Customer Segmentation Using Machine Learning: A Credit Card Usage Clustering Approach

Rahul Kumar[1], Raj Vardhan[2], Suraj Nikhil[3], Suvam Chakraborty[4], Elaiyaraja P[5]
[1,2,3,4,] *UG Student, Department of Computer Science and Engineering, Sir M Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India*
[5]*Associate Professor, Department of Computer Science and Engineering, Sir M Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India*

***Abstract*—This study applies unsupervised machine learning techniques to segment credit card users based on 18 behavioral variables. Using clustering algorithms like K-Means, Agglomerative Clustering, and Gaussian Mixture Models, we identify distinct customer groups. Principal Component Analysis (PCA) enhances clustering performance and visualization. The resulting segments—such as Big Spenders, Average Users, and High Riskers—offer actionable insights for marketing and risk management. Our approach demonstrates the potential of data-driven segmentation in financial analytics.**

***Index Terms*—Behavioral Segmentation, Unsupervised Profiling, Cluster Interpretability, Dimensionality Reduction Analytics, Customer Microtargeting, PCA-Enhanced Clustering, Financial Data Stratification**

I. INTRODUCTION

Credit cards are central to modern financial behaviour, providing convenience while generating rich transactional data. However, traditional segmentation techniques based solely on demographics fail to capture the nuanced financial habits of users. To overcome this, the "Credit Card Clustering" project utilizes machine learning-based clustering to categorize customers based on behavioural variables such as purchase frequency, payment history, and credit utilization. This approach enables financial institutions to better understand customer needs, mitigate risks, and implement more tailored marketing campaigns. By turning complex transactional data into actionable insights, the project demonstrates the value of data science in modern banking.

II. LITERATURE SURVEY

Recent research demonstrates the growing application of machine learning in financial data analysis. Sharma et al. (2023) used K-Means and DBSCAN for credit card fraud detection, showing clustering's advantage over rule-based systems. Patel et al. (2022) explored K-Means and Hierarchical Clustering for customer segmentation, proving useful in targeted marketing. Iyer et al. (2024) compared clustering methods for transaction analysis and found Agglomerative Clustering effective for risk assessment. These studies validate the use of clustering for uncovering patterns in complex datasets, which inspired the methodology used in our project to provide segmentation-based financial strategies.

III. METHODOLOGY

The methodology employed in this study revolves around the analysis of a dataset containing behavioral data from over 9000 active credit card holders. The dataset includes 18 variables that describe aspects such as spending patterns, payment behavior, and credit usage. Initial efforts were focused on data cleaning, which involved treating missing values using median imputation and normalizing numerical features to ensure uniformity across scales. This preprocessing stage was crucial for preparing the dataset for clustering algorithms.

Following data preparation, exploratory data analysis was conducted to understand variable distributions, detect correlations, and uncover patterns within the dataset. These insights informed the selection and tuning of clustering models. A variety of unsupervised

learning algorithms were then applied, including K-Means, Agglomerative Hierarchical Clustering, and Gaussian Mixture Models. These models enabled the identification of natural groupings within the customer base based on similarities in financial behavior.

To improve the quality of clustering and reduce dimensional complexity, Principal Component Analysis (PCA) was utilized. PCA transformed the original dataset into a lower-dimensional space while retaining the most informative features, thereby enhancing both model performance and visual interpretability. The number of clusters was determined using inertia and silhouette score metrics to ensure optimal segmentation.

Once clustering was complete, each resulting group was analyzed in detail to interpret distinct behavioral traits. These insights were used to propose personalized marketing strategies and risk management practices tailored to the specific characteristics of each customer segment.

I Data Collection

The dataset was obtained from a public source, Kaggle, comprising credit card usage data of approximately 9,000 customers over the past six months. The data includes 18 behavioral variables such as balance, purchases, cash advances, and credit limits. Each entry represents an individual credit card holder. This dataset provides a strong basis for clustering and segmentation, allowing for deep behavioral analysis of customers, which is essential for designing personalized marketing strategies and improving customer relationship management.

II Data Preprocessing and Cleaning

The dataset contained missing values, which were imputed using the median of respective columns to avoid data skew. Non-numeric and identifier columns like CUST_ID were removed as they do not contribute to clustering. To ensure feature parity, all numerical variables were normalized using StandardScaler(). This standardization process scaled all features to have a mean of zero and a standard deviation of one, making the clustering algorithms more effective and improving convergence during iterative processes.

III Feature Extraction using PCA

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset while retaining maximum variance. Various PCA configurations (2, 3, and 4 components) were tested to identify the optimal structure for clustering. PCA significantly improved the clustering metrics by transforming the data into a lower-dimensional space that highlights the most informative variance patterns. This step also reduced noise and redundancy, allowing better separation and interpretation of clusters during visualization and subsequent analysis.

IV Clustering Algorithms Implementation

Multiple clustering techniques were implemented to ensure robustness in segmentation. These included KMeans, Agglomerative Hierarchical Clustering, and Gaussian Mixture Models (GMM). Each method was applied on both the original and PCA-transformed datasets. KMeans served as the baseline, using inertia and silhouette scores for cluster validation. Agglomerative Clustering provided hierarchical insights, while GMM allowed probabilistic interpretation of cluster memberships, offering a flexible view of customer profiles. The goal was to identify optimal groups with distinct behavioral traits.

V Model Evaluation

After clustering, each model was evaluated using both visual and quantitative metrics. The elbow method was used to determine the optimal number of clusters based on inertia. Silhouette analysis provided clarity on how well-separated the clusters were. Visual tools such as silhouette plots, pairplots, and cluster-wise breakdowns helped to validate the segmentation qualitatively. These evaluations ensured that each cluster formed a meaningful group with distinguishable characteristics, critical for actionable marketing strategies.

VI Customer Segmentation Analysis

Customers were grouped into six distinct clusters based on usage behavior. Each cluster was profiled to extract key financial behaviors and usage patterns, such as purchase frequency, balance levels, credit limit usage, and payment behaviors. The clusters were labeled with intuitive names: The Average Joe, The Active Users, The Big Spenders, The Money Borrowers, The High Riskers, and The Wildcards. Each group revealed unique traits, helping businesses target segments with personalized marketing efforts and better risk control.

VII Marketing Strategy Formulation

After identifying cluster profiles, tailored marketing strategies were developed. For example, "Average Joes" can be encouraged to increase card usage through incentives, while "Big Spenders" may benefit

from loyalty programs. High-risk groups were addressed by limiting credit exposure or offering financial literacy resources. The aim was to not only maximize customer value but also align promotions and services with specific needs and behaviors, increasing engagement, reducing churn, and improving customer satisfaction.

IV RESULTS ANALYSIS AND DISCUSSION

The clustering analysis revealed meaningful groupings of credit card users based on their financial behaviors. Through PCA-enhanced clustering and multiple algorithms, six distinct customer segments were identified.

Each segment provides actionable insights that can directly inform data-driven marketing and credit policies.

- Notable Improvement with PCA

Applying PCA dramatically enhanced the clustering results. On reducing features to 2 components, silhouette scores improved from ~0.25 to ~0.46, and visual separation became more distinct. This validated that PCA helps in eliminating noise and redundancy, which often masks true clusters in high-dimensional spaces.

- Optimal Cluster Count: Six

Based on inertia plots and silhouette scores, six clusters consistently offered the best trade-off. The elbow method showed diminishing returns beyond six clusters, and silhouette scores confirmed strong inter-cluster separation. This number also allowed meaningful customer segmentation without overfitting or losing generality.

ACKNOWLEDGMENT

We thank Mr. Elaiyaraja P, Assistant Professor, Department of CSE at Sir MVIT, for his guidance and insights throughout this project. We also extend our gratitude to the Department of Computer Science and Engineering, Sir M Visvesvaraya Institute of Technology, for providing the necessary resources and infrastructure. Lastly, we acknowledge the open dataset and Python libraries that made the analysis possible.

REFERENCES

[1] Iyer, N., Mehta, S., Nair, R. (2024). "Clustering Techniques for Credit Card Transaction Analysis", International Journal of Business Analytics.

[2] Sharma, A., Desai, P., Gupta, R. (2023). "Credit Card Fraud Detection Using Machine Learning Clustering Techniques", Journal of Financial Technology

[3] Patel, S., Roy, M., Jain, V. (2022). "Customer Segmentation in Credit Card Data Using ML", International Journal of Data Science.

[4] Andrew Ng (2019). "Machine Learning" Course on Coursera.

[5] McKinney, W. (2017). "Python for Data Analysis".

[6] Bishop, C. M. (2006). "Pattern Recognition and Machine Learning".