



**Clustering For Customer Segmentation & Understanding**

**High Level Design Document**

Domain: Machine Learning

Creator: Rumanshu Chandekar

Shreyas Dongre

Date: 29.04.2024

**Document Version Control**

|  |  |  |  |
| --- | --- | --- | --- |
| Date issued | Version | Description | Author |
| April 29, 2024 | 1.1 | First Draft | Shreyas Dongre  Rumanshu Chandekar |
| May 10, 2024 | 1.2 | Added unit test cases | Shreyas Dongre  Rumanshu Chandekar |
| May 20, 2024 | 1.3 | Correcting font sizes | Shreyas Dongre  Rumanshu Chandekar |

Contents

[Abstract 3](#_Toc110595993)

[Introduction 3](#_Toc110595994)

[What is High-Level Design Document? 3](#_Toc110595995)

[Scope 4](#_Toc110595996)

[General Description 4](#_Toc110595997)

[Definitions 4](#_Toc110595998)

[Product Description 4](#_Toc110595999)

[Problem Statement 4](#_Toc110596000)

[Proposed solution 5](#_Toc110596001)

[Further improvements 5](#_Toc110596002)

[Data requirements 5](#_Toc110596003)

[Tools used 5](#_Toc110596004)

[Hardware Requirements 7](#_Toc110596005)

[Constraints 7](#_Toc110596006)

[Assumptions 7](#_Toc110596007)

[Design Details 7](#_Toc110596008)

[Process Flow 7](#_Toc110596009)

[Event Log 8](#_Toc110596010)

[Error Handling 9](#_Toc110596011)

[Performance 9](#_Toc110596012)

[Reusability 9](#_Toc110596013)

[Application Compatibility 9](#_Toc110596014)

[Resource Utilization 9](#_Toc110596015)

[Dashboards 10](#_Toc110596016)

[KPIs (Key Performance Indicators) 10](#_Toc110596017)

[Conclusion 10](#_Toc110596018)

# Abstract

# This project employs clustering techniques to segment customers effectively, aiming to understand their behaviors 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

## What is High-Level Design Document?

The goal of this HLD or a high-level design document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

* Present all of design aspects and define them in detail
* Describe all user interfaces being implemented
* Describe the hardware and software interfaces.
* Describe the performance requirements.
* Include design features and architecture of the project.
* List and describe the non-functional attributes such as security, reliability, maintainability, portability, reusability, application compatibility. resource utilization, serviceability

## Scope

The HLD documentation presents the structure of the system, such as database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly technical terms which should be understandable to the administrators of the system.

# General Description

## Definitions

**Clustering**: A machine learning technique used to group a set of objects in such a way that objects in the same group are more like each other than to those in other groups.

**Customer Segmentation**: The practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing.

## System Overview

The system will use PyCaret to preprocess data, select optimal clustering algorithms, and apply these to segment customers.

Outputs will include cluster labels for each customer, key characteristics of each segment, and visualizations of the clusters.

## System Architecture

## High Level Architecture

Data ingestion module to collect and clean customer data.

Clustering module where PyCaret automates the selection and tuning of clustering algorithms, Visualization module to display clusters and interpret results.

## Detailed System Design

Data Preprocessing: Handling missing values, normalizing data, and encoding categorical variables.

Model Selection: Using PyCaret's comparison module to evaluate multiple clustering algorithms.

|  |  |
| --- | --- |
| **Term** | **Description** |
| Pycaret | Source for machine learning Algorithm |
| Database | Collection of the Information |
| Cloud | A data center full of services connected to the internet performing service |
| Python | Python Environment for coding |
| EDA | Exploratory data Analysis |
| Flask | Flask framework is used for developing web applications using python. |

Result Interpretation: Tools to analyze and visualize the characteristics of each customer segment.

## Product Description

Our project simplifies customer understanding by grouping them based on customer habits. We collect data, analyze it with smart algorithms, and create clusters of similar customers. This helps businesses personalize marketing efforts, leading to happier customers.

## Problem Statement

Our project aims to tackle the challenge of understanding diverse customer behaviors in businesses. By implementing clustering techniques, we seek to efficiently group customers based on their habits. This will enable businesses to tailor their strategies and offerings to meet the unique needs of each customer segment, ultimately improving customer satisfaction and driving business growth.

## Proposed solution

Our proposed solution involves utilizing clustering algorithms to analyze customer data and group them based on similarities in the customers’ behaviors. By implementing this approach, businesses can gain valuable insights into their customer base and tailor their marketing strategies accordingly. This allows for personalized communication, targeted promotions, and improved customer satisfaction.

## Further improvements

To improve our project, we can use smarter methods to group customers based on shopping habits. We'll make our system faster and more accurate. Also, we'll add automatic adjustments for changes in data and use predictions to anticipate customer behavior. These upgrades will keep our project valuable for businesses seeking to understand and serve their customers better.

Data requirements

Data requirement completely depend on our problem statement. We need the dataset from credit card to improve accuracy of the model. The required dataset should contain the following features:

**Customer Identification:** Unique identifier for each customer.

**Financial Metrics:**

BALANCE: The balance amount left in the account to make purchases.

BALANCE\_FREQUENCY: How frequently the balance is updated. PURCHASES: Total number of purchases made.

ONEOFF\_PURCHASES: Total amount of one-off purchases. INSTALLMENTS\_PURCHASES: Total amount of installment purchases. CASH\_ADVANCE: Total amount of cash advances taken.

**Purchase Behavior:**

PURCHASES\_FREQUENCY: Frequency of purchases. ONEOFF\_PURCHASES\_FREQUENCY: Frequency of one-off purchases. PURCHASES\_INSTALLMENTS\_FREQUENCY: Frequency of installment purchases. CASH\_ADVANCE\_FREQUENCY: Frequency of cash advances. CASH\_ADVANCE\_TRX: Number of transactions made with cash advances. PURCHASES\_TRX: Number of purchase transactions.

**Credit and Payments:** Credit limit for the customer.

**PAYMENTS:** Total payments made by the customer.

MINIMUM\_PAYMENTS: Minimum payments made by the customer. PRC\_FULL\_PAYMENT: Percentage of full payments made by the customer.

**Customer Tenure:** Tenure of the customer with the bank.

Tools Used

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, Azure, Pycaret, and a few other libraries were used to build the whole model.





 

* For visualization tasks, matplotlib, seaborn and plotly were used.
* Azure Cloud was used for deploying machine learning.
* Pycaret was used to predict final model.
* GitHub is used as version control system.
* NumPy and Pandas were used to clean and interpret data.
* Scikit-learn was used to cross-validate and compare different models.
* K-means Clustering was used to build the final model.

## Hardware Requirements

For our project, we'll need a computer with enough memory and processing power to handle large datasets. A modern laptop or desktop computer should suffice. Additionally, access to cloud computing services like AWS or Google Cloud can be beneficial for for scalability and faster computation. No special hardware is required beyond a standard computer setup.

* Windows Server, Linux, or any operating system that can run as a webserver, capable of delivering HTML5 content.
* Minimum 1.10 GHz processor or equivalent.
* Between 1-2 GB of free storage
* Minimum 512 MB of RAM
* 3 GB of hard-disk space

## Constraints

The front-end must be user friendly and should not need anyone to have any prior knowledge to use it.

## Assumptions

The main objective of this project is to implement the use case as previously mentioned (2.3 problem statement) for new dataset that comes through the UI. It is assumed that all aspects of this project can work together as the designer is expecting and also the data on which our model is trained is as correct as possible.

# Design Details

## Process Flow

For accomplishment of the task, we will use a trained Machine Learning model. The process flow diagram is shown below:

**Data Preparation**

**Model**

**Development**

**Deployment**

**Deployment**

Event Log

The system should log every event so that the user will know what process is running internally. Initial step-by-step description: 1. The system identifies at what level logging is required 2. The system should be able to log each system flow 3. Developer can choose logging method. You can choose database logging/ File logging as well 4. System should not hang even after so many loggings. Logging just because we can easily debug issues, so logging is mandatory to do.

## Error Handling

Errors should be encountered; an explanation will be displayed as to what went wrong? An error will be defined as anything that falls outside the normal intended usage.

# Performance

## For our "Clustering for Customer Segmentation & Understanding" project, accuracy is crucial in providing reliable insights to businesses. We focus on regular model retraining to enhance performance and ensure up-to-date customer segmentation. This approach prevents any misleading interpretations of customer behaviors, enabling businesses to make informed decisions confidently.

## Reusability

The code written and the components used should have the ability to be reused with no problems.

## Application Compatibility

The different components for this project will be using Python and pycaret as an interface between them, each component will have its own task to perform, and it is the job of Python and pycaret to ensure proper transfer of information.

## Resource Utilization

When any task is performed, it will likely use all the processing power available to it until finished.

## KPIs (Key Performance Indicators)

* Key Performance Indicators of MITVP
* Latency or the amount of time the application takes to display results for some specific input.
* The processing power our application takes to run
* The memory and RAM our application takes to run on a web server.

# Conclusion

In our project on clustering for customer segmentation and understanding demonstrates the effectiveness of data-driven approaches in enhancing business strategies. By efficiently utilizing computing resources and optimizing data processing techniques, we have successfully segmented customers based on their behaviors and preferences. This has empowered businesses to tailor their marketing efforts, leading to improved customer satisfaction and business outcomes. Moving forward, continuous monitoring and refinement of our methods will ensure ongoing success in leveraging data for customer insights and strategic decision-making. The project leverages PyCaret to streamline the application of clustering algorithms for effective customer segmentation. This approach will enhance marketing strategies by providing deeper insights into customer behaviors and preferences.