**High-Level Design ( HLD )**

Customer Personality Analysis

Revision Number – 1.0

# Last Date of Revision: 11 – 05 -2024

Harshal K Zarikar

**Document Version Control**

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Version | Description | Author |
| 11– 5 - 2024 | 1.0 | First Draft | Harshal Zarikar |

# **Contents**

Document Version Control……………………………………………………….. 2

Abstract…………………………………………………………………………… 4

1. Introduction ……………………………………………………………………. 5
   1. Why this HLD Document ? ………………………………………….. 5
   2. Scope ………………………………………………………………… 5
   3. Definitions ………………………………………………………… 5
2. General Description ……………………………………………………………… 6
   1. Problem Perspective …………………………………………………. 6
   2. Problem Statement ……………………………………………………… 6
   3. Proposed Solution ……………………………………………………….. 6
   4. Further Improvements …………………………………………………6
   5. Technical Requirements ……………………………………………………. 6
   6. Data Requirements …………………………………………………… 7
   7. Tools Used ……………………………………………………………….. 7
   8. Constraints ………………………………………………………………. 7
   9. Assumptions ………………………………………………………….. 7
3. Design Flow ……………………………………………………………………… 8
   1. Model Creation and Evaluation …………………………………………. 8
   2. Deployment Process …………………………………………………… 8
   3. Logging …………………………………………………………………… 8
   4. Error Handling …………………………………………………………… 9
4. Performance Evaluation ………………………………………………………… 9
   1. Reusability …………………………………………………………… 9
   2. Application Compatibility ……………………………………………... 9
   3. Resource Utilization …………………………………………………… 9
   4. Deployment …………………………………………………………….. 9
5. Conclusion ………………………………………………………. 9

**Abstract**

This High-Level Design (HLD) document outlines the design of a Customer Personality Analysis system. The system aims to segment the customer data into clusters to better understand customer preferences towards a product. The document describes the proposed solution, technical and data requirements, tools and technologies used, design details, and deployment process.

**1.Introduction**

The Customer Personality Analysis project is a machine learning-based solution to segment customers based on their information. The project aims to help companies to tailor their marketing and sales efforts to better target a product towards certain group of customers.

**1.1 Why this High-Level Design Document?**

This document serves as a blueprint for the development team to understand the proposed solution and how the system will be built.

**1.2 Scope**

The system will be designed to analyse the customer personality based on their income and expenses, education level. The system will use machine learning algorithms to analyze historical data and generate accurate quotes.

**1.3 Definitions**

**Machine learning:** A subset of artificial intelligence that uses algorithms to learn from data and make predictions or decisions.

Customer Analysis: process of identifying and understanding the unique characteristics and traits that make up an individual customer's personality

**2. General Description**

Customer personality analysis is the process of identifying and understanding the unique characteristics and traits that make up an individual customer's personality. This information can be used by companies to tailor their marketing and sales efforts to better target and serve each customer's specific needs and preferences. Traditionally, customer personality analysis has been done manually by marketing and sales teams, who would use their expertise and experience to identify common patterns and trends among customers. However, with the advent of data mining and machine learning, it is now possible to automate this process using algorithms that can analyse large amounts of data and identify common patterns and traits among customers. One type of machine learning algorithm that can be used for customer personality analysis is unsupervised learning. Unsupervised learning algorithms are trained on a large amount of data and can automatically detect patterns and similarities among customers without being explicitly told what to look for. This makes them particularly well suited for customer personality analysis, as they can uncover subtle differences and trends that might not be immediately apparent to humans

**2.1 Product perspective**

The system will be a standalone application that integrates with the company's existing systems to group customers based on their data.

**2.2 Problem statement**

The given problem statement is to perform clustering to summarize customers’ segments. Using clustering techniques to group similar customers together, companies can create customer segments with distinct personality characteristics, allowing them to tailor their marketing and sales efforts to better target each segment

**2.3 Proposed solution**

The proposed solution is to use machine learning algorithms to analyze historical data and generate clusters that group customers’ data.

**2.4 Technical requirements**

The technical requirements for the Customer Personality Analysis system include:

**Programming Language**: Python will be used as the primary programming language for building the system.

**Machine Learning Libraries**: Scikit-learn, Pandas, NumPy, and Matplotlib will be used as the primary machine learning libraries to develop and train the predictive models.

**Jupyter Notebook**: Jupyter Notebook will be used for data exploration, feature engineering, and model training.

**Data Storage**: The system will require a database to store customer information and historical policy data.

**Cloud Computing Platform**: A cloud computing platform like Streamlit will be used to deploy the system and to ensure scalability and availability.

**2.5 Data requirements**

The Customer Personality Analysis system will require the following data for generating accurate quotes:

**Customer personal information (age, gender, etc.):** The system will require information such as income, date of birth, education, marital status, and number of children of the customer.

**Expense information (fruits, gold year, etc.):** The system will require information such as the expenses on fruits, beverages, meat, gold, fish sweets.

**Accepted product data**: The system will require data on whether the customer accepted the product on first instance or on subsequent instances.

**2.6 Tools and technologies used**

**Python:** Python is a high-level programming language that is widely used for machine learning and data analysis.

**Scikit-learn:** Scikit-learn is a machine learning library in Python that provides simple and efficient tools for data mining and data analysis.

**Pandas:** Pandas is a Python library for data manipulation and analysis.

**NumPy**: NumPy is a Python library for scientific computing that provides support for large, multi-dimensional arrays and matrices.

**Matplotlib:** Matplotlib is a Python library for data visualization and plotting.

**Jupyter Notebook:** Jupyter Notebook is an open-source web application that allows users to create and share documents that contain live code, equations, visualizations, and narrative text.

**2.7 Constraints**

The system must be scalable to handle large amounts of data and capable of creating clusters in real-time. The system must maintain the confidentiality, integrity, and availability of customer data. The system should have security measures such as access controls, encryption, and data backups.. It should be able to integrate seamlessly with existing systems without disrupting the workflow.

**2.8 Assumptions**

Sufficient historical data is available for training and testing the machine learning algorithms. The model will be trained on the assumption that historical data accurately represents current and future trends. The system will be integrated with the insurance company's existing systems for seamless deployment. The system will be able to handle large amounts of data in real-time. The system will be deployed on the cloud for scalability and accessibility. The machine learning algorithms used in the system will be effective in accurately predicting insurance premiums. The system will be able to handle various types of insurance policies and customer information. The performance of the system will depend on the quality and quantity of input data. The system will adhere to industry standards and regulations regarding the use of customer data and privacy.

**3.0 Design Details**

**3.1 Process flow**

**Data collection and pre-processing:** The system will collect customer information from various sources such as online forms, databases, and insurance policies. The data will be pre-processed to remove any missing values or errors.

**Feature engineering:** The pre-processed data will be used to create new features such as age of the customers, re-organising the marital and education status. These features will be used as input variables for the machine learning algorithms.

**Model training and testing:** The system will use machine learning algorithms such PCA, KNN, DBSCAN, Agglomerative clustering algorithm, to train and test the models. The models will be evaluated based on metrics such as silhouette score.

**Model deployment:** The best-performing model will be deployed as a standalone application and integrated with the company's existing systems to creating clusters.

**3.2 Event log**

The system will log events such as data collection, feature engineering, model training, and model deployment for monitoring and troubleshooting purposes.

**3.3 Error handling**

The system will handle errors such as missing data, invalid data, and model prediction errors. Error messages will be logged and displayed to the user.

**3.4 Performance**

The system will be designed to handle large amounts of data and generate insurance premium quotes in real-time.

**3.5 Reusability**

The system will be designed to be reusable and scalable to handle different types of insurance policies and customer information.

**3.6 Application compatibility**

The system will be compatible with the insurance company's existing systems and databases.

**3.7 Recourse utilization**

The system will be designed to use minimal resources such as CPU and memory.

**3.8 Deployment**

The system will be deployed on the cloud to allow for scalability and accessibility.

**4. Dashboards:**

A dashboard is a visual representation of data that displays key metrics and information in an easily understandable format. It is typically designed to provide a quick and easy overview of the performance of a business or specific department. Dashboards can be customized to show relevant data and KPIs, such as sales, revenue, customer satisfaction, and productivity.

**4.1 Key Performance Indicators (KPIs)**

Key Performance Indicators (KPIs) for Customer Personality Analysis can vary depending on the specific goals and objectives of the company.

**5. Conclusion**

This HLD document outlines the design of an Customer Personality Analysis system that uses machine learning algorithms to create clusters to summarize customers segments based on their personal information. The system will be designed to be scalable, reusable, and user-friendly.