FORMATIVE 2: GROUP 11 PROJECT SUMMARY REPORT

STEPS TAKEN IN PREPROCESSING

1. **Data Loading & Cleaning:** Loaded the dataset, identified 10 null values in customer_rating, and applied mean imputation.
2. **Synthetic Data Generation:** Created additional transaction data with small variations.
3. **Class Imbalance Handling:** Applied SMOTE to balance product category distribution.
4. **Data Transformation:** Performed log transformation for normalization.
5. **Synthetic Transaction Generation:** Modeled new transactions based on existing customer behaviors.
6. **Dataset Merging:** Mapped customer_id_legacy and customer_id_new using a key dataset.
7. **Feature Engineering:** Created a **Customer Engagement Score** and extracted predictive behavioral features using moving averages, time-based aggregation, and TF-IDF on reviews.
8. **Deduplication:** Removed duplicate entries in the customer_social_profiles.csv dataset.
9. **Feature Selection:** We used a correlation heatmap to identify the top 10 features for model training.

KEY INSIGHTS FOUND

1. Only customer_rating had missing values, which were fixed via mean imputation.
2. The dataset was imbalanced, requiring SMOTE for balance.
3. Duplicate entries in customer_social_profiles.csv required careful handling.

CHALLENGES FACED AND SOLUTIONS TO THEM

1. **Merging datasets with different IDs:** Used a mapping file to link customer_id_legacy and customer_id_new.
2. **Handling missing values:** Mean imputation ensured data completeness.
3. **Balancing data classes:** Addressed via SMOTE to enhance model performance.
4. **Feature selection:** We used correlation analysis to refine relevant features.