**Architecture Document**

Low Level Design (LLD)

**Clustering For Customer Segmentation & Understanding**

Revision Number: 1.8

Last date of revision: 30/05/2024

Rumanshu Chandekar

Shreyas Dongre

# Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| Date Issued | Version | Description | Author |
| 25 April 2024 | 1.1 | First Draft | Shreyas Dongre  Rumanshu Chandekar |
| 26 April 2024 | 1.2 | Added Workflow chart | Shreyas Dongre  Rumanshu Chandekar |
| 27 April 2024 | 1.3 | Added Exception Scenarios Overall, Constraints | Shreyas Dongre  Rumanshu Chandekar |
| 28 April 2024 | 1.4 | Added KPIs | Shreyas Dongre  Rumanshu Chandekar |
| 30 April 2024 | 1.5 | Added user I/O flowchart | Shreyas Dongre  Rumanshu Chandekar |
| 10 May 2024 | 1.6 | Added EHR, LSTM model diagrams | Shreyas Dongre  Rumanshu Chandekar |
| 15 May 2024 | 1.7 | Added dataset overview and updated user I/O flowchart. | Shreyas Dongre  Rumanshu Chandekar |
| 20 May 2024 | 1.8 | Restructure and reformat LLD | Shreyas Dongre  Rumanshu Chandekar |

**Contents**

[Document Version Control 2](#_gjdgxs)

[**Abstract** 4](#_30j0zll)

[1](#_1fob9te) Introduction 5

[1.1](#_3znysh7) Why this Low-Level Design Document? 5

[1.2](#_2et92p0) Scope 6

[1.3](#_tyjcwt) Constraints 6

[1.4](#_3dy6vkm) 1.3 Risks 6

[1.5](#_1t3h5sf) 1.4 Out of Scope 7

[2](#_4d34og8) Technical specifications 7

[2.1 Predicting Disease 7](#_2s8eyo1)

[2.2 Logging 7](#_17dp8vu)

[2.3 Database 7](#_3rdcrjn)

[3](#_26in1rg) Technology stack 8

[4](#_lnxbz9) Proposed Solution 9

[5](#_35nkun2) Model training/validation workflow 10

[6](#_1ksv4uv) User I/O workflow 12

[7](#_44sinio) Exceptional scenarios 13

**Abstract**

# This project employs clustering techniques to segment customers effectively, aiming to understand their behaviours and preferences. Through the analysis of various customer attributes such as purchase history, frequency, and spending patterns, we utilize clustering algorithms to identify groups of customers with similar characteristics. By categorizing customers into distinct segments, businesses gain insights into their diverse customer base, allowing for tailored marketing strategies and personalized services. This approach facilitates targeted communication, product recommendations, and promotions, ultimately enhancing customer engagement and satisfaction. Through the implementation of clustering for customer segmentation, this project demonstrates how businesses can leverage data-driven techniques to optimize customer relationships and drive business growth. This project utilizes the PyCaret machine learning library to implement clustering algorithms for customer segmentation. The objective is to identify distinct groups within the customer base to tailor marketing strategies effectively. Insights derived from this analysis will guide strategic decisions and enhance customer engagement.

# Introduction

## Why this Low-Level Design Document?

The purpose of this document is to present a detailed description of the "Clustering for Customer Segmentation & Understanding" project. It will explain the purpose and features of the system, the interfaces of the system, what the system will do, the constraints under which it must operate, and how the system will react to external stimuli.

This document is intended for both the stakeholders and the developers of the system and will be proposed to higher management for its approval. The main objective of the project is to segment customers into distinct groups based on their behaviours and characteristics using clustering techniques. Effective customer segmentation allows businesses to tailor their marketing strategies, improve customer service, and enhance overall customer satisfaction.

**Customer segmentation through clustering can:**

* Group customers with similar behaviours and preferences, enabling targeted marketing campaigns.
* Identify high-value customers and those at risk of churn for proactive engagement.
* Enhance product recommendations and personalized offers based on customer segments.
* Provide insights into customer demographics, purchasing patterns, and preferences.

Key customer information for segmentation includes:

* Customer Demographics: Age, gender, income, location, etc.
* Purchase History: Transaction records, frequency of purchases, and monetary value.
* Browsing Behaviour: Pages visited, duration of visits, and products viewed.
* Feedback and Reviews: Customer satisfaction scores and qualitative feedback.
* Engagement Metrics: Interaction with marketing campaigns, email open rates, and social media activity.

This project aims to utilize clustering algorithms to analyse and segment customer data effectively, providing actionable insights for strategic decision-making.

## Scope

This software system will be a web application designed to perform customer segmentation using clustering algorithms for better customer understanding and targeted marketing strategies. The system aims to analyse customer data to identify distinct segments based on behaviours and characteristics. More specifically, effective customer segmentation is crucial for personalized marketing, improving customer satisfaction, and optimizing business strategies.

The system is designed to segment customers based on information such as demographics, purchase history, browsing behaviour, and engagement metrics. The insights gained from these segments will help in:

* Targeted Marketing: Creating personalized marketing campaigns tailored to specific customer segments.
* Customer Retention: Identifying high-value customers and those at risk of churn to implement retention strategies.
* Product Recommendations: Offering personalized product recommendations based on segment preferences.
* Business Strategy: Informing business decisions with a deeper understanding of customer needs and behaviours.

## Constraints

1. Data Quality: The effectiveness of customer segmentation depends on the quality and completeness of the data collected.
2. Scalability: The system must be scalable to handle large volumes of customer data and complex clustering computations.
3. Privacy: Ensuring customer data privacy and compliance with data protection regulations is mandatory.
4. Real-Time Processing: Implementing real-time data processing capabilities for dynamic customer segmentation.

## Risks

1. Data Inaccuracy: Inaccurate or incomplete data can lead to incorrect segmentation results.
2. Algorithm Selection: Choosing inappropriate clustering algorithms may result in poor performance and misleading insights.
3. Integration Challenges: Integrating the segmentation system with existing CRM and marketing platforms may pose technical challenges.

## Out of Scope

# Predictive Analytics: While the focus is on clustering for segmentation, predictive analytics based on these segments is not within the scope of this project.

# Manual Data Collection: The project assumes automated data collection from various sources and does not cover manual data entry processes.

# 2.Technical specifications

## 2.1 Dataset Overview

The dataset comprises various attributes related to customer demographics, behaviour, and interactions, which are essential for effective clustering and segmentation. The primary goal is to identify distinct customer segments based on these attributes to enable targeted marketing and personalized customer experiences.

1. **Explanation CUST\_ID**: Unique identifier for each customer.
2. **BALANCE**: The balance amount on the customer's account.
3. **BALANCE\_FREQUENCY**: How frequently the balance is updated.
4. **PURCHASES**: Total amount of purchases made.
5. **ONEOFF\_PURCHASES**: Number of one-off purchases.
6. **INSTALLMENTS\_PURCHASES**: Number of purchases made in installments.
7. **CASH\_ADVANCE**: Amount of cash advances taken.
8. **PURCHASES\_FREQUENCY**: Frequency of purchases.
9. **ONEOFF\_PURCHASES\_FREQUENCY**: Frequency of one-off purchases.
10. **PURCHASES\_INSTALLMENTS\_FREQUENCY**: Frequency of installment purchases.
11. **CASH\_ADVANCE\_FREQUENCY**: Frequency of cash advances.
12. **CASH\_ADVANCE\_TRX**: Number of cash advance transactions.
13. **PURCHASES**\_TRX: Number of purchase transactions.
14. **CREDIT\_LIMIT**: Credit limit of the customer.
15. **PAYMENTS**: Number of payments made by the customer.
16. **MINIMUM\_PAYMENTS**: Minimum number of payments required.
17. **PRC\_FULL\_PAYMENT**: Percentage of full payments made.
18. **TENURE**: Tenure of the customer in months.

## 2.2 Predicting Disease

1. System Displays Segment Options: The system shows the available customer segments, such as "High Spend - Young Adults," "Budget Shoppers," and "Loyal Customers."
2. User Selects Segment Criteria: The user chooses criteria like "Age," "Annual Income," and "Spending Score" to focus on for segment prediction.
3. System Presents Input Fields: Based on the selected criteria, the system generates the necessary input fields for the user to fill out.
4. User Provides Information: The user enters their details, such as age, annual income, and spending score.
5. Data Handling and Validation: The system validates and preprocesses the user input to ensure it matches the required format for the clustering model. Predicting Customer Segment: The system uses a trained clustering model to predict which customer segment the user belongs to base on the provided information.
6. Display Predicted Segment: The system displays the predicted segment to the user and provides relevant insights or recommendations based on that segment.

## 2.3 Logging

In the "Clustering for Customer Segmentation & Understanding" project, logging every user activity is crucial for debugging and smooth operation. The system should log key steps and the entire process flow, with developers choosing between database or file logging. Extensive logging must not affect system performance. Effective logging helps monitor user actions, diagnose issues, and ensure system reliability, enabling quick problem resolution and an enhanced user experience.

## 2.4 Database

In the "Clustering for Customer Segmentation & Understanding" project, the system must store every user request and interaction in a database to facilitate easy model retraining. When a user selects criteria and provides required information, the system should log all data inputs into a database, such as MongoDB or MySQL. This ensures that all user data is captured systematically, allowing for efficient retrieval and analysis to improve and retrain the clustering model. Effective data storage is essential for maintaining accurate customer segmentation and enhancing the overall system performance.

**2.5 Deployment**

1. AWS EC2



# 3.Technology stack

For the "Clustering for Customer Segmentation & Understanding" project, a robust technology stack is essential to ensure efficient data handling, model training, and system performance. The technology stack includes:

Frontend: Stream lit for building an interactive and user-friendly web interface.

Backend: Python for implementing machine learning algorithms and clustering models using libraries such as Scikit-Learn and Pandas.

Database: MySQL for structured relational data storage, depending on the specific requirements for data retrieval and model retraining.

Machine Learning: Scikit-Learn for clustering algorithms (e.g., K-Means, DBSCAN). Pandas for data manipulation and preprocessing.

Deployment: Docker for containerizing applications to ensure consistency across different environments. AWS or Azure for scalable cloud hosting and database management.

|  |  |
| --- | --- |
| **Front End** | Stream lit |
| **Backend** | Python/ML |
| **Database** | MySQL |
| **Deployment** | AWS |

# 4.Proposed Solution

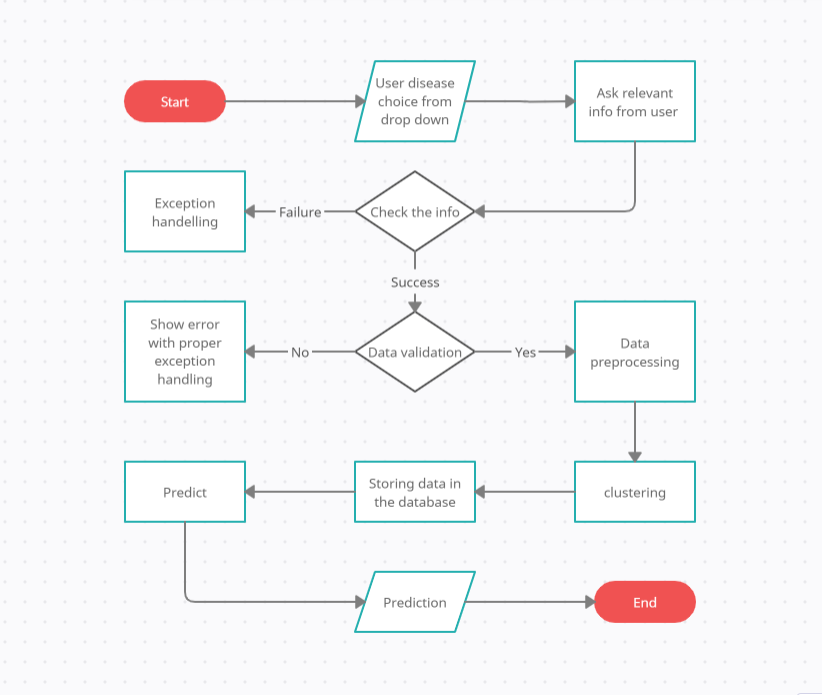
Your project, "Clustering for Customer Segmentation & Understanding," is all about organizing customers into groups based on their behaviours and preferences so that businesses can understand them better. Imagine you're sorting a big pile of puzzle pieces into different piles based on their colours and shapes. Similarly, you'll be sorting customers based on factors like what they buy, how often they buy it, where they live, and how they interact with the business. This way, you'll create clear groups of customers who are like each other in some way.

To do this sorting, you'll use clever computer techniques that can automatically group customers together. These techniques look at all the information about each customer and decide which group they belong to. You might also use some special tricks to make sure the groups make sense and are useful to the business. For example, you might combine similar groups or split big groups into smaller ones to get a clearer picture. Ultimately, the goal is to help businesses understand their customers in a detailed way, so they can offer products and services that meet their needs and keep them happy.

# 5.Model training/validation workflow



# 6.User I/O workflow

****

# 7.Exceptional scenarios

Exceptional scenarios in the context of "Clustering for Customer Segmentation & Understanding" encompass challenges such as handling outliers, imbalanced datasets, and noisy data, all of which can affect the accuracy of clustering results. Additionally, ensuring the interpretability of clusters and aligning them with business understanding poses another scenario. By proactively addressing these challenges through robust preprocessing, validation techniques, and visualization, the project aims to enhance the reliability of clustering for businesses seeking actionable insights into their customer base.

# 8.Key performance indicators (KPI)

* Efficiency and Accuracy: Measure the efficiency gains and accuracy improvement achieved through clustering techniques compared to traditional segmentation methods.
* Customer Insight: Gain in-depth understanding of customer segments' characteristics, preferences, and behaviours to inform targeted marketing strategies and personalized offerings.
* Marketing Effectiveness: Assess the effectiveness of marketing campaigns tailored to specific customer segments, evaluating response rates and overall campaign performance.
* Customer Retention and Satisfaction: Monitor customer retention rates, loyalty metrics, and satisfaction levels within each segment to optimize customer relationship management and enhance loyalty.
* Business Impact: Track the impact of clustering-driven strategies on revenue growth, profitability, and overall business performance, demonstrating the value of customer segmentation and understanding initiatives.