User Segmentation Based on Purchasing Behavior

By: Ooi Li Yoong

https://www.linkedin.com/in/liyoongooi

https://www.github.com/liyoongooi

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

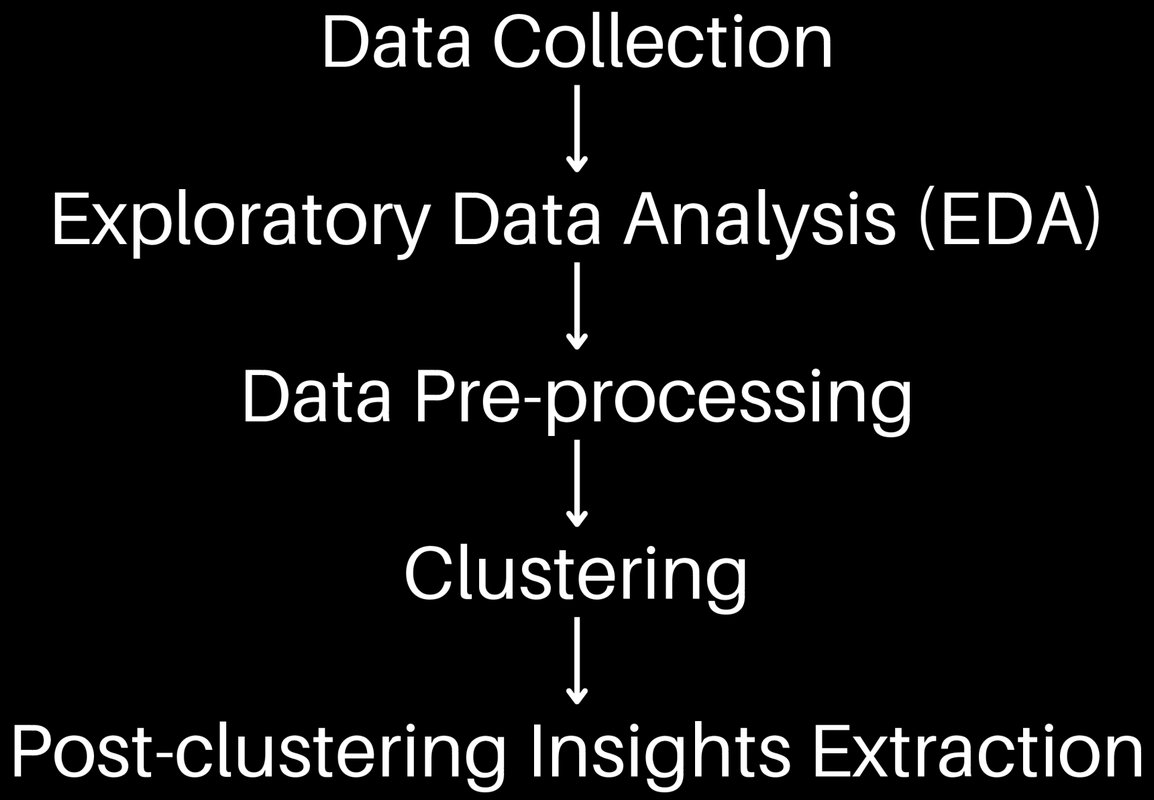
Introduction - Background

- Each customer has different needs
- Given the large number of customers, it is impractical to target each of them
- User segmentation effectively divide customers into several groups based on their similarities between each other
- Tailored business strategies for each group can be formulated

Introduction - Objective

- To identify segments of customers

Methodology - Workflow



Methodology - Data Collection

- Data is obtained from Kaggle
  - https://www.kaggle.com/datasets/refiaozturk/online-shopping-dataset?resource=download

Methodology - EDA

- The dataset contains 7 columns
  - Categorical: gender, country, purchase category
  - Numerical: user ID, age, purchase amount
  - Timestamp: purchase date
- Completeness: remove rows with missing value(s)
- Uniqueness: remove duplicated rows
- Validity: check the ranges of values

Methodology - EDA

- Univariate analysis
  - Bar chart
  - Histogram
- Bivariate analysis
- Multivariate analysis

Methodology - Data Pre-processing

- Encoding
  - One-hot encoding
  - One-hot encoding + Autoencoding
- Scaling
  - Min-max scaling
  - Standard scaling

Methodology - Clustering

- Clustering Algorithms
  - k-Means
    - Elbow method
  - DBSCAN
    - Grid search

Methodology - Clustering

k-Means

Model	Encoding	Scaling	Clustering Algorithm
K1	OHE	None	k-Means
K2	OHE + Autoencoding	None	k-Means
K3	OHE	Min-max	k-Means
K4	OHE + Autoencoding	Min-max	k-Means
K5	OHE	Standard	k-Means
K6	OHE + Autoencoding	Standard	k-Means

DBSCAN

Model	Encoding	Scaling	Clustering Algorithm
D1	OHE	None	DBSCAN
D2	OHE + Autoencoding	None	DBSCAN
D3	OHE	Min-max	DBSCAN
D4	OHE + Autoencoding	Min-max	DBSCAN
D5	OHE	Standard	DBSCAN
D6	OHE + Autoencoding	Standard	DBSCAN

Methodology - Clustering

- Clustering Models Evaluation
  - Silhouette
    - A value close to 1 is desired
  - Calinski-Harabasz
    - A large value is desired

Methodology - Post-Clustering Insights Generation

- Clustering to classification
  - Train-test split
    - 70% training data, 30% testing data
  - Classification algorithm
    - Logistic Regression, Random Forest, Light Gradient-Boosting Machine
  - Evaluation
    - Accuracy, Precision, Recall, F1-score

Methodology - Post-Clustering Insights Generation

- SHapley Additive exPlaination (SHAP)
  - shap.LinearExplainer
  - shap.TreeExplainer
- Identify important features from the SHAP summary plot
- Analyze central tendencies of the important features across different clusters
- Develop descriptive personas

Results - EDA

- Each customer performed exactly one purchase
- All numerical columns contain no outlier

Results - EDA

- Numerical columns have varying scales
- Values of purchase amount differ minimally across different categories

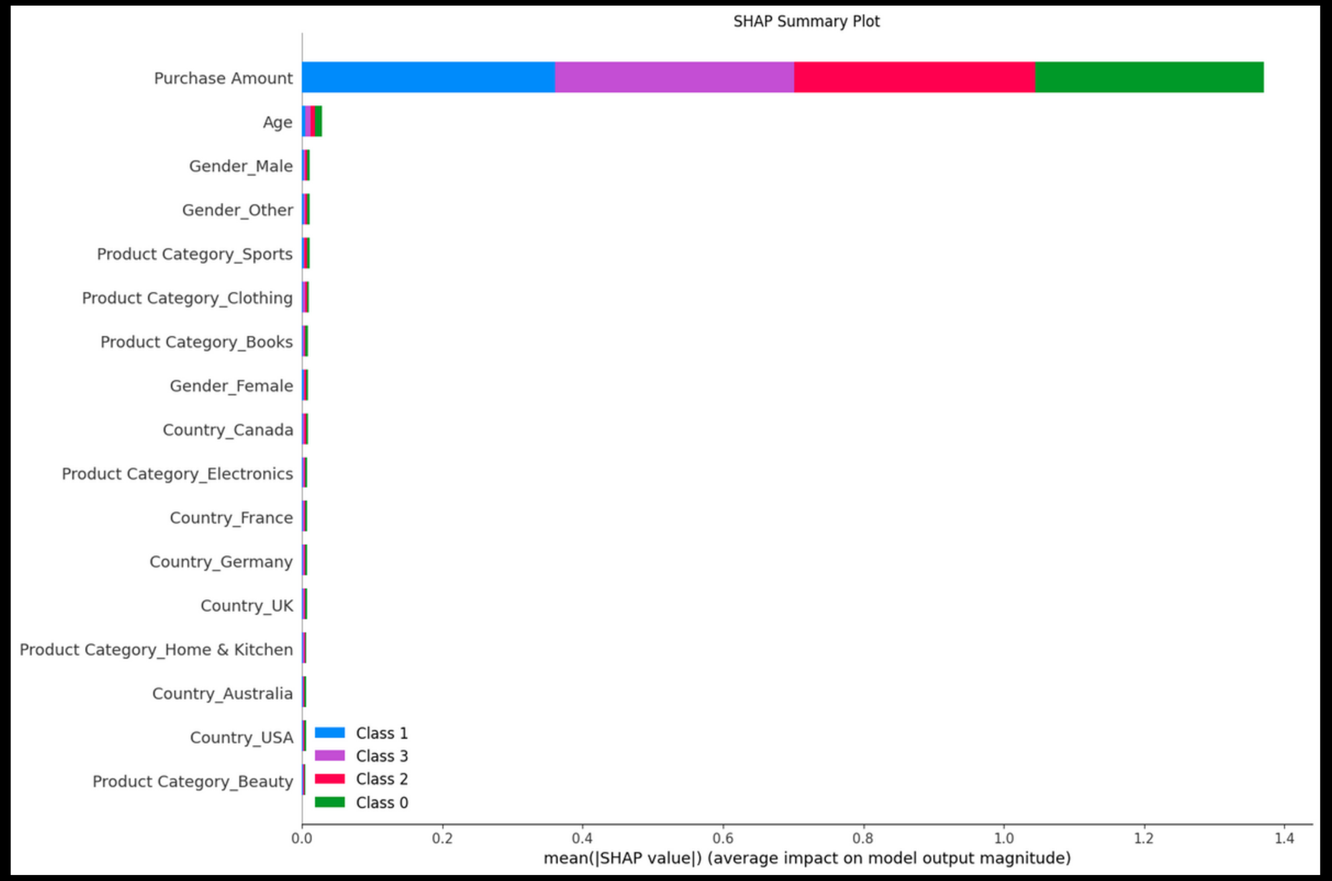
Results - Clustering

- One-hot encoding is sufficient
- Scaling is not required
- k-Means is optimal
- Optimal clustering model: K1
  - One-hot encoding
  - k-Means

Results - Classification

- Optimal classifier: Random Forest

Results - SHAP



Results - Personas

Cluster 0: Moderate spender

Cluster 1: Top spender

Cluster 2: Minimum spender (Below 50 years old)

Cluster 3: Above-average spender

Conclusion

- Optimal clustering model: k-Means with OHE
- Optimal classifier: Random Forest
- Future work
  - Obtain more related data. Try to collect customers' multiple purchasing records
  - Include business experts to formulate tailored marketing strategies

Thank You

 https://www.linkedin.com/in/liyoongooi

 https://www.github.com/liyoongooi