# ML MINI PROJECT SUMMARY AND IMPLEMENTATION

**Submitted by:**

* **Names:** Aarit Mehta & Akshat Panchal
* **Roll No:** 16014223002 & 16014223008
* **Division:** TY AIDS A1
* **GitHub Repository:**<https://github.com/aaritmehta15/amazon-customer-prediction>

This document provides a summary of the research paper "Exploration and Analysis of Amazon Customer Behavior" and explains how we implemented its findings in our code and then extended the project with new, valuable features.

### 1. Summary of the Research Paper

The research paper we based our initial work on, "Exploration and Analysis of Amazon Customer Behavior," focuses on understanding and segmenting Amazon customers to improve business performance.

Here’s a breakdown of what the paper does:

* **Objective**: To analyze customer data, identify key factors influencing customer satisfaction, and group similar customers together.
* **Methodology**:
  + It uses a public dataset of Amazon customer behavior from Kaggle.
  + It cleans the data by handling missing values and converting text-based categories into numbers.
  + It applies **Principal Component Analysis (PCA)** to reduce the number of variables, making the data easier to work with without losing important information.
  + It uses two main machine learning algorithms, **K-Means Clustering** and **Hierarchical Clustering**, to identify distinct customer segments.
  + The **Silhouette Score** and **Elbow Method** are used to determine the ideal number of customer groups (clusters).
* **Conclusion**: The paper successfully identifies several customer segments and concludes that these insights can be used to make strategic business decisions, such as creating targeted marketing campaigns. The final step mentioned is visualizing these insights using Business Intelligence (BI) tools.

### 2. How We Implemented the Paper's Findings

In our Jupyter notebook, clustering-amazon-customers-pca-k-means.ipynb, we built a direct and practical implementation of the core technical analysis described in the research paper. We not only replicated the methodology but also quantified the results to produce a clear, data-driven segmentation model.

Here is what we did in our code to align with the paper:

* **Data Processing and Dimensionality Reduction**: We implemented a full preprocessing pipeline, including handling missing values, scaling features, and encoding categorical variables. We then correctly applied PCA to reduce the dimensionality of the dataset, which was a key step in the paper's methodology.
* **Clustering, Evaluation, and Segmentation**: We implemented both K-Means and Hierarchical clustering algorithms. Through the Elbow Method and Silhouette analysis, we determined the optimal number of clusters to be **two**. Our model produced the following key performance metrics:
  + **Silhouette Score**: 0.094
  + **Davies-Bouldin Index**: 2.91
  + **Calinski-Harabasz Index**: 69.49
* **Identified Cluster Profiles**: Our analysis successfully segmented the customers into two distinct, nearly equal-sized groups:
  + **Cluster 0 (50.3% of customers)**: Identified as "Beauty-Focused Shoppers," with a strong preference for Beauty and Personal Care products.
  + **Cluster 1 (49.7% of customers)**: Identified as "Diverse Category Shoppers," with more varied purchasing patterns across multiple categories.
  + We further validated these segments by performing statistical tests, which confirmed significant differences (p < 0.05) in shopping satisfaction and the importance placed on customer reviews between the two groups.
* **Visualization of Clusters**: To make the results clear and easy to understand, we created scatter plots and other visualizations to effectively display the identified customer segments, providing a clear visual representation of the findings.

### 3. How We Extended the Analysis (Beyond the Paper)

After replicating the paper's analysis, we took the project a step further in our second notebook, advanced\_customer\_segment\_analysis.ipynb. While the research paper focuses on identifying and describing customer segments, we extended the analysis to create a tool that generates a broader range of actionable business insights directly from the clustered data.

Here are the key functionalities we developed that go beyond the original paper:

1. **Marketing Strategy Suggestions**:
   * We built on the paper's idea by creating the generate\_marketing\_strategy function. This provides targeted marketing ideas and slogans for each customer segment based on their characteristics, offering a starting point for campaigns.
2. **Customer Churn Risk Scoring**:
   * To provide a more forward-looking view, we developed the predict\_churn\_risk function. This function creates a churn risk score for customers based on their purchasing behavior, helping to identify at-risk customers who might need proactive engagement.
3. **Product Recommendation Insights**:
   * We created the analyze\_product\_recommendations function, which identifies the top product categories purchased by each customer segment. This analysis provides the data needed to power personalized product recommendation engines.
4. **Customer Lifetime Value (CLV) Estimation**:
   * We developed the estimate\_clv function to calculate the estimated lifetime value for each customer cluster. This is a critical business metric that helps in prioritizing marketing spend and strategic efforts on the most valuable customer groups.
5. **Estimating Customer Service Demand**:
   * We also added the forecast\_service\_demand function to help with operational planning. This function analyzes the average number of customer service interactions for each segment, providing a valuable estimate of which groups require more support.

By adding these features, we have evolved the academic analysis from the paper into a powerful, practical tool that can deliver direct business value in marketing, customer retention, and operations.