Customer Journey Analysis Using Clustering and Dimensionality Reduction: Enhancing User Experience

Phase 1: Problem Definition and Data Understanding

---

1. Project Overview

In today's competitive e-commerce landscape, understanding the customer journey is essential for improving the user experience. The goal of this project is to analyze the customer journey on a cloth shopping website using advanced techniques like clustering and dimensionality reduction. By doing so, we aim to identify distinct customer segments based on their behaviors, preferences, and interactions with the website, ultimately improving personalization and targeting.

- **Objective**:
    - To enhance the overall shopping experience by understanding customer behavior, segmenting users into meaningful groups, and providing tailored marketing strategies and recommendations.

- **Why this is Important**:
    - Customers interact with websites in a variety of ways, and these interactions often reveal unique preferences and needs. By analyzing these patterns, businesses can improve customer engagement, create personalized experiences, and increase conversion rates.

---

2. Objective of the Project

The central objective of this project is to uncover hidden patterns within customer behaviors and use them to group similar customers together. This will allow us to create segments based on different attributes, such as purchasing patterns, age, gender, and engagement, all without needing predefined labels.

- **Core Objective**:
    - Implement clustering algorithms to group customers based on their behaviors, preferences, and demographics.

- **Target Audience**:
    - **E-commerce Businesses**: Businesses looking to enhance their marketing strategies, improve user experience, and boost sales by targeting specific customer segments.

- **Marketing Teams**: Teams who can use the segments for tailored email campaigns, promotions, and recommendations.

- **Customer Support Teams**: Support teams can use segmentation to better understand customer issues and offer personalized solutions.

- **Potential Benefits**:

  - **Personalized Shopping Experiences**: Providing customers with tailored product recommendations based on their past behaviors.

  - **Targeted Marketing Campaigns**: Segmenting customers allows marketing teams to craft specific offers for different groups, making marketing more effective and efficient.

  - **Improved Customer Service**: Understanding different customer segments helps the customer support team anticipate issues and provide better support based on customer history.

  - **Better Inventory Management**: With clear customer segments, businesses can stock more of what specific customer groups are likely to buy.

---

3. Dataset Overview and Data Requirements

For the project to be successful, the dataset must contain rich customer data, including demographics, behavioral information, and engagement metrics. This dataset will be crucial in training the clustering models and will require careful preprocessing to ensure that it's ready for analysis.

- **Key Features to Include**:

  - **Demographic Information**: Age, gender, location, etc. These provide context about who the customers are.

  - **Behavioral Data**: Insights into how customers behave on the website—things like purchase frequency, product categories purchased, total amount spent, and session duration. This helps in understanding customer habits.

  - **Engagement Data**: Information on how customers interact with marketing campaigns, product reviews, website clicks, and other digital touchpoints. It's critical to understand what drives customer actions.

  - **Shopping Preferences**: What products do customers prefer based on their behavior (e.g., size, color, or category)?

- **Dataset Format**:

- The dataset should be structured in a tabular format (like CSV or Excel), where each row represents a unique customer and each column represents a different attribute or behavior.

- **Data Types**:

  - **Categorical**: These columns will represent categories, such as "Gender", "Payment Method", "Product Category", etc. These will need to be encoded into numerical values for machine learning.

  - **Numerical**: Features like "Age", "Purchase Amount (USD)", and "Review Rating" that represent quantities and need to be scaled for analysis.

---

4. Data Sources

The success of the project will depend on high-quality data. While public datasets can provide a good starting point, proprietary data from real business systems will give us a more accurate and useful foundation for customer segmentation.

- **Public Datasets**:

  - **Kaggle**: Offers a range of datasets related to customer behavior, including the "Customer Shopping Trends Dataset" that we can use as a baseline.

  - **UCI Machine Learning Repository**: This platform provides many datasets on market segmentation that we can analyze and adapt to our project.

- **Proprietary Data**:

  - **E-commerce Platforms**: Many online stores (Amazon, eBay, etc.) gather customer data, which can be scraped (with permission) for insights into purchasing behaviors.

  - **CRM Systems**: Businesses often collect rich customer data via CRM systems, which can provide detailed purchase histories, customer demographics, and feedback.

  - **Social Media**: Scraping data from platforms like Instagram and Twitter can offer engagement insights related to customer behavior.

---

5. Initial Data Exploration

Before diving into complex modeling, it's important to explore and understand the data we have. This involves checking for any issues such as missing values, outliers, or patterns that may require special handling.

- **Key Exploration Tasks**:

- **Missing Data**: We'll check for any missing or null values across the dataset and apply imputation strategies like mean imputation for numerical columns or mode imputation for categorical ones.

- **Outlier Detection**: We will look for data points that fall far outside the normal range (e.g., extremely high or low purchase amounts) and decide whether they should be removed or adjusted.

- **Data Distribution**: We'll examine the distribution of key features (e.g., purchase amounts, age) and determine if transformations (like scaling or normalization) are necessary to improve clustering performance.

- **Correlation Analysis**: Identifying correlations between variables will help us understand how features relate to one another, which can guide us in feature selection.

- **Visual Exploration**: Visual tools such as histograms, box plots, and scatter plots will allow us to identify patterns, trends, or issues in the data that might need to be addressed before modeling.

---

6. Preprocessing Objectives

Data preprocessing is crucial for ensuring that the machine learning models can effectively work with the data. This phase focuses on cleaning the data and transforming it into a format suitable for clustering.

- **Feature Scaling**:

  - We'll scale numerical features (such as Age, Purchase Amount) using techniques like **Min-Max scaling** or **Z-score normalization** to ensure all features have similar scales.

- **Categorical Encoding**:

  - We will apply **One-Hot Encoding** to convert categorical features (such as Gender, Location, Payment Method) into numerical representations that can be used in the clustering algorithms.

- **Feature Selection/Reduction**:

  - We may remove irrelevant or highly correlated features to avoid overfitting and reduce the dimensionality of the data. This helps improve the efficiency and accuracy of the clustering algorithms.

- **Outlier Handling**:

  - We'll remove or adjust outliers that could skew the results of our clustering analysis, using methods like **Z-score** or **Interquartile Range (IQR)**.

7. Conclusion of Phase 1

By the end of Phase 1, we will have:

- Defined the problem clearly, focusing on customer journey analysis and enhancement of user experience.

- Identified the necessary features and requirements for the dataset.

- Explored and understood the data through initial exploration and preprocessing tasks.

This sets a strong foundation for the next phases of the project, where we will apply clustering algorithms and dimensionality reduction techniques to segment the customers and ultimately improve their shopping experience.