Advanced Market Segmentation Using Deep Clustering

Phase 1: Problem Definition and Data Understanding

1.1 Project Overview

The global goal of this project is to transform the way that market segmentation is currently done, with new opportunities offered by deep learning and unsupervised machine learning. Classic clustering techniques although useful in certain instances fail to address problems of high dimensionality and non-linearity inherent based on customer purchasing behavior and preferences. This work fills this gap by using deep clustering algorithms which combine the strength of deep learning and clustering.

The primary contribution is based on the employ of using autoencoder deep learning models to obtain latent features from the considered data sources. It is conducive in making clusters that reflect real customer segments as they are in the market. The ultimate goal is to help businesses tailor and improve all sorts of customer engagements, sales and marketing tactics, and their products based on deeper audience insights. Further, the project investigates the possibility of applying these clusters when analyzing changing, growing datasets, which makes them relevant to the quickly changing markets.

1.2 Objective of the Project

Core Objective: The goal of this work is to propose a clustering framework that is easily extensible, and can effectively capture hidden customer segments. It involves grouping customers through employ of multidimensional behavioral, demographic as well as psychographic data employing techniques of unsupervised learning.

Secondary Objectives:

- Improve the decision-making value of customer clusters at the firm for better business outcomes.
- Create an elastic system that can take streaming data and help adapt segmentation techniques in near-real time to assist businesses in making better decisions.

Target Users:

- This project is valuable for a broad spectrum of users, including:
- Businesses and Marketers: To create specific approaches and adverting campaigns which meet a particular audience's needs.
- Data Scientists and Analysts: Looking for an opportunity to use modern clustering approaches to obtain operational knowledge.
- Product Teams: To feed back into the product development and lifecycle decision making process with regards to the clusters.

1.3 Dataset Overview and Data Requirements

To achieve meaningful segmentation, the dataset must encompass a wide range of customer-related attributes:

Features:

- Demographics: Basic demographics such as age, sex, income and level of education.
- Behavioral Data: The buying behaviour which includes previous purchase behaviour, consumption behaviour and product behaviour.
- Engagement Metrics: Website engagement, email open rates, and response rates, time spent within an application.
- Psychographic Information: Information that customers and their values, attitudes and interests and that can be collected by surveys or through other means of user-generated data.
- Temporal Patterns: Any data that has a time component including the time of transaction, seasonality and cyclic characteristics.

Dataset Format:

- A combination of ordered (numeric) data and unordered (alphabetic) data.
- Categorical variables, numerical variables, and temporal variables are supported.
- Data needs to be pre-processed so it fits into deep learning models; categorical data must be encoded, and numerical data normalized.

1.4 Data Sources

A robust segmentation project requires data from diverse sources to ensure comprehensive coverage:

Public Datasets:

- For the initial model development and to compare with the best for large datasets, sources such as UCI Machine Learning and Kaggle are appropriate.
- Examples of such data include, financial transaction data or e-commerce customers' records of transactions.

Proprietary Data:

- CRM systems that have the detailed record of customers' interactions.
- Longitudinal bases of loyalty programs that contain information about customers' behavior.

Crowdsourced and Social Media Data:

- Yelp's reviews, ratings, and comments or the reviews, ratings, and comments on products in Amazon.

- Social media post, tweet or status engagement data from Twitter or Instagram gathered through API or web crawling.

Web Scraping:

- Live feeds from online stores to monitor patterns of products' demand and pricing mechanisms.

1.5 Initial Data Exploration

The first step is always a process of checking the data and getting a feel of the data set. The exploration process will include:

- **Missing Data Analysis:**
  As a result of data missing, some features might be missing values and to fill these gap the following imputation techniques are used depending on the data type for numerical type median imputation is used.
- **Outlier Detection:**
  By first employing mathematical techniques to detect outliers and employ data representations such as box plots to manage the facet.
- **Data Distribution Analysis**:
  Looking at the distribution of features, how far they move away from the mean and if some basic transformation like logarithm or power transformation will help.
- **Correlation Matrix Analysis:**
  To enhance the feature selection and the dimensionality reduction, the importance level of variables and their relations is determined.
- **Exploratory Visualizations:**
  Other methods like using t-SNE or UMAP to try to visualize other concepts or directions hidden in the data set.

1.6 Preprocessing Objectives

To prepare the dataset for deep clustering, preprocessing steps will involve:

- **Feature Engineering:**
  Starting new variables that define useful and hidden characteristics of customers or using multiple features to improve predictive models.
- **Scaling and Normalization:**
  Using techniques such as RobustScaler to adequately work on outliers or skewed data distributions real-life data.
- **Dimensionality Reduction:**
  What was done were in employing PCA or similar procedures to decrease the dataset complexity while preserving important information.

- **Text Preprocessing:**
  For unstructured data, the text data, natural language processing (NLP) methods such as tokenization, stemming and vectorization will be used.
- **Handling Class Imbalances:**
  Despite being unsupervised which could lead to overwhelming clusters especially for one, this can be handled by ensuring that each cluster has enough data by duplicating the data where necessary.

1.7 Conclusion of Phase 1

The first phase tries to lay the foundation for the realization of high-level clustering for the attainment of advanced market segmentation. Finally, this phase makes sure that all the objectives of the project, technical and business, are met by making sure the problem domain and data characteristics are well understood. The information that will be obtained in this phase will be used immediately in the construction of further deep clustering models which will enhance the accuracy of the segmentation.