Clustering-for-Customer-Segmentation-Understanding

High Level Design (HLD)

[Pick the date]

Contents

Abstract

Not all customers are the same. To know which group is your customer and their Preferences are a big part of success in your business. Unsupervised machine learning can help marketers 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 and offline, and cluster them into a specific group of their interest.

Introduction

Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) 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 the design aspects and define them in detail
- Describe the user interface being implemented
- Describe the software interfaces
- Describe the performance requirements
- Include design features and the architecture of the project

List and describe the non-functional attributes like:

- security
- reliability
- maintainability
- portability
- reusability
- application compatibility
- resource utilisation
- serviceability

Scope

The HLD documentation presents the structure of the system, such as the 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

Product Perspective

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.

Problem statement

A case requires developing a customer segmentation to give recommendations like saving plans, loans, wealth management, etc. on target customers groups.

PROPOSED SOLUTION

In the context of customer segmentation, customer clustering analysis is the use of a mathematical model to discover groups of similar customers based on finding the smallest variations among customers within each group. These homogeneous groups are known as "customer archetypes" or "personas".

The goal of cluster analysis in marketing is to accurately segment customers in order to achieve more effective customer marketing via personalization. A common cluster analysis method is a mathematical algorithm known as *k-means cluster analysis*, sometimes referred to as scientific segmentation. The clusters that result assist in better customer modelling and predictive analysis and are also used to target customers with offers and incentives personalized to their wants, needs, and preferences.

Technical Requirements

- **Scalability**: We need highly scalable clustering algorithms to deal with large databases.
- **Ability to deal with different kinds of attributes:** algorithms should be capable of being applied to any kind of data, such as interval-based (numerical) data, categorical data, and binary data.
- **Discovery of clusters with attribute shape**: The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bound to only distance measures that tend to find spherical clusters of small sizes.
- **High dimensionality**: The clustering algorithm should not only be able to handle low-dimensional data but also high-dimensional space.
- **Ability to deal with noisy data:** databases contain noisy, missing, or erroneous data. Some algorithms are sensitive to such data and may lead to poor-quality clusters.
- **Interpretability**: The clustering results should be interpretable, comprehensible, and usable.

Dataset

The sample dataset summarizes the usage behaviour of about 9,000 active credit cards. Holders during the last 6 months. The file is at the customer level, with 18 behavioural variables.

- Variables of the Dataset
- Balance
- Balance Frequency
- Purchases
- One-off Purchases
- Instalment Purchases
- Cash Advance
- Purchases Frequency
- One-off Purchases Frequency
- Purchases Instalment Frequency
- Cash Advance Frequency
- Cash Advance
- TRX Purchases
- TRX Credit Limit
- Payments
- Minimum Payments
- PRC
- Full Payment
- Tenure Cluster

.

Tools used:

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, Jupiter notebook, Git-Github, VScode, Stremlit are used to build the whole model.

- The **Python** programming language used to build machine learning algorithms.
- **Pandas** are chiefly used for machine learning in the form of DataFrames. Pandas allows for importing and exporting tabular data in various formats, such as CSV or JSON files.
- **NumPy** is a Python library used for working with arrays.
- The **sklearn library** contains a lot of efficient tools for machine learning and statistical modelling including classification, regression, and clustering and dimensionality reduction.
- Building the **Streamlit App**. Creating the Streamlit UI. Loading the Saved Model & Making Real-Time Predictions. Deploying Machine Learning Models with Python and Streamlit.
- **GitHub** hosts a large collection of open-source machine-learning projects. A GitHub repository is the Git folder inside a project. This repository tracks all changes made to files in your project, building history over time.
- Python **Seaborn** library is a widely popular data visualization library that is commonly used for data science and machine learning tasks.

Constraints

It requires specifying the number of clusters (k) in advance. It cannot handle noisy data and outliers. It is not suitable to identify clusters with non-convex shapes.

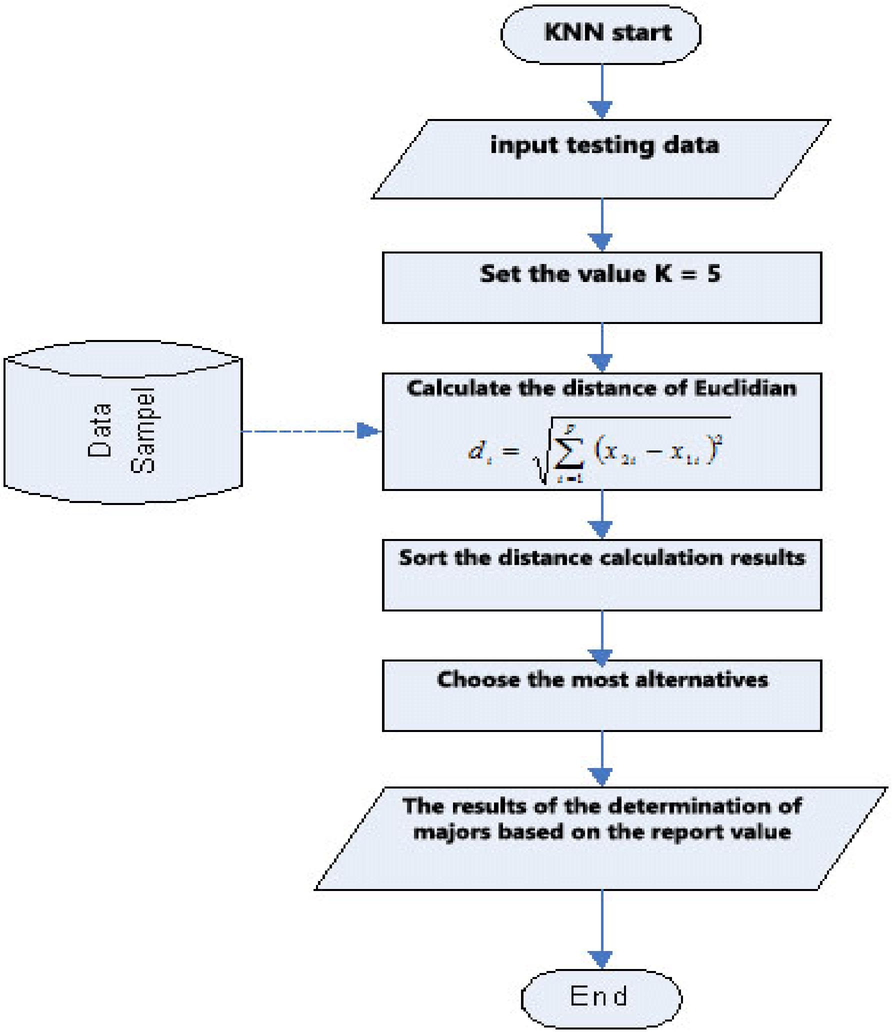
Assumptions

The goal of cluster analysis in marketing is to accurately segment customers in order to achieve more effective customer marketing via personalization. A common cluster analysis method is a mathematical algorithm known as *k-means cluster analysis*, sometimes referred to as scientific segmentation. The clusters that result assist in better customer modelling and predictive analysis and are also used to target customers with offers and incentives personalized to their wants, needs, and preferences.

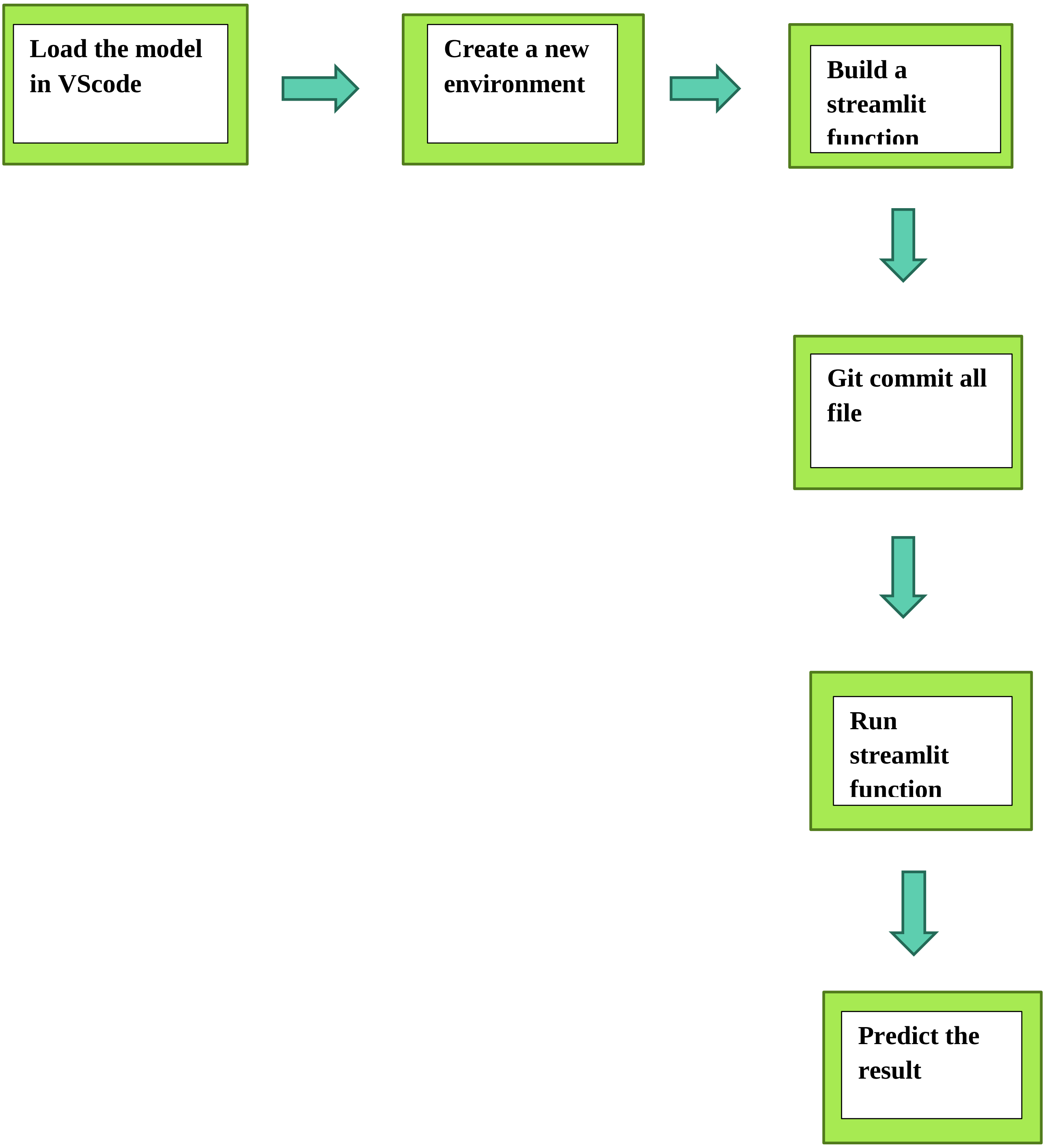
Design detail

Process Flow

The flow chart for the KNN algorithm given below


KNN start
input testing data
Set the value K = 5
Data Sampel
Calculate the distance of Euclidian
$d_i = \sqrt{\sum_{i=1}^{p} (x_{2i} - x_{1i})^2}$
Sort the distance calculation results
Choose the most alternatives
The results of the determination of majors based on the report value
End

Deployment Process



Load the model in VScode

Create a new environment

Build a streamlit function

Git commit all file

Run streamlit function

Predict the result

Event log

The system should log every event so that the user will know what process is running internally.
Initial Step-By-Step Description:

- The System identifies at what step logging required
- The System should be able to log each and every system flow.
- Developer can choose logging method. You can choose database logging/ File logging as well.
- System should not hang even after using so many loggings. Logging just because we can easily debug issues so logging is mandatory to do.

Error Handling

Should errors be encountered, an explanation will be displayed as to what went wrong? An error will be defined as anything that falls outside the normal and intended usage.

Performance

The goal of cluster analysis in marketing is to accurately segment customers in order to achieve more effective customer marketing via personalization. A common cluster analysis method is a mathematical algorithm known as *k-means cluster analysis*, sometimes referred to as scientific segmentation. The clusters that result assist in better customer modelling and predictive analysis and are also used to target customers with offers and incentives personalized to their wants, needs, and preferences. Also, model retraining is very important to improve the performance.

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 as an interface between them. Each component will have its own task to perform, and it is the job of the Python to ensure proper transfer of information.

Resource Utilization

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

Deployment

Building the **Streamlit App**. Creating the Streamlit UI. Loading the Saved Model & Making Real-Time Predictions. Deploying Machine Learning Models with Python and Streamlit.

Key performance indicator

- Simple and easy to implement: The k-means algorithm is easy to understand and implement, making it a popular choice for clustering tasks.
- Fast and efficient: K-means is computationally efficient and can handle large datasets with high dimensionality.
- Scalability: K-means can handle large datasets with a large number of data points and can be easily scaled to handle even larger datasets.
- Flexibility: K-means can be easily adapted to different applications and can be used with different distance metrics and initialization methods.

Conclusion

Based on this clustering method, I have developed a Streamlit application where I'll take customer information and determine which cluster the consumer belongs to. Better consumer modelling and predictive analysis are made possible by the resulting clusters, which are also utilized to target customers with offers and incentives that are catered to their wants, requirements, and preferences.

Reference

- https://link.springer.com/article/10.1007/s42452-019-1356-9
- https://www.google.com/
- https://www.youtube.com/
- https://www.javatpoint.com/