High Level Design (HLD)

High Level Design (HLD)

CLUSTERING FOR CUSTOMER SEGMENTATION AND UNDERSTANDING

Revision Number – 1.3

Last Date of Revision – 15/05/2022

Naveen Gupta

(Document Version Control

HIGH LEVEL DESIGN (HLD)

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Version | Description | Author |
| 12-05-2022 | 1.0 | Abstract, | Naveen |
|  |  | Introduction |  |
|  |  | Problem |  |
|  |  | Statement |  |
| 13-05-2022 | 1.1 | Design Flow | Naveen |
| 14-05-2022 | 1.2 | Performance | Naveen |
|  |  | Evaluation |  |

HLD)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Contents** | | | | | | | | | | | | | | | | | | | | | |
| [**Abstract**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.gjdgxs) | | | | | |  | | | | | | | | | | | | | | | |
|  |  |  |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| [**INTRODUCTION**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.30j0zll) | | | | | | | | | |  | | | | | | | | | | | |
|  |  |  |  | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | [**Why this HLD documentation?**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.1fob9te) | | | | | | | | | | | | | | | | | | | | |
|  |  |  |  | |  | |  |  |  | |  |  |  |  |  |  |  |  |  |  |  |
| [**1 Description**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.2et92p0) | | | | | | | |  | | | | | | | | | | | | | |
|  |  | |  |  |  | |  |  |  | |  |  |  |  |  |  |  |  |  |  |  |
|  | [**1.1**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.tyjcwt) | |  |  | [**Problem Perspective**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.tyjcwt) | | | | | | | | | | | | | | | | |
|  |  | |  | |  | |  | |  | |  |  |  |  |  |  |  |  |  |  |  |
|  | [**1.2 Problem Statement**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.3dy6vkm) | | | | | | | | | | | | | | |  | | | | | |
|  |  | | | | | |  | |  | |  |  |  |  |  |  |  | |  |  |  |
|  | [**1.3 Proposed Solution**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.1t3h5sf) | | | | | | | | | | | | | | |  | | | | | |
|  |  | | | | | |  | |  | |  |  |  |  |  |  |  | |  |  |  |
|  | [**1.4 Solution Improvements**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.4d34og8) | | | | | | | | | | | | | | | | | | | | |
|  |  | | | | | |  | |  | |  |  |  |  |  |  |  | |  |  |  |
|  | [**1.5 Technical Requirements**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.2s8eyo1) | | | | | | | | | | | | | | | | | | |  | |
|  | [**1.6 Data Requirements**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.17dp8vu) | | | | | | | | | | | | | | |  | | | | | |
|  |  | | | | | |  | |  | |  |  |  |  |  |  |  | | | |  |
|  | [**1.7 Tools Used**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.3rdcrjn) | | | | | | | | | |  | | | | |  | | | | | |
|  |  | |  |  | | |  | |  | |  |  |  |  |  | |  | | | |  |
|  | [**1.8**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.26in1rg) | |  | [**Constraints**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.26in1rg) | | | | | | | |  | | | | | | | | | |
|  |  | | | | | |  | |  | | |  |  |  |  | |  | | | |  |
|  | [**1.9 Assumptions**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.lnxbz9) | | | | | | | | | | | | | | | | | | | | |
|  | |  | | | | |  | |  | | |  |  |  |  | |  | | | |  |
| [**2 Design Flow**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.35nkun2) | | | | | | | | |  | | | | | | | | | | | | |
|  |  | | | | | |  | |  | | |  |  | |  | |  | | | |  |
|  | [**2.1 Modelling Process**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.1ksv4uv) | | | | | | | | | | | | | | | | | | | | |
|  |  | | | | | |  | | | | |  |  | |  | |  | | | |  |
|  | [**2.2 Deployment Process**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.44sinio) | | | | | | | | | | | | | | | | | | | | |
|  | |  | |  | | |  | | | | |  |  | | | |  | | | |  |
| **2.3** | | | | **Logging** | | | | | | | | | | | | | | | | | |
| **2.2** | | | | **Error Handling** | | | | | | | | | | | | | | | | | |
| [**3 Performance Evaluation**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.4i7ojhp) | | | | | | | | | | | | | | | | |  | | | | |
|  | | |  |  | | |  | | | | |  |  | | | |  | | | |  |
|  | [**3.1**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.2xcytpi) | |  | [**Reusability**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.2xcytpi) | | | | | | | |  | | | | |  | | | | |
|  |  | |  |  | | |  | | | | |  |  | | | |  | | | |  |
|  | [**3.2**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.1ci93xb) | |  | [**Application Compatibility**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.1ci93xb) | | | | | | | | | | | | | | | | | |
|  | [**3.3**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.3whwml4) | |  | [**Resource Utilization**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.3whwml4) | | | | | | | | | | | | |  | | | |  |
|  |  | | | | | |  | | | | | |  | | | |  | | | | |
|  | [**3.2 Deployment**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.2bn6wsx) | | | | | | | | | | | | | | | |  | | | | |
|  |  | | | | | |  | | | | | |  | | | | | | | | |
| [**Conclusion**](https://docs.google.com/document/d/1ZLlZEVdMuxYl3o0PziQjZcUbNmAcz-PR/edit#heading=h.qsh70q) | | | | | | | | | | | | | | | | | | | | | |

**Abstract**

Effective decisions are mandatory for any company to generate good revenue.

In these days competition is huge and all companies are moving forward with their own different strategies .We should use data and take a proper decision. Every person is different from one another and we don’t know what he/she buys or what their likes are.

But, with the help of machine learning technique one can sort out the data and can find the target group by applying several algorithms to the dataset .Without this ,It will be very difficult and no better techniques are available to find the group of people with similar character and interests in a large dataset .Here ,The customer segmentation using K-Means clustering helps to group the data with same attributes which exactly helps to business the best .We are going to us elbow method to find the number of clusters and at last we visualize the data . And at the end we can do prediction on groups using classification machine learning algorithms.

**1 Introduction**

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

The main purpose of this HLD documentation is to feature the required details of the project and supply the outline of the machine learning model and also the written code. This additionally provides the careful description on however the complete project has been designed end-to-end.

**1.2 Description**

**Problem Perspective**

The Customer Segmentation Prediction may be a machine learning model that helps Business to forecast the sale of the store’s products and helps the users to manage the sale of the products.

**1.3 Problem Statement**

Not all customers are same. To know which group is your customer and their

preferences is a big part for success in your business. Unsupervised machine learning

can help marketers to know their audience globally and engage them with their products

accordingly.

Here we can classify millions of people’s interests through their social media activity

and also through other surveys online & offline and cluster them in specific group of

their interest.

**1.4. Project Solution**

The solution proposed to take the required input of user from the created interface and process all the provided data to meet the requirements of the machine learning model and finally display the output saying customer segmentation prediction based on the new customer enteries.

**1.5 Technical needs**

There are not any hardware needs needed for victimization this application, the user should have AN interactive device that has access to the web and should have the fundamental understanding of providing the input. And for the backend half the server should run all the package that's needed for the process and provided information to show the results.

**1.6 Information needs**

The info demand is totally supported the matter statement. and also, the information set is accessible on the Bank Website within the type of standout sheet(.xlsx), because the main theme of the project is to induce the expertise of real time issues, we have a tendency to once more mercantilism {the information into the prophetess data base and commerce it into csv format.

**1.8Tools Used**

* Python 3.9 is employed because the programming language and frame works like K-means , numpy , pandas, sklearn and alternative modules for building the model.
* Visual Studio Code is employed as IDE.
* For visualizations seaborn , plotly and components of matplotlib are getting used.
* For information assortment prophetess info is getting used.
* Front end development is completed victimization Streamlit.
* Flask is employed for each information and backend readying.
* GitHub is employed for version management.
* AWS is employed for deployment.
* SQL SERVER IS USED FOR DATABASE.
* Power Bi is used for creating a report.

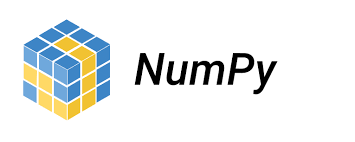
**1.9 Constraints**

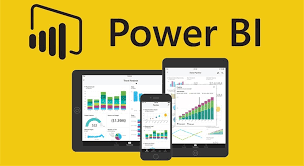
The Customer Segmentation Prediction answer should be user friendly, as automatic as attainable and also the user should not be needed to understand any of the operating.

**1.10 Assumptions**

The most objective of the project is to implement the utilization cases as for the new dataset that user provides through the programme. Machine learning model is employed for process the on top of computer file. It's additionally assumed that each one aspects of this project have the flexibility to figure along within the approach as the designer is expecting.





**2.1 and 2.2 Design Flow and Deployment Process**

PYTHON

SQL SERVER

Data (CSV)

EDA

DATA PREPROCESSING

IMPORT PYTHON LIBRARIES & READ DATA

FEATURE SELECTION

FITTING MODEL WITH K-MEANS & LIGHTGBM

FEATURE ENGINEERING

CREATING A WEB PAGE FOR DEPLOYMENT

SAVE MODEL IN PICKLE FILE

HYPER PARAMETER TUNNING

DEPLOYMENT ON LOCAL HOST USING STREAMLIT AND VS CODE

EXPORT DATA BACK TO SQL SERVER

DEPLOYMENT ON AWS

CREATING POWER BI REPORT

INPUT VALUE & PREDICT FINAL RESULT

ARCHITEC

**2.3 Logging**

Each step is being logged within the system that runs internally, that shows the date time and therefore the processed that has been performed, work is completed in several layers as information, DEBUG, ERROR, WARNINGS. this provides US the perceive of the logged info.

**2.4 Error Handling**

Once ever a slip is occurred, the reason are logged in its several log file, in order that the developer will rectify the error.

ARCHITECTURE

**3 Performance analysis**

**3.1 Reusability**

Elements of the code written is accustomed different applications and therefore the rest is changed and be reused.

**3.2 Application Compatibility**

The various parts for this project are exploitation python as associate interface between them. Every element can have its own tasks to perform, and it's the work of the python to make sure correct transfer of data.

**3.3 Resource Utilization**

Once any task is performed, it'll doubtless; use all the process power offered till that performs is finished.

**3.4 Deployment**

The model is being deployed on AWS.

**Conclusion**

The CUSTOMER SEGMENTATION PREDICTION will predict the worth supported the trained knowledge set within the rule. Therefore, the user will recognize the approximate result for his or her product.