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

Phase 2: Data Preprocessing and Model Design

College Name: Sapthagiri College Of Engineering

Group Members:

- **Name: Karthik L**
  CAN ID Number:CAN_33260330
- **Name: Dhanush J R**
  CAN ID Number: CAN_33274239
- **Name: Rakesh R**
  CAN ID Number: CAN_3326030

2.1 Overview of Data Preprocessing

After finishing the initial data exploration in Phase 1, Phase 2 concentrates on preparing the dataset for deep learning. This preparation involves cleaning, transforming, and scaling the data to ensure it is suitable for training the deep autoencoder model. The primary objectives of this phase are to address missing values, outliers, and data inconsistencies. Additionally, it includes applying the appropriate transformations, such as feature scaling, encoding, and dimensionality reduction. Understanding customer journeys involves preparing datasets to uncover patterns and behavioral insights. This process includes cleaning, transforming, and scaling data to ensure its suitability for advanced techniques like clustering and dimensionality reduction. The aim is to enhance the user experience by identifying actionable insights.

2.2 Data Cleaning: Handling Missing Values, Outliers, and Inconsistencies

Cleaning the dataset is a vital step to ensure that the input data is accurate and ready for modelling. In this phase, we address the following issues:

- **Missing Values:** Handling missing values is critical to ensure the accuracy of the customer journey analysis. For numerical data, missing values were imputed using the mean for normally distributed features and the median for skewed data. For categorical data, missing values were imputed using the mode (most frequent category) to maintain consistency in categorical distributions. This approach helped ensure the dataset remained representative and accurate for analysis.
- **Outliers:** Outliers can distort clustering and dimensionality reduction results. We detected outliers using boxplots and Z-scores. Extreme values beyond a Z-score of 3 or -3 were flagged

as outliers. To handle these, we used capping (limiting extreme values to a threshold) for valid outliers and removal for extreme values that skewed the data. This ensured that the model wasn't influenced by abnormal data points.

- **Inconsistencies:** Inconsistent data, such as duplicates or contradictory entries, can undermine the analysis. We identified and removed **duplicate records** to avoid redundancy and flagged **contradictory data** (e.g., mismatched customer details) for review. If inconsistencies couldn't be corrected, the records were removed. This process ensured that the dataset was coherent and reliable for clustering and dimensionality reduction.

2.3 Feature Scaling and Normalization

Scaling ensures feature comparability:

- **Standardization:** Applied to numerical features using Z-score normalization.
- **Normalization:** Skewed features are scaled to a range [0, 1] using Min-Max scaling.
- **Categorical Features:** Encoded using One-Hot Encoding to create binary representations.

2.4 Feature Transformation and Dimensionality Reduction

Dimensionality reduction improves model performance and efficiency:

- **Encoding:** One-Hot Encoding ensures categorical variables are represented accurately.
- **Principal Component Analysis (PCA):** Reduces dimensions while retaining key variance. Principal Component Analysis (PCA) is applied to reduce data dimensionality while capturing maximum variance. This helps visualize and cluster data effectively.
- **Feature Selection:** Identifies and removes redundant or low-variance features.
- The code employs Standard Scaler to normalize the features, ensuring each feature contributes equally to the clustering process.

2.5 Clustering Techniques

Clustering groups customers with similar behavior using K-Means Clustering:

- The code iterates through a range of cluster numbers (k) using the K-Means algorithm. The inertia (within-cluster sum of squared distances) is computed for each iteration to determine the optimal number of clusters.
- Based on the inertia analysis (often visually inspected using an elbow plot), a final K-Means model is created with the chosen k value (e.g., 3 in this case).
- The K-Means model assigns cluster labels to each data point, segmenting visitors into distinct customer journey groups.

2.6 Enhancing User Experience

Insights from clustering and dimensionality reduction can:

- Personalize recommendations based on customer segments.
- Improve user interfaces by identifying common pain points.
- Optimize marketing strategies for targeted outreach.

2.7 Conclusion

By leveraging clustering and dimensionality reduction, businesses can analyze customer journeys effectively. These techniques provide actionable insights to enhance user satisfaction and drive growth.