Customer Segmentation and Classification

By: Thomas Antony

Revision No:0.1

Last date of revision:22/01/2024

Customer Segmentation and Classification

Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Version | Description | Author |
| 22/02/2024 | 0.01 | First Draft | Thomas Antony |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Contents

[General Description 4](#_Toc160109299)

[Product Description 4](#_Toc160109300)

[**Key Features:** 4](#_Toc160109301)

[2. System Architecture: 5](#_Toc160109302)

[Components: 5](#_Toc160109303)

[Interactions: 5](#_Toc160109304)

[Workflow : 6](#_Toc160109305)

[ Data Ingestion. 6](#_Toc160109306)

[ Data Validation: 6](#_Toc160109307)

[ Data Transformation: 6](#_Toc160109308)

[ Apply K-Means clustering to group customers based on demographics and psychographics features. 6](#_Toc160109309)

[ Model Evaluation Module: 6](#_Toc160109310)

[ Develop API endpoints to provide real-time marketing recommendations. 6](#_Toc160109311)

[. Key Technologies: 7](#_Toc160109312)

[ Clustering Algorithm: 7](#_Toc160109313)

[ Programming Languages: 7](#_Toc160109314)

[ Tools and Libraries: 7](#_Toc160109315)

[Design Details 8](#_Toc160109316)

[Model Training and Evaluation 8](#_Toc160109317)

[. API Module: 8](#_Toc160109318)

[Data and Model Versioning 9](#_Toc160109319)

[ DVC 9](#_Toc160109320)

[ MLFLOW 9](#_Toc160109321)

[Event Log 9](#_Toc160109322)

[Error Handling 9](#_Toc160109323)

[**Performance** 10](#_Toc160109324)

[Reusability 10](#_Toc160109325)

[Application Compatibility 10](#_Toc160109326)

[Resource Utilization 10](#_Toc160109327)

# General Description

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.

The project is an Unsupervised Machine learning project, this project can be made into semi supervised as well by using the segmented or clustered predictions as the target variables for classification

## Product Description

An application for predicting customer segments for the given data to help marketers in analysis of key performance indicators of the business to provide tailored and personalised experience with the product or services

## **Key Features:**

* Customer Segmentation
* API Endpoint for Predicting new data
* Segmented Data for Dashboarding

# 2. System Architecture:

## Components:

* + Data Ingestion Module
  + Data Validation Module
  + Data Transformation Module
  + Customer Segmentation Module/Model Training Module
  + Model Evaluation Module
  + API Module/Prediction Module

## Interactions:

* + Data Ingestion Module collects transaction and customer data.
  + Data Validation Module Validated the data structure to see if the data format and features match to the fixed features
  + Data Transformation Module transforms the data to the required format for the machine learning model
  + Customer Segmentation Module uses K-Means clustering to create segments or fits or trains the model based on the data.
  + Model Evaluation model keeps track of the performance of the model
  + API Module enables real-time predictions for targeted marketing.

# Workflow :

### Data Ingestion.

* + Utilize the Data Ingestion Module to fetch data from the specified Amazon S3 bucket.
  + Use DVC for data versioning to track changes in the ingested data.

### Data Validation:

* + Employ the Data Validation Module to ensure that the ingested data adheres to the expected structure.
  + Verify if the data format and features match the predefined fixed features.

### Data Transformation:

* + Leverage the Data Transformation Module to preprocess and transform the data into the required format for machine learning model input.

Customer Segmentation / Model Training for Targeted Marketing:

### Apply K-Means clustering to group customers based on demographics and psychographics features.

* + Within the Customer Segmentation Module, use K-Means clustering to create segments based on customer features.
  + If necessary, utilize the same module to train a supervised model, although we are not implementing semi supervised as the project description does not imply too

Model Evaluation:

### Model Evaluation Module:

* + Introduce a Model Evaluation Module to keep track of the performance of the clustering and supervised models.
  + We have used mlflow here for tracking experiments with different logged parameters and logged evaluation metrics

API Module / Prediction Module:

### Develop API endpoints to provide real-time marketing recommendations.

* + In the API Module (or Prediction Module), create endpoints that utilize the clustering results or the trained supervised model to provide real-time marketing recommendations.

# . Key Technologies:

### Clustering Algorithm:

* + K-Means

### Programming Languages:

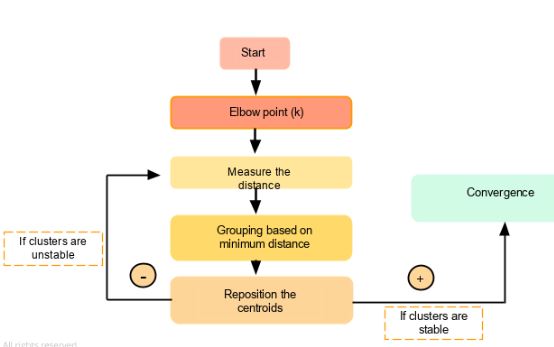
* + Python for data processing, clustering, and API development.
  + Html for simple API UX for predicting user inputs

### Tools and Libraries:

* + Numpy
  + Pandas
  + Scikit learn
  + python-box==6.0.2
  + pyYAML
  + ensure==1.0.2
  + types-PyYAML
  + Flask
  + joblib
  + boto3
  + python-dotenv
  + AWS S3
* Version Control
  + Git
  + Mlflow
  + DVC

# Design Details

## Model Training and Evaluation



## . API Module:

* **Responsibilities:**
  + Create RESTful API using Flask/FastAPI.
  + Implement endpoints for real-time Customer Segments predictions
* **Endpoints:**
  + / : Index.html links to predict page
  + **/predict :** Html form for inputs by the client
  + **/train :**For retraining of the whole application if model degradation spotted
  + **/results :**Redirected to results after submition is prediction page
* **Input:**
  + Customer data (features) as the customer enters the shop.
* **Output:**
  + Identifies the Customer segment of the entered customer

## Data and Model Versioning

## DVC

The data must be version control to identify data drifts that could potentially influence model performance and DVC help us to keep versions of the data we are only tracking the hash key of the current version and the cache and other files are not being tracked

## MLFLOW

MLFLOW help us to log the model , evaluation metrics and parameters used so that we can track different model versions and ml flow and also help us to compare different models and to select the best model .This will also help us to track model drift

## Event Log

The system should log every event so that the user will know what process is running internally and where has the systems produced errors

**Initial Step-By-Step Description**:

1. The System identifies **at** what step logging required

2. The System should be able to log **each** and every **system flow**.

3. Developer can choose **logging** method. You **can** choose database logging**/** File

logging as **well**.

4. 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.

We have created a custom exception function for extra ease of exception handling

## **Performance**

The Customer Segmentation app provides segmented results the performance of the model is measured using silhouette score and its been tracked by mlflow for different experiments performed

### 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.

**Conclusion**

The Customer Segmentation prediction application performs quite well and clients can use it for

Targeted marketing, customer satisfaction and personalised experience etc

**Constraints and Further Improvements:**

Constraints I encountered was the lack of system resource for creating docker containers and implementing airflow for automated training and these both constraints blocked me from hosting the web application on AWS these are further improvements that can be build upon the current application next improvements include UX design implementing a targeted Marketing API endpoint for the predicted customer UX wise there are lot to be done and ML wise Supervised Algorithms can be used on top of semi supervised and use semi supervised only for Api end point prediction but this can further increase the complexity of the project