Customer Segmentation Model

Utkarsh Tripathi (2025EM1100146)
Juwaria Qadri (2025EM1100132)
Mohammed Omar (2025EM1100235)
Merin Ann Cherian (2025EM1100211)

Phase 1

Segmenting Blinkit customers based on their spending behaviour and delivery experience to identify distinct customer clusters (high, medium, and low spending) for targeted marketing strategies and increasing sales

Proposal

1. **Project Statement**
   Across customer-centric businesses, one of the major goals for the organization is being able to predict the needs of its customers and how to best cater to their needs while maintaining high profitability. However, to make the solution feasible, the way forward would be to create groups of consumers that share similar spending patterns, allowing the businesses to create targeted marketing campaigns, improve customer satisfaction, demand-oriented product development, and therefore improve *profitability*.

2. **Business Goal**
   The objective of this project is to develop a machine learning model that can create suitable and usable clusters/segments given the profile and purchases of the customer. The model should then be able to predict, within bounds of acceptable error, which segment a customer is likely to belong to. The model should allow stakeholders to make informed and data-backed decisions on product demand, campaign successes and customer retention.

3. **Data Source**
   We will use the "Blinkit Sales Dataset" for this project. This dataset provides detailed information on product sales, visibility, item types, and outlet performance, making it ideal for performing sales data analysis and gaining insights into business trends. The

dataset is well-structured and suitable for data preprocessing, exploratory data analysis, and predictive modeling tasks.

- **Source Platform: Kaggle**
- **Dataset**: Blinkit Sales Data (Vaghasiya 2025)

4. **Tools and Technology**
   Following languages and libraries are planned to be used for this project. More libraries may be used and every non trivial library shall be updated.

   - Python
     - Core Libraries
       * Data Manipulation: Pandas, NumPy
       * ML: scikit-learn
       * Data Visualization & Storytelling: Matplotlib, Seaborn
     - Development Environment: VS Code, Google Colab, GitHub
   - Dashboard: Power BI

5. **Project Workflow**
   The project is to follow a standard data science development lifecycle,

   1. Data Acquisition: Fetch the dataset from **Kaggle** using its API.
   2. Preprocessing: Handle missing values (if any), encode categorical variables, and check for data inconsistencies.
   3. EDA: Analyze features to understand their relationship with attrition using statistical summaries and visualizations.
   4. Feature Engineering: Create new features from existing ones if necessary to improve model performance.
   5. Modeling: Train several clustering models (e.g., K Means Clustering, PCA, Decision Trees).
   6. Evaluation: Assess model performance using metrics like Accuracy, Precision, Recall, and F1-Score. Select the best-performing model.

7. Visualization: Create an interactive dashboard in Power BI to present the key findings and predictions to stakeholders.

6. **Data Extraction**
   The "*Blinkit Sales Dataset*" is acquired directly from the Kaggle repository. To ensure a professional and reproducible workflow, manually downloading the files is not done. Instead, we will perform the following steps:

   - Automate the Process: We will write a Python script that utilizes the official Kaggle API to connect to the source and download the dataset.
   - Ensure Reproducibility: This scripted approach guarantees that the data extraction process is consistent and can be easily re-run by any team member or reviewer.
   - Prepare for Analysis: The script will handle the unzipping of the downloaded files and load the data directly into a Pandas DataFrame, making it immediately available for the next phase of our project.
   - Notebook: Customer Segmentation Notebook

7. **Schema / Data Dictionary**
   This data dictionary was created after inspecting the dataset.

	Feature Name	Data Type	Description	PK (Yes/No)
0	order_id	int64	Unique identifier for each order	Yes
1	customer_id	int64	Unique identifier for each customer	No
2	order_date	object	The timestamp when the order was placed	No
3	promised_delivery_time	object	The time informed to the customer for completi...	No
4	actual_delivery_time	object	The actual time when the order was delivered	No
5	delivery_status	object	The delivery status of the order	No
6	order_total	float64	The total price of the order (in Rupees)	No
7	payment_method	object	The payment method used by the customer	No
8	delivery_partner_id	int64	Unique identifier representing the delivery pa...	No
9	store_id	int64	Unique identifier for the store that fulfilled...	No

Source: Data Dictionary

Phase 2

Data Extraction

Vaghasiya, Akshit. 2025. “Blinkit Sales Dataset.” https://www.kaggle.com/datasets/akxiit/blinkit-sales-dataset.