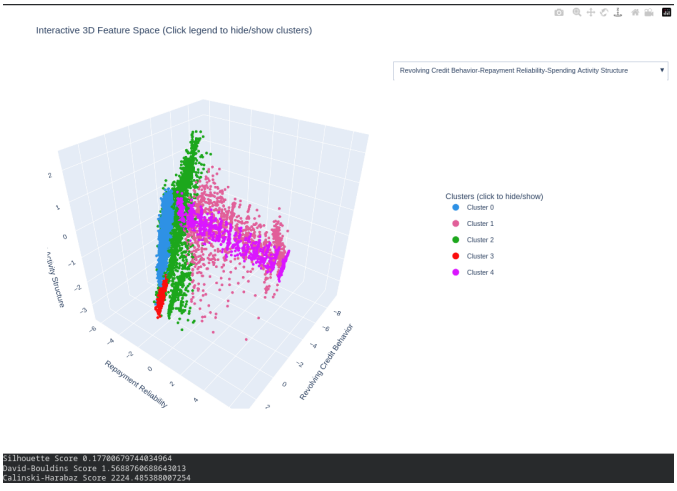

Fig. 12. GMM Labeled 3D Scatter Plot 2

of 3520.80, indicating weaker cluster separation and compactness. Average linkage with k = 7 achieved the highest Silhouette score of 0.4173 and the lowest Davies–Bouldin index of 0.922, demonstrating superior separation quality, though its Calinski–Harabasz score of 3120.91 was lower than that of Ward and K-Means. Based on these results, Ward and Average linkage were retained as strong alternatives due to their robust metric performance.

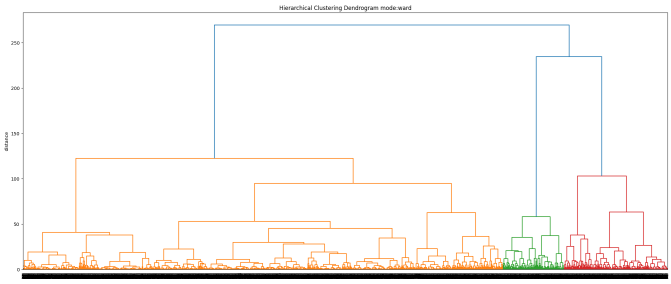

Fig. 13. Agglomerative(ward) Dendogram

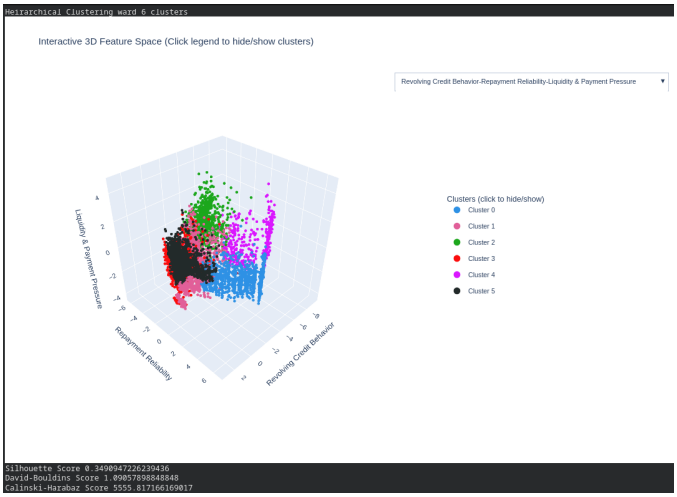

Fig. 14. Agglomerative(ward) Labeled 3D Scatter Plot 1

K-Means Clustering

K-Means clustering was evaluated over k = 1–29 using inertia (within-cluster sum of squares), with the elbow method suggesting an optimal choice at k = 6. At this configuration, the model achieved a Silhouette score of approximately 0.3769, a Davies–Bouldin index of 0.9642, and a Calinski–Harabasz score of 6289.11. K-Means produced the highest Calinski–Harabasz score among all models evaluated, reflecting strong between-cluster separation relative to within-cluster compactness, while also maintaining a low Davies–Bouldin index and a competitive Silhouette score. Based on this overall performance, K-Means with k = 6 was selected as the final clustering model, offering the best balance between cluster separation, compactness, and interpretability.

Final Model Justification

K-Means with k = 6 was selected as the final clustering solution due to its combination of strong global separation, compact cluster structure, and clear interpretability in PCA

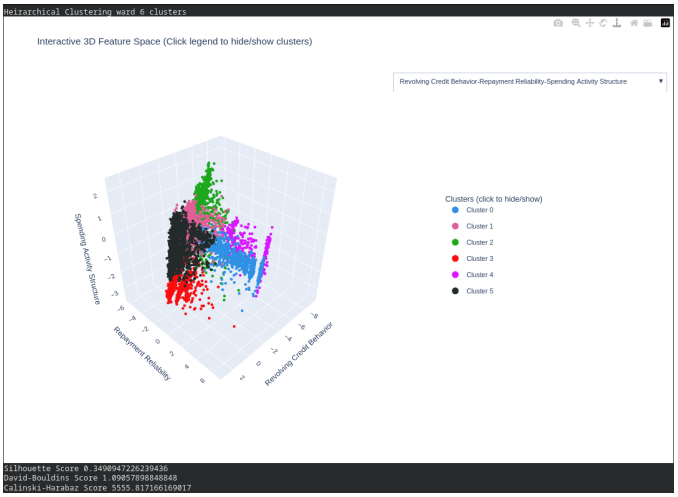

Fig. 15. Agglomerative(ward) Labeled 3D Scatter Plot 2

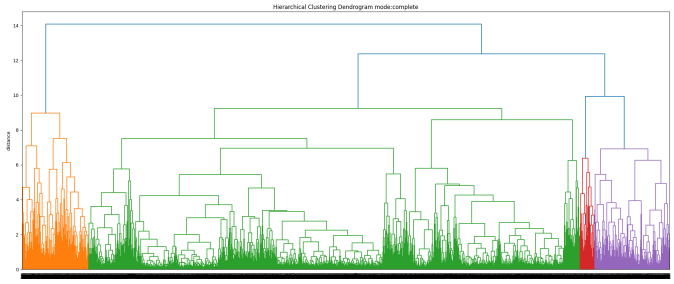

Fig. 16. Agglomerative(complete) Dendogram

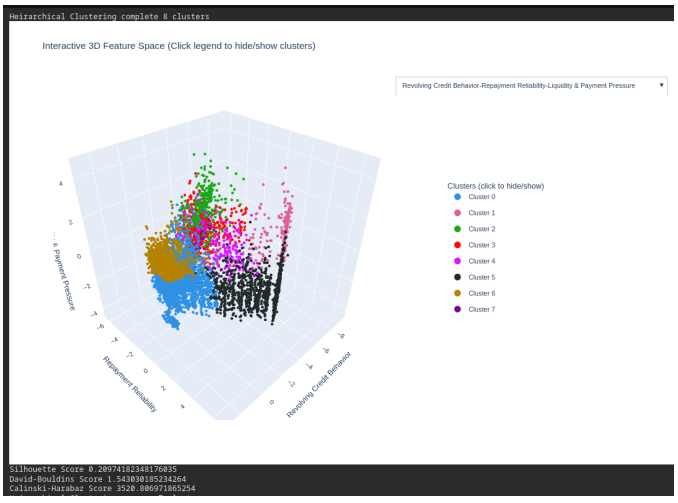

Fig. 17. Agglomerative(complete) Labeled 3D Scatter Plot 1

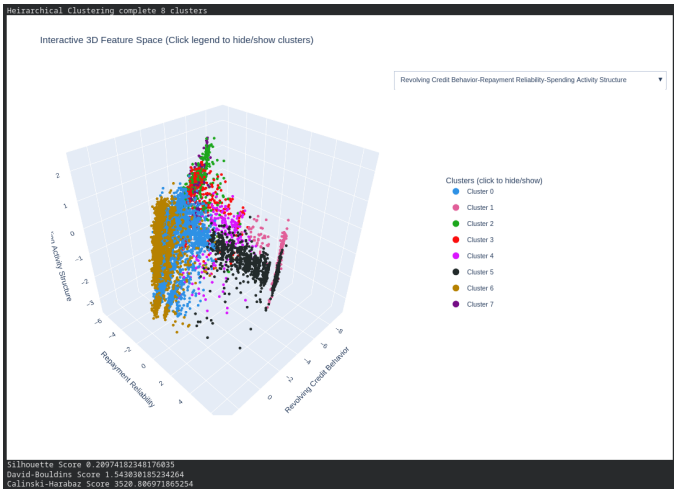

Fig. 18. Agglomerative(complete) Labeled 3D Scatter Plot 2

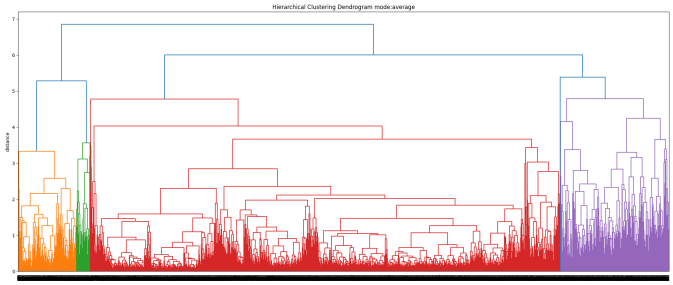

Fig. 19. Agglomerative(average) Dendogram

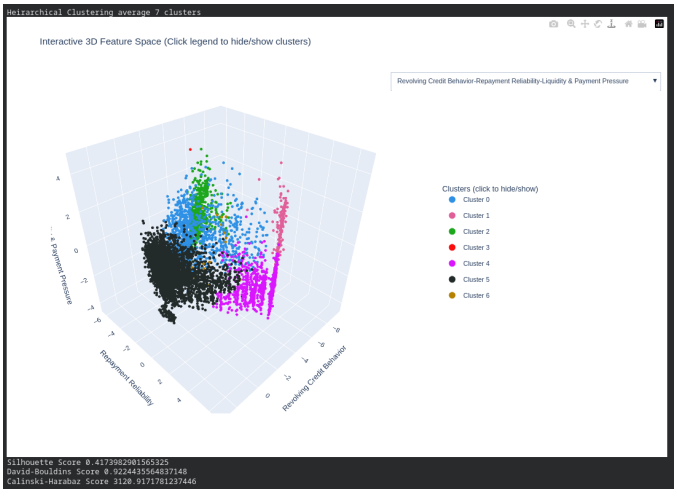

Fig. 20. Agglomerative(average) Labeled 3D Scatter Plot 1

space. It achieved the highest Calinski–Harabasz index, indicating robust between-cluster dispersion, alongside a low Davies–Bouldin index, reflecting compact and well-defined clusters. While Hierarchical Average linkage attained the highest Silhouette score, K-Means provided more balanced clusters witdh superior global dispersion, making it better suited for segmentation and practical deployment. In contrast, DBSCAN was deemed unsuitable because of its sensitivity to density parameters, and GMM consistently underperformed across evaluation metrics.

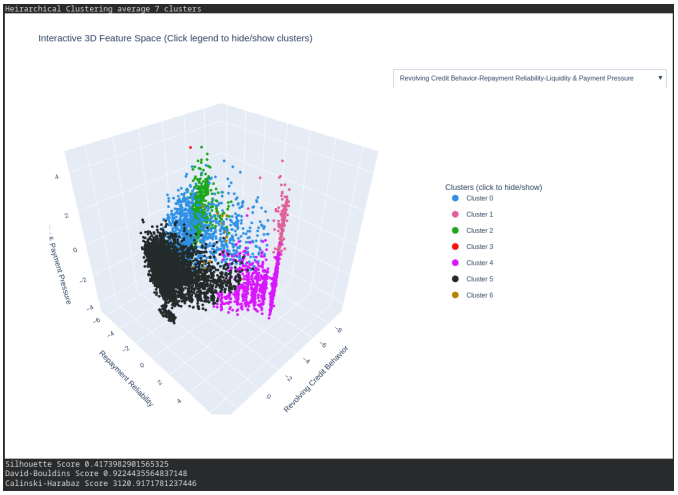

Fig. 21. Agglomerative(average) Labeled 3D Scatter Plot 2

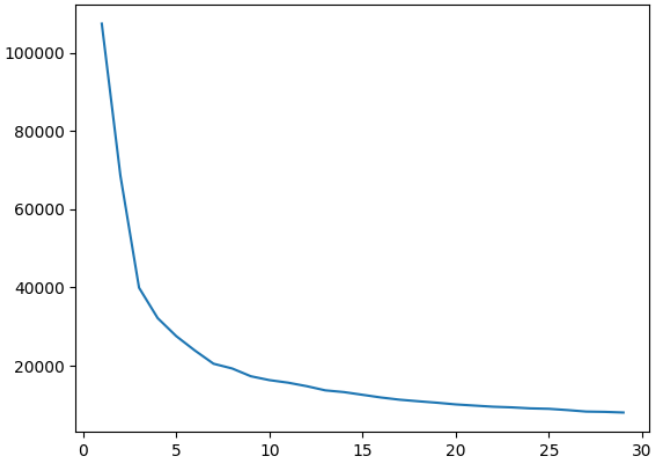

Fig. 22. WCSS Elbow Graph

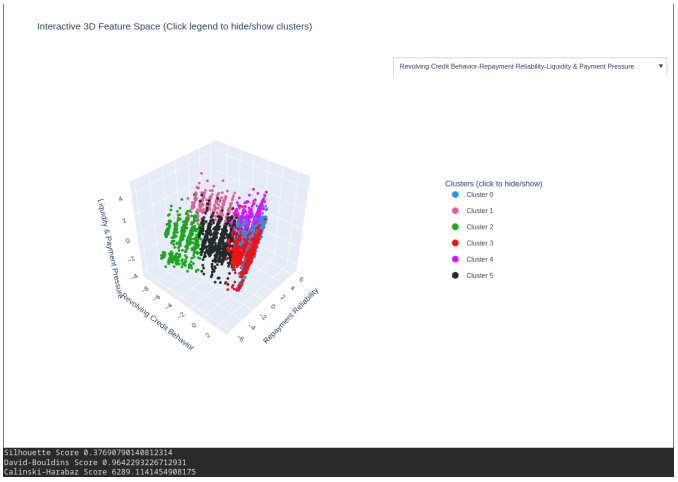

Fig. 23. KMEANS Labeled 3D Scatter Plot 1

IV. Results and Discussion

Cluster-Level Interpretations (K-Means, k=6)

The six-cluster solution reveals distinct credit behavior profiles:

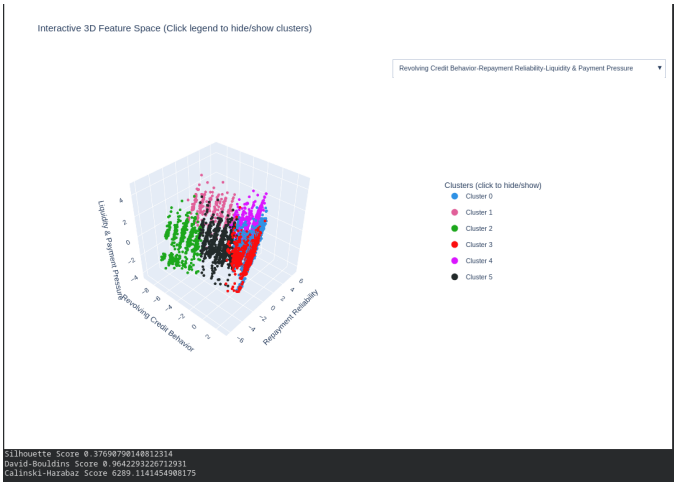

Fig. 24. KMEANS Labeled 3D Scatter Plot 2

Cluster 0 – Balanced Revolvers

Moderate balance carrying with acceptable repayment behavior and moderate spending levels. Liquidity stress appears situational rather than chronic.

- Revolving Credit Behavior: Moderate (-1.3 → 2.8)
- Repayment Reliability: Generally reliable (-1.46 → 2.82)
- Spending Activity Structure: Moderately structured (-0.77 → 2.18)
- Liquidity & Payment Pressure: Mixed liquidity (-4.21 → 3.35)

Customers with balanced credit use and decent repayment behavior. Moderate spenders, sometimes under liquidity stress.

Cluster 1 – Cautious Credit Users

- Revolving Credit Behavior: Very low (-9.21 → -2.36)
- Repayment Reliability: Moderate (-1.32 → 4.28)
- Spending Activity Structure: Low structured spending (-3.39 → 0.91)
- Liquidity & Payment Pressure: Moderate pressure (-1.86 → 3.93)

Customers who rarely use revolving credit but repay reliably when they do. Low spending activity and careful with finances.

Cluster 2 – High Credit Risk

- Revolving Credit Behavior: Very low (-9.21 → -2.36)
- Repayment Reliability: Moderate (-1.32 → 4.28)
- Spending Activity Structure: Low structured spending (-3.39 → 0.91)
- Liquidity & Payment Pressure: Moderate pressure (-1.86 → 3.93)

Customers who rarely use revolving credit but repay reliably when they do. Low spending activity and careful with finances.

Cluster 3 – Moderate Credit Users

- Revolving Credit Behavior: Moderate (-1.19 → 2.73)
- Repayment Reliability: Moderate (-2.39 → 2.23)
- Spending Activity Structure: Low to Moderate (-3.04 → -0.03)
- Liquidity & Payment Pressure: Mixed (-4.18 → 3.39)

Customers who use credit moderately, with average repayment reliability. Spending is conservative; liquidity varies.

Cluster 4 – High Repayers

- Revolving Credit Behavior: Low (-4.56 → 0.12)
- Repayment Reliability: High (1.20 → 6.41)
- Spending Activity Structure: Low (-3.41 → 1.14)
- Liquidity & Payment Pressure: Moderate (-2.59 → 3.40)

Customers who may not heavily use credit but repay consistently and reliably. Spending is structured and controlled.

Cluster 5 – Opportunistic Borrowers

- Revolving Credit Behavior: Low (-4.91 → -0.01)
- Repayment Reliability: Moderate (-3.88 → 1.43)
- Spending Activity Structure: Moderate (-2.46 → 1.92)
- Liquidity & Payment Pressure: Moderate (-3.77 → 4.58)

Customers who borrow selectively, repay moderately, and have flexible spending patterns. Liquidity can fluctuate

V. Conclusion

This study evaluated multiple clustering approaches on a PCA-reduced representation of credit card customer behavior. Among the tested models, K-Means with six clusters applied to the 4-principal-component space delivered the strongest overall balance of separation, compactness, and interpretability.

Final performance metrics for K-Means (k = 6):

- Silhouette Score ≈ 0.377
- Davies–Bouldin Index ≈ 0.964
- Calinski–Harabasz Index ≈ 6289

Compared to DBSCAN, Gaussian Mixture Models, and hierarchical variants, K-Means demonstrated superior global dispersion (highest Calinski–Harabasz), low cluster overlap (low Davies–Bouldin), and competitive cohesion (Silhouette). The resulting segmentation structure was stable, interpretable, and operationally actionable.

The six-cluster structure uncovered in this study demonstrates that credit card customers are not merely separated by spending volume, but by multidimensional behavioral patterns spanning revolving behavior, repayment discipline, spending structure, and liquidity pressure. By reducing complex financial activity into four principal behavioral axes and identifying six distinct and interpretable customer segments, this research provides a structured and actionable framework for portfolio differentiation.

Importantly, the segmentation isolates both high-risk profiles (Cluster 2) and high-value, low-risk profiles (Cluster 4), while also revealing nuanced middle segments such as opportunistic borrowers and balanced revolvers. This granularity moves beyond simplistic risk scoring and enables targeted credit policy, product design, monitoring strategies, and revenue optimization initiatives.

The findings demonstrate that combining rigorous preprocessing, dimensionality reduction, and systematic model benchmarking can transform raw transactional data into operational intelligence. The resulting segmentation is not only statistically robust but strategically meaningful, offering financial institutions a scalable and data-driven foundation for risk management, customer engagement, and long-term portfolio stability.

In essence, this study shows that well-structured unsupervised learning can convert complex credit behavior into clear, interpretable, and business-aligned insights.

References

[1] H. N. Trinh, H. H. Tran, and D. H. Q. Vuong, "Determinants of consumers' intention to use credit card: a perspective of multifaceted perceived risk," *J. Retailing Consum. Serv.*, vol. 56, p. 102177, 2020.

[2] J. Lauer, "Plastic surveillance: Payment cards and the history of transactional data, 1888 to present," *Big Data & Soc.*, vol. 7, no. 1, Jan.-Jun. 2020.

[3] eCompareMo, "A brief history of the credit card and its usage in the Philippines," Website, Apr. 2018, accessed: 2026. [Online]. Available: https://www.ecomparemo.com/info/a-brief-history-of-the-credit-card-and-its-usage-in-the-philippines

[4] M.-H. Yang, Y.-S. Hsu, and H.-C. Hsu, "Enhanced EMV security: Preventing credit card fraud from a distance," *IEEE Access*, vol. 13, pp. 78 345–78 363, 2025.

[5] M. F. G. Ng, "Implications of alternative payment methods in the Philippine financial technology ecosystem: A research note," 2025, unpublished manuscript.

[6] TransUnion, "Amidst stubbornly high inflation, consumers continue to turn to credit cards, home equity to maintain stability: Q4 2022 TransUnion credit industry insights report explores latest credit trends," TransUnion Newsroom, Feb. 2023, [Accessed 12 Feb. 2026]. [Online]. Available: https://newsroom.transunion.com/q4-2022-ciir/

[7] GlobalData, "Philippines card payments set to grow 18.8% in 2025, forecasts GlobalData," GlobalData Media Center, Nov. 2025, [Accessed 12 Feb. 2026]. [Online]. Available: https://www.globaldata.com/media/banking/philippines-card-payments-set-to-grow-18-8-in-2025-forecasts-globaldata/

[8] ——, "Philippines cards and payments: Opportunities and risks to 2029," GII Research, Product Code: GDFS0930CI, Dec. 2025, [Accessed 12 Feb. 2026]. [Online]. Available: https://www.giiresearch.com/report/gd1901436-philippines-cards-payments-opportunities-risks.html

[9] Malaya Business Insight, "PH card payments to hit P4.2T in 2025 - research firm," Malaya.com.ph, Nov. 2025, [Accessed 12 Feb. 2026]. [Online]. Available: https://malaya.com.ph/business/business-news/ph-card-payments-to-hit-p4-2t-in-2025-research-firm/

[10] Manila Standard, "Philippines card payments seen hitting P4.2 trillion in 2025," ManilaStandard.net, Nov. 2025, [Accessed 12 Feb. 2026]. [Online]. Available: https://manilastandard.net/business/314673553/philippines-card-payments-seen-hitting-p4-2-trillion-in-2025.html

[11] A. M. C. Sy, "PHL card payments seen growing to P4.2 trillion," BusinessWorld Online, Nov. 2025, [Accessed 12 Feb. 2026]. [Online]. Available: https://www.bworldonline.com/page/30/?m=freezone

[12] Y. B. Kurata *et al.*, "Analysis of Filipino consumer perception on shopping using credit cards in the Philippines: A multi regression approach," in *Proc. 5th African Int. Conf. Ind. Eng. Oper. Manage.*, Johannesburg/Pretoria, South Africa, Apr. 2024, pp. 1–12.

[13] Wise, "Best PAL mabuhay credit card Philippines - which should you choose?" Wise.com/ph, Oct. 2025, [Accessed 12 Feb. 2026]. [Online]. Available: https://wise.com/ph/blog/mabuhay-miles-credit-card

[14] Metrobank, "Apply for a credit card," Metrobank.com.ph, 2025, [Accessed 12 Feb. 2026]. [Online]. Available: https://www.metrobank.com.ph/cards/credit-cards

TABLE III
CLUSTER PROFILES: STANDARDIZED FEATURE MEANS

Cluster	Revolving Credit Behavior		Repayment Reliability		Spending Activity Structure		Liquidity & Payment Pressure	
	min	max	min	max	min	max	min	max
0	-1.318145	2.787138	-1.464013	2.819588	-0.771241	2.179589	-4.208039	3.353653
1	-9.211489	-2.358364	-1.320539	4.281275	-3.38784	0.909144	-1.864756	3.928846
2	-7.601124	-3.007953	-6.030268	-1.149589	-1.755792	2.284722	-2.81594	3.980898
3	-1.191002	2.727644	-2.394996	2.234825	-3.039669	-0.027701	-4.177333	3.393885
4	-4.562509	0.121774	1.203143	6.405423	-3.409421	1.140459	-2.585946	3.39667
5	-4.907974	-0.011259	-3.87586	1.426781	-2.456858	1.922426	-3.768863	4.578556

[15] Chinabank, "Credit cards — application requirements," Chinabank.ph, 2025, [Accessed 12 Feb. 2026]. [Online]. Available: https://www.chinabank.ph/credit-cards-eligibility-requirements

[16] R. S. Sá and J. B. M. S. Jardim, "Machine learning for credit card user segmentation in the financial sector," M.S. thesis, Dept. Inf. Manage., Universidade Nova de Lisboa, Lisbon, Portugal, 2025. [Online]. Available: http://hdl.handle.net/10362/179216

[17] M. M. Rahman, S. S. Akhi, S. Hossain, M. I. Ayub, M. T. Siddique, A. Nath, P. C. Nath, and M. M. Hassan, "Evaluating machine learning models for optimal customer segmentation in banking: A comparative study," *Amer. J. Eng. Technol.*, vol. 6, no. 12, pp. 68–83, Dec. 2024.

[18] Fraud Detection Handbook, "Introduction to model validation and selection," Reproducible Machine Learning for Credit Card Fraud Detection, 2024, [Accessed 12 Feb. 2026]. [Online]. Available: https://fraud-detection-handbook.github.io/fraud-detection-handbook/Chapter_5_ModelValidationAndSelection/Introduction.html

[19] T. Yachamaneni, U. Kotadiya, and A. S. Arora, "Credit card customer profiling using self-supervised representation learning on multi-source financial data," *IJAIDSML*, vol. 6, no. 1, pp. 164–173, 2025.

[20] A. Agarwal, A. Rana, N. Verma, and K. Gupta, "A comparative study and enhancement of classification techniques using principal component analysis for credit card dataset," in *Proc. 2020 Int. Conf. Intell. Eng. Manage. (ICIEM)*, 2020, pp. 443–448.

[21] A. S. Asmath, "Enhancing credit card fraud detection accuracy by optimisation of anomaly detection algorithms and resampling techniques," M.Sc. thesis, School of Computing, National College of Ireland, Dublin, Ireland, 2024. [Online]. Available: https://norma.ncirl.ie/8674/1/aafreenshanasmath.pdf

[22] J. W. Osborne, "Notes on the use of data transformations," *Practical Assessment, Research, and Evaluation*, vol. 8, no. 1, p. 6, 2002.

[23] G. E. P. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society: Series B*, vol. 26, no. 2, pp. 211–252, 1964.

[24] D. I. Cendana and T. D. Palaoag, "The potential of designing a digital payment framework for philippine heis," *IOP Conference Series: Materials Science and Engineering*, vol. 803, p. 012045, 2020.

[25] Coursera Staff, "What is Kaggle and what is it used for?" Coursera, Jan. 2025, [Web page]. [Online]. Available: https://www.coursera.org/in/articles/kaggle