Advanced Market Segmentation Using Deep Clustering

Phase 2: Data Preprocessing and Model Design

2.1 Overview of Data Preprocessing

Focusing on Phase 1, this phase gets at the process of converting the obtained customer data into a fine set of data for deep clustering. It involves dealing with the inconsistencies, standardizing the future values and simplification which comes with the complication of brought high dimensionality data. The focus is on guaranteeing that spanning is done to satisfy the architectural demand of the clustering model. Some of the important concerns include data representation preservation, noise control and maximizing the extent to which the extracted hidden features are comprehensive during the model building process.

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

The first and most important step is to perform a data cleaning process on all the data collected as is depicted by Ng and Wong (2012). The following advanced strategies were implemented:

Missing Data Treatment:

- The feature of endogenously exposed missing numerical data utilized the more-complicated imputation approaches, such as K-Nearest Neighbours (KNN), which maintains the point's local contour around the data set.
- Handling of categorical data involved the use of probabilistic models that is used to predict missing values for the benefit of data trends.

Outlier Detection and Management:

- To improve the performance and to get more reliable and stable results, basic statistical procedures as for instance the IQR with dynamic thresholds, to identify context-specific outliers, were employed.
- To investigate the multi-dimensional outliers impacting multiple associated features, Machine learning models were used and one of them named Isolation Forest was used.

Resolving Inconsistencies:

- To make them consistent, particularly in the case when some records were split or joined together in the dataset, record-linkage methods were employed.
- Local validation checks, as well as simple validation and verification techniques for time sequential data sets were performed to normalize the data set for resulting incoherencies including chronological mismatch of timestamps.

2.3 Feature Scaling and Normalization

Appropriate scaling and normalization techniques were employed to align feature distributions and enhance model convergence:

Advanced Scaling Techniques:

- Logarithmic transformation and excision of outliers are two of the main forms of data normalization that were used throughout this study, and quantile-based scaling was specifically used to bring all features to the same scale since certain features of the dataset were skewed.
- To reduce skewness and kernel transform the data, Some power transformation methods like Yeo-Johnson were used.

Encoding Innovations:

- According to the features that could be in the different categories, multi-label encoding was applied, for example: the same customers also participate in the several programmes.
- The concept of target encoding was examined where categorical predictors were recoded with corresponding effects on some critical business measures.

2.4 Feature Transformation and Dimensionality Reduction

Efficient feature engineering and dimensionality reduction were vital for optimizing computational resources and enhancing model performance:

Hierarchical Encoding:

- Some independent variables, like customer region that can be broken down into city and locality levels, were coded hierarchically to preserve geographical extent.

Feature Interactions:

- Non-linear interactions between the given variables, including combined effects of income level and purchase rate, were introduced by means of polynomial feature creation.

Dimensionality Reduction:

- t-SNE was used in this work for data visualization and also as a feature preprocessing step that aimed at revealing the underlying clustering structure of high-dimensional data.
- In the implementation of feature reduction, autoencoder-based reduction was performed by an auxiliary autoencoder with weights that reduces features, creating the best input data set for a primary clustering model.

2.5 Advanced Autoencoder Model Design

The autoencoder architecture was customized to align with project goals and handle dataset complexity:

Custom Encoder:

- The encoder network had convolution layers in order to incorporate spatial structures in sequential information like transaction histories.
- To eliminate the problem with gradient loss in deep layers, skip connections were incorporated.

Innovative Decoder:

- The decoder was designed to output multiple types of data simultaneously; numerical, categorical and sequences all in parallel.
- Additional side outputs were proposed to give feature-wise reconstruction errors improving model validation and interpretability.

Optimization Strategy:

- To increase the probability of identifying the efficient local optima, the training process contained cyclic learning rates.
- To reduce bias in latent representations, which are used in the model for the ETM, custom loss functions that included Mean Squared Error (MSE) along with the regularization terms were incorporated.

2.6 Model Training and Validation

An iterative training and validation approach ensured model performance and prevented overfitting:

Advanced Validation Techniques:

- Due to the stratified sampling, the validation set's features represented the dataset's features.
- In the current work, K-Fold cross-validation is used for the assessment of the model to have a deeper insight of the model.

Dynamic Monitoring:

- Such measures as reconstruction precision and clustering density were also observed in real time using scoreboards.
- Reconstruction loss based early stopping was used to prevent the network from continuing iterations that are unnecessary.

2.7 Conclusion of Phase 2

In phase 2, we laid down a flowchart that sequentially moved from raw to deeper clustered dataset for deep clustering. The rely mechanical cleaning approaches as well as novel preprocessing methods to provide high quality inputs to the autoencoder. Thus, using the techniques of advanced dimensionality reduction and carefully developed autoencoder architecture, the phase provides a solid basis for accurate and sensible segmentation. Subsequent phases will consider clustering algorithms specific to latent features as well as segment quality assessment and business perspectives on decision making.