Customer Segmentation And Predicting Behavior

1. Business Situation

CRISA is an Asian market research agency that specializes in tracking consumer purchase behavior in consumer goods (both durable and nondurable). In one major research project, CRISA tracks numerous consumer product categories (e.g., "detergents"), and, within each category, perhaps dozens of brands. To track purchase behavior, CRISA constituted household panels in over 100 cities and towns in India, covering most of the Indian urban market. The households were carefully selected using stratified sampling to ensure a representative sample; a subset of 600 records is analyzed here. The strata were defined on the basis of socioeconomic status and the market (a collection of cities).

2. Key Problems and Objective

CRISA has traditionally segmented markets on the basis of purchaser demographics. They would now like to segment the market based on two key sets of variables more directly related to the purchase process and to brand loyalty:

1- Purchase behavior (volume, frequency, susceptibility to discounts, and brand loyalty)

2- Basis of purchase (price, selling proposition)

Doing so would allow CRISA to gain information about what demographic attributes are associated with different purchase behaviors and degrees of brand loyalty, and thus deploy promotion budgets more effectively.

3. Dataset Overview

CRISA has both transaction data (each row is a transaction) and household data (each row is a household), and for the household data it maintains the following information:

- Demographics of the households (updated annually)
- Possession of durable goods (car, washing machine, etc., updated annually).
- an "affluence index" is computed from this information)
- Purchase data of product categories and brands (updated monthly)

4. Tools and Techniques Used

- Languages: Python
- Libraries: pandas, NumPy, scikit-learn, XGBoost, matplotlib, seaborn, joblib
- Environment: Jupyter Notebook
- Deployment: Flask API for real-time customer input and prediction

5. Data Preprocessing

- Categorical variables encoded using one-hot and label encoding
- Numerical features standardized using StandardScaler
- Train-test split performed with stratification to preserve segment distribution

6. Exploratory Data Analysis

- Cluster profiling showed distinct differences in Promo Usage, Purchase Volume, and Loyalty
- Visualizations helped validate cluster labeling

7. Model Building

- K-means clustering to identify the ideal number of clusters and assign label to each cluster
- Interpret segments based on brand loyalty, and purchase behavior
- Clustering evaluation using silhouette scores and visualization (2D plots using PCA)
- Supervised classification based on only demographic inputs to classify new or households, using Logistic Regression, Random Forest, and XGBoost
- Classification evaluation using accuracy, precision, recall, and confusion matrix
- Feature importance used for explainability

8. Evaluation Metrics

Best model: XGBoost Classifier

Accuracy: 0.57

Precision by Segment:

- Loyalists: 0.49
- Variety Seekers: 0.33
- Promo Shoppers: 0.64

9. Key Takeaways

- XGBoost outperformed other models in overall accuracy and segment-specific precision
- Feature importance revealed that Affluence Index, Food Eating Habits and number of children were key drivers
- Flask app successfully deployed for real-time labeling of new households

10. Resources

GitHub Repo

Kaggle Notebook

Flask App Demo