Retail Buyer Segmentation — Project Documentation

Project Summary

This project discovers and profiles natural customer groups in a retail dataset, then trains predictive models to assign new customers to those groups. The goal is to create actionable segments for targeted marketing (e.g., campaign offers, channel preferences), reduce marketing waste, and support business rules for personalization.

Objectives

1. Conduct exploratory data analysis (EDA) and preprocess the dataset, including cleaning, missing-value imputation, and applying appropriate transformations to prepare the data for modeling.
2. Utilize clustering methodologies (at least one clustering model) to uncover latent customer segments, followed by a comprehensive interpretation and documentation of the defining demographic and behavioral attributes associated with each segment.
3. Train classification models (at least one classification model) to predict cluster membership for new customers.
4. Build reproducible code, support your work with clear visualizations, and a final report.

Data overview

Download the data from here

- `customer_id` — unique customer identifier
- `birth_year` — year of birth
- `education_level` — categorical (e.g., Basic, HighSchool, Graduate, Postgraduate, Unknown)
- `marital_status` — categorical (Single, Married, Divorced, Widowed, Unknown)
- `annual_income` — numeric (household yearly income)
- `num_children` — integer (number of children at home)
- `num_teenagers` — integer (number of teenagers at home)
- `signup_date` — date when customer first engaged with company
- `days_since_last_purchase` — numeric (number of days since customer's last purchase)
- `has_recent_complaint` — binary flag (1/0)
- `spend_wine`, `spend_fruits`, `spend_meat`, `spend_fish`, `spend_sweets`, `spend_gold` — spending amounts for the last 24 months
- `num_discount_purchases` — integer (number of purchases made using a discount)
- `accepted_campaign_1` ... `accepted_campaign_5` — binary flags indicating acceptance of each promotional campaign
- `accepted_last_campaign` — binary flag indicating acceptance of the final promotional campaign

Channels & activity:

- `num_web_purchases` — integer
- `num_catalog_purchases` — integer
- `num_store_purchases` — integer
- `web_visits_last_month` — integer

Models Evaluation

Use the entire dataset for building clustering models, split the data into train, and test. And use the final 20% of the data as the test set (make sure you are shuffling the training set).

Calculate Silhouette Score for clustering models, and accuracy, precision, recall, f1 score (macro), and confusion matrix for classification models.

Bonus

Additional analyses or methodological enhancements that extend beyond the outlined requirements will be recognized and awarded accordingly.