WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Customer Personality Analysis using PCA and KMeans Clustering

Tin Tin Htar*

University of Computer Studies, Yangon, Myanmar

A B S T R A C T :

Understanding customer personality and behavior has become a cornerstone of intelligent marketing, retention strategies, and product design. With rising volumes of consumer data, traditional analytics tools fall short in offering insights that are both precise and actionable. This study uses the publicly available "Customer Personality Analysis" dataset from Kaggle, consisting of demographic, lifestyle, and transactional attributes. The methodology begins with data cleaning and transformation, followed by dimensionality reduction through Principal Component Analysis (PCA) and clustering using KMeans to identify consumer groups in order to predict customer responses to marketing campaigns.

Keywords: customer behavior, PCA, KMeans

1. Introduction

Customer Personality Analysis (CPA) is a strategic discipline in modern data science that focuses on identifying patterns in consumer behavior, motivations, and preferences. With businesses increasingly shifting toward data-driven operations, understanding customer personality has become essential for developing personalized marketing strategies, enhancing customer experience, and improving product offerings. The exponential growth in customer data generated through online transactions, social media interactions, and digital marketing campaigns presents both opportunities and challenges. Traditional analytics approaches are often limited in their ability to process high-dimensional, nonlinear data. In contrast, machine learning (ML) techniques offer scalable solutions for extracting insights from complex datasets.

This study leverages a combination of clustering, classification, to create an end-to-end pipeline for profiling customers based on their behavior and predicting their response to marketing efforts. The key objectives are to predict customer campaign responses and to segment users using PCA and KMeans.

2. Related Work

The technique of understanding and predicting customer behavior has long been a central topic in marketing analytics and data science. Several studies have proposed machine learning and statistical methods for segmenting consumers, modeling campaign responses, and providing actionable insights. Recent work emphasizes not only predictive accuracy but also model transparency through explainable artificial intelligence.

Kotler, P. and K. L. Keller highlighted the importance of customer segmentation in modern marketing, advocating for behavioral data-driven segmentation over static demographic approaches. Traditional segmentation methods like RFM (Recency, Frequency, Monetary) analysis, while still used, have been complemented by machine learning clustering techniques.

Aggarwal, C.C. discussed the utility of unsupervised learning models—particularly KMeans and hierarchical clustering—for customer grouping, which provide more nuanced insights when combined with dimensionality reduction techniques like PCA.

Dohan, M., Ates, M., and Hossain, M. demonstrated the methods how models such as Logistic Regression, Random Forest, and XGBoost can effectively predict campaign response, especially when complemented by proper data balancing techniques. Their study, also based on the Kaggle Customer Personality dataset, emphasized data preprocessing, feature engineering, and model evaluation through F1 Score and AUROC.

Saranya, G. and Sivaranjani, S. explored the linear SVM for customer purchase prediction and showed that supervised learning techniques outperform traditional heuristics in online shopping environments, particularly for imbalanced datasets.

3. Background Theory

Customer behavior modeling is rooted in statistical learning, pattern recognition, and model interpretability. This study integrates core machine learning methods, unsupervised learning for segmentation.

PCA is a dimensionality reduction technique that transforms high-dimensional data into a smaller set of uncorrelated components, while preserving most of the variance (information) in the data. It simplifies the dataset while keeping key patterns, removes redundancy among correlated features,

speeds up clustering and modeling processes. enables visualization of high-dimensional data in 2D or 3D. PCA takes a dataset with many features and finds a new set of features (principal components) that are combinations of the original ones. These new features are arranged in order of how much of the original data's variability they explain. The first few components usually capture the most important information, allowing for simplification without losing too much detail.

KMeans is an unsupervised learning algorithm that groups customers into k clusters based on feature similarity. It assigns each data point to the cluster with the nearest mean (centroid). It groups similar customers based on behavior or demographics, enables segmented targeting in marketing, works well after applying PCA (reduces noise). While various types of clustering algorithms exist, including exclusive, overlapping, hierarchical and probabilistic, the k-means clustering algorithm is an example of an exclusive or "hard" clustering method. This form of grouping stipulates that a data point can exist in just one cluster. This type of cluster analysis is commonly used in data science for market segmentation, document clustering, image segmentation and image compression. The k-means algorithm is a widely used method in cluster analysis because it is efficient, effective and simple.K-means is an iterative, centroid-based clustering algorithm that partitions a dataset into similar groups based on the distance between their centroids. The centroid, or cluster center, is either the mean or median of all the points within the cluster depending on the characteristics of the data.

4. Proposed System

The proposed system is a modular, scalable machine learning pipeline designed to analyze customer personality traits and predict their response to marketing campaigns. The process begins with the acquisition of a real-world dataset, which includes demographic, lifestyle, and transactional attributes of customers. This data is collected from the Kaggle "Customer Personality Analysis" dataset. This step involves cleaning missing values, removing irrelevant fields (like ID and timestamp fields), and encoding categorical variables using one-hot encoding. Then the prepossessed dataset is splitted with Training data (80%) and Testing data (20%). Numerical features are normalized using MinMaxScaler to ensure uniform feature scaling for distance-based algorithms like KMeans. Derived variables (e.g., total spend, average frequency) are created to enrich the dataset. Correlation and importance analysis is conducted to identify influential features for modeling. Principal Component Analysis (PCA) is used to reduce dimensionality while preserving variance. KMeans is then applied to group customers into segments. These clusters are visualized to identify behavioral patterns and are useful for marketing personalisation.

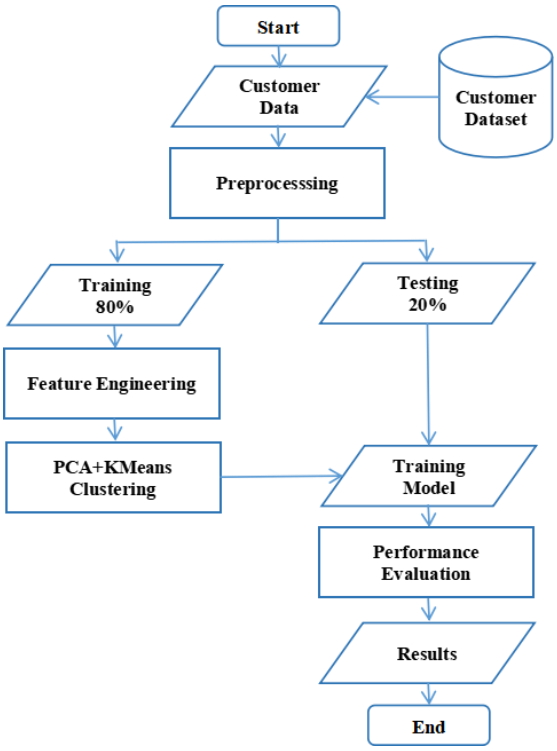

Fig. 1 - System Flow Diagram.

4.1. Preprocessing

- Missing Value Imputation: Missing values in features such as Income are filled using the median to avoid skewing the distribution.
- Categorical Encoding: Features like Education and Marital_Status are converted into numerical form using One-Hot Encoding to make them compatible with data mining algorithms. Machine learning models require numerical inputs to perform calculations. However, many real-world datasets—like the Customer Personality Analysis dataset—contain categorical variables, which are text-based or symbolic.Two such features in your dataset are: Education (e.g., Graduate, PhD, Basic) and Marital_Status (e.g., Single, Married, Together, Widow). One-Hot Encoding is a method to convert each category in a column into a new binary (0 or 1) column. Instead of assigning arbitrary numbers (which could imply an order), One-Hot Encoding creates a separate column for each unique category, marking 1 if that category is present, and 0 otherwise.

Table 1 - Before One-Hot Encoding

Customer_ID	Education	Martial status
001	Graduate	Married
002	PhD	Single
003	Basic	Together

Table 2 -After One-Hot Encoding

Customer_ID	Education_ Graduate	Education_ PhD	Education _Basic	Martial status-Married	Martial status-Single	Martial status-Together
001	1	0	0	1	0	0
002	0	1	0	0	1	0
003	0	0	1	0	0	1

- Feature Selection: Irrelevant columns like ID, Dt_Customer, and other constants are dropped.
- Feature Scaling: MinMaxScaler is applied to normalize numerical attributes (e.g., MntWines, Income, Recency) within a 0–1 range. In real-world datasets like the Customer Personality Analysis, numerical features often have different scales or units. For example:
  - Income: ranges from 20,000 to 120,000
  - MntWines: ranges from 0 to 1500
  - Recency: ranges from 0 to 100 (days since last purchase). If these features are used directly, models like KMeans (which depend on distance) or gradient-based models (like Logistic Regression) may be biased toward features with larger numeric ranges. This ensures all features contribute equally during modeling. MinMaxScaler rescales each numeric feature to a [0, 1] range using this formula:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Table 3 - Before Scaling

Customer_ID	Income	MntWines	Recency
001	60,000	500	10
002	120,000	1500	100
003	30,000	0	60

Table 4 -After MinMax scaling

Customer_ID	Income-scaled	MntWines-scaled	Recency-scaled
001	0.375	0.333	0.0
002	1.0	1.0	1.0
003	0.0	0.0	0.555

4.2. Feature Engineering

- Derived Features: Aggregated variables such as total monetary value spent, average frequency of purchases, and engagement scores are created. Derived features are new variables created by combining or transforming existing features. They provide additional insights and help models learn patterns more effectively. In the Customer Personality Analysis dataset, we can engineer useful features like:

1. Total Monetary Value Spent : Customers spend across different product categories are MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds.

Total_Spending = MntWines + MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts + MntGoldProds (2)

Table 5 - Total_Spending

Customer_ID	MntWines	MntMeatProducts	MntGoldProds	Total_Spending
001	300	200	150	650

2. Average Purchase Frequency: How often a customer engages through different channels are NumWebPurchases, NumCatalogPurchases, NumStorePurchases

Avg_Purchase_Frequency = (NumWebPurchases + NumCatalogPurchases + NumStorePurchases) / 3 (3)

Table 6 - Avg_Purchase_Frequency

Customer_ID	WebVisits	Catalog	Store	Avg_Purchase_Frequency
002	5	2	4	3.67

3. Engagement Score : Combines several behavioral features into a single engagement metric are NumWebVisitsMonth, AcceptedCmp1 to AcceptedCmp5, Response.

Engagement_Score = NumWebVisitsMonth + Total_Campaign_Accepted + Response (4)
(where Total_Campaign_Accepted = sum of AcceptedCmp1 to AcceptedCmp5)

Table 7 - Engagement_Score

Customer_ID	WebVisits	Campaigns_Accepted	Response	Engagement_Score
003	6	2	1	9

- Correlation Analysis: Pearson correlation and SHAP importance values are used to rank and retain only relevant features for model building. Correlation analysis helps determine how strongly two numerical variables are related. Pearson correlation coefficient (ρ) ranges from -1 to +1:
  - +1: strong positive relationship
  - -1: strong negative relationship
  - 0: no linear relationship
  -

In customer data, it's useful for identifying features that are: Redundant (highly correlated with each other) and Relevant (highly correlated with the target variable, e.g., Response). Correlation gives a global linear relationship.

Table 8 -After MinMax scaling

Feature	Recency	MntWines	Income	NumWebVistsMonth	Response
Recency	1.00	-0.32	-0.15	0.12	-0.42
MntWines	-0.32	1.00	0.47	0.06	0.38
Income	-0.15	0.47	1.00	0.02	0.21
NumWebVists Month	0.12	0.06	0.02	1.00	-0.29
Response	-0.42	0.38	0.21	-0.29	1.00

  - Recency has a negative correlation with Response → longer inactivity means less chance to respond.
  - MntWines is positively correlated → wine buyers are more responsive.
  - Income is moderately correlated but may not be a strong predictor alone.

4.3. Clustering using PCA+KMeans

- PCA Application: Principal Component Analysis is used to reduce dimensionality while retaining >85% variance.
- KMeans Clustering: The optimal number of clusters (k=3) is chosen based on the silhouette score and elbow method. Step by step process of KMeans are:
  - Choose the number of clusters k (e.g., using the elbow method or silhouette score)
  - Randomly initialize k centroids
  - Assign each data point to the nearest centroid
  - Recalculate centroid positions
  - Repeat steps 3–4 until assignments stabilize (convergence)
- Cluster Profiles are:
  - Cluster 0: High-income, high-spending customers with frequent web interactions.
  - Cluster 1: Low-to-mid income customers with sporadic engagement.
  - Cluster 2: Older cust
- After applying PCA and then KMeans with k=3:

Table 9 -Customer Description

Customer	Description
0	High-income, high-spending online buyers
1	Low-spending, low-engagement consumers
2	older customers with moderate spend

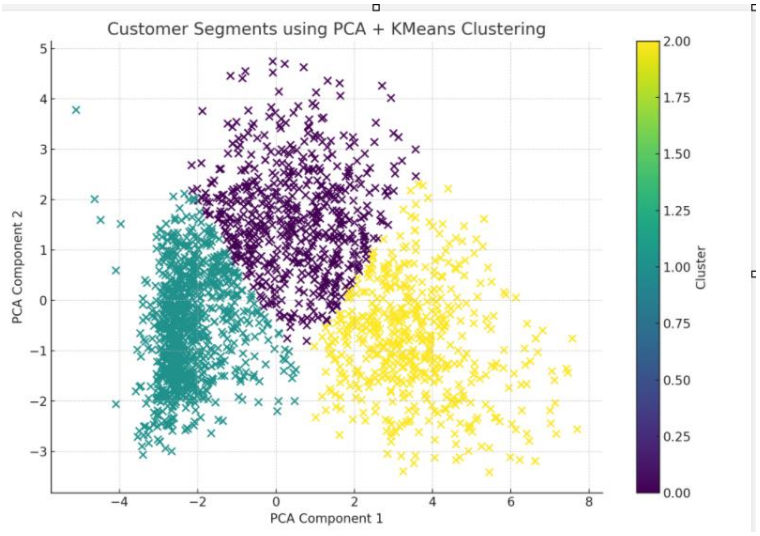

Fig. 2 - Customer Segments using PCA + KMeans Clustering.

4.4. Model Evaluation

- Metrics: Accuracy, Precision, Recall, F1 Score and ROC-AUC are used.
- ROC AUC, or Area Under the Receiver Operating Characteristic curve, is a performance metric used to evaluate binary classification models. It quantifies how well a model can distinguish between two classes (e.g., positive/negative, diseased/healthy). A higher ROC AUC indicates better model performance.

$$\text{Accuracy} = \frac{\text{True Poistive+True Negative}}{\text{True Poistive+Trure Negative+False Positive+False Negative}} \quad (5)$$

$$\text{Precision} = \frac{\text{Ture Positive}}{\text{True Positive+False Positive}} \quad (6)$$

$$\text{Recall} = \frac{\text{Ture Positive}}{\text{True Positive+False Negative}} \quad (7)$$

$$\text{F} - \text{measure} = 2 * \frac{\text{Precision*Recall}}{\text{Precision+Recall}} \quad (8)$$

Table 10 -Performance Metrics

Metrics	Value
F1 Score	0.8635
ROC-AUC	0.860
Top Features	Recency, NumWebVistsMonth, MntMeatProducts

5. Conclusion

This study successfully demonstrated a comprehensive and interpretable machine learning framework for Customer Personality Analysis, focusing on marketing campaign response prediction and behavioral segmentation. Using a real-world dataset from Kaggle, we implemented an end-to-end system that integrates:Data preprocessing and feature engineering, PCA + KMeans clustering for customer segmentation. This system can be extended to integrate behavioral (e.g., clickstream) data, use deep learning (e.g., LSTM, transformers), make real-time segmentation with Spark and perform cross-domain analysis across industries.

REFERENCES

1. Kotler, P. and K. L. Keller, Marketing Management, 15th ed. Upper Saddle River, NJ, USA: Pearson Education, 2016.
2. Aggarwal, C.C., Data Mining: The Textbook. Springer, 2015, doi:10.1007/978-3-319-14142-8.
3. Dohan, M., Ates, M., and Hossain, M.,"Customer Personality Analysis Using Machine Learning and Explainable AI," in Proc. IEEE Int. Symp. Digital Forensics and Security (ISDFS), 2025.

4. Saranya, G. and Sivaranjani, S.,"Prediction of Customer Purchase Intention Using Linear SVM," in J. Phys.: Conf. Ser., vol. 1712, 2020, Art. no. 012024.

5. Imakash3011, "Customer Personality Analysis Dataset," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis