Customer Personality Analysis
Machine Learning

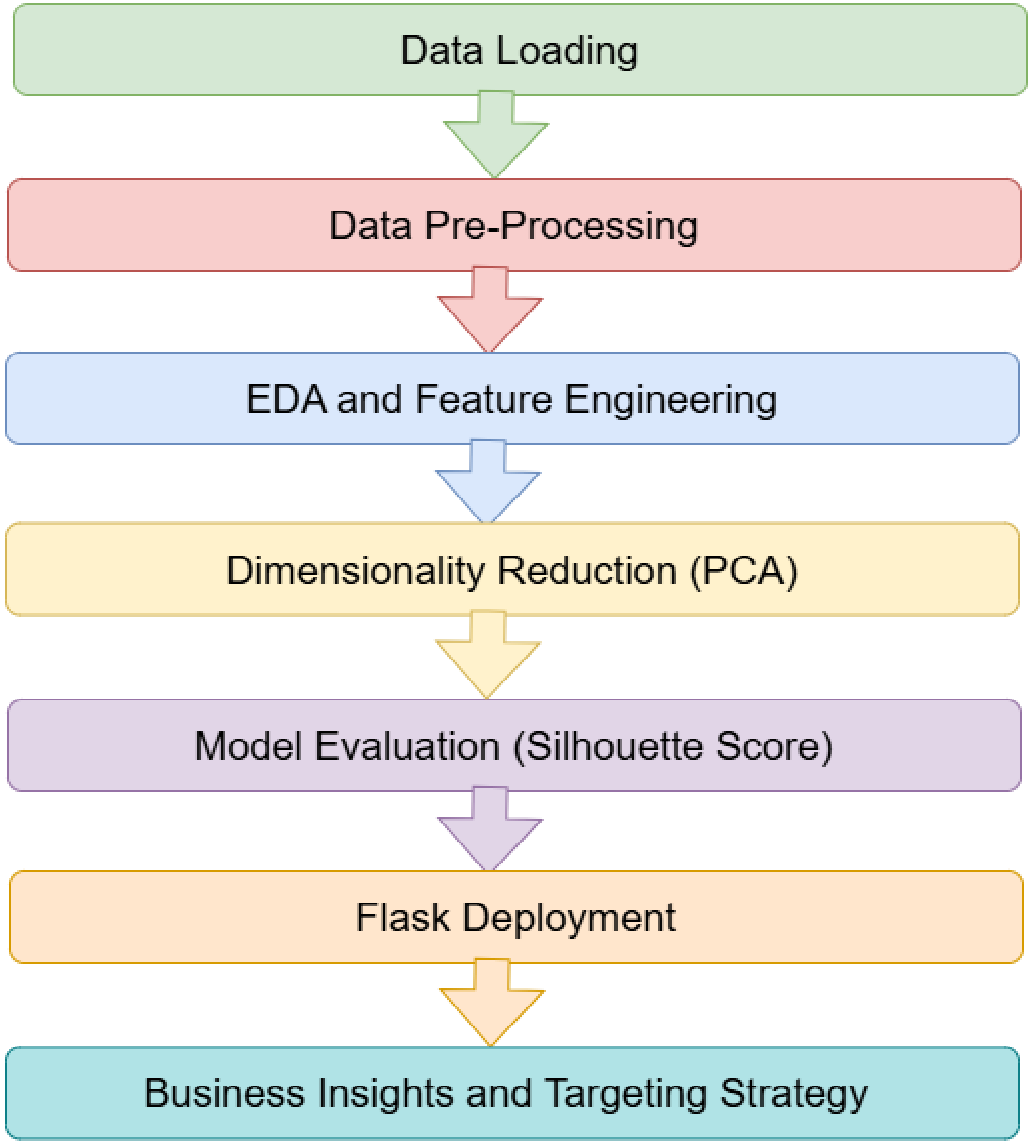
Machine learning project

Document Type : Architecture

Author : Nisha Doshi

Date : 06-02-2025

Architecture



Data Loading

Data Pre-Processing

EDA and Feature Engineering

Dimensionality Reduction (PCA)

Model Evaluation (Silhouette Score)

Flask Deployment

Business Insights and Targeting Strategy

Data Collection & Preprocessing

The first step involves loading the raw dataset from Kaggle, which is typically in CSV or Excel format. After loading, various preprocessing steps are performed. Missing values in the Income feature are handled by replacing them with the median of that column. Unnecessary columns such as 'ID', 'Z_CostContact', and 'Z_Revenue' are dropped to streamline the dataset. Label Encoding is applied to categorical columns like Marital_Status and Education to convert them into numerical values. Furthermore, several new features are created, such as the Age column (derived from Year_Birth), Kids (a combination of Teenhome and Kidhome), Total_Amount_Spent (aggregation of all product spending columns), and Number_of_Purchases (count of product purchase columns). The Accepted_Campaign column is created by combining all campaign acceptance columns. Any rows with outliers in the Income feature are removed using the Interquartile Range (IQR) method.

Data Exploration & Visualization

In the next phase, the dataset undergoes Exploratory Data Analysis (EDA). This step uses libraries such as Plotly, Seaborn, and Matplotlib for creating visualizations like histograms, scatter plots, and heatmaps to better understand the data distribution, correlations, and patterns. Insights gained from EDA help identify which features are important for clustering and which ones can be dropped due to redundancy or lack of significant variance. For example, the Recency and Days_Customer_Spent columns are excluded due to lack of peaks or drops, indicating that they do not contribute meaningfully to customer segmentation.

Feature Selection & Transformation

Following EDA, a subset of features is selected for clustering. These include Age, Education, Marital_Status, Kids, Income, Total_Amount_Spent, Accepted_Campaign, Number_of_Purchases, and Complain. The features that do not contribute significantly to the model's goal are excluded. Additionally, dimensionality reduction is performed using Principal Component Analysis (PCA) to reduce the number of features while retaining most of the variance in the data, making the clustering model more efficient.

Clustering

With the transformed data, clustering algorithms such as K-Means and Hierarchical Clustering are applied to group customers based on their spending behaviors and characteristics. The optimal number of clusters is chosen using the Elbow Method, where the cost (inertia) is plotted against the number of clusters to determine where the curve begins to flatten. K-Means is used to assign each customer to one of the identified clusters, while Hierarchical Clustering provides additional insights into how clusters are related to each other. After clustering, the model’s performance is evaluated using the Silhouette Score, which helps assess how well-separated the clusters are.

Business Insights

Once the clustering model is trained, it provides valuable business insights. For example, Cluster 1 might represent high-income, high-spending customers, while Cluster 2 could represent budget-conscious, low-income customers. Based on these insights, targeted marketing strategies are developed. For Cluster 1, premium products and loyalty programs could be suggested, while for Cluster 2, discounts and budget-friendly offerings might be appropriate. These insights guide how the business can improve its customer engagement and increase sales.

Deployment

The final model, along with the targeting strategies, is deployed as a web application using Flask. This allows the model to serve predictions in real-time. New customer data can be input into the system, which predicts the appropriate customer cluster and provides a marketing strategy. This deployment makes the segmentation model accessible for use by the business, allowing decision-makers to leverage real-time data to target customers more effectively.