**Project Title:** A Customer segmentation model for ecommerce company to drive marketing and upselling strategies.

**1. Problem Statement:**

The e-commerce retailer aims to utilize customer purchase history to segment its customer base into actionable groups. The goal is to drive targeted marketing and upselling strategies that enhance customer engagement, retention, and lifetime value. The retailer seeks analytical models and processes that provide meaningful segments, prescribe optimal marketing actions for each, and allow ongoing measurement of business impact.

**2. Business Goal:**

The primary objective is to enhance customer engagement and maximize the lifetime value of each customer by:

* **Understanding customer behavior** through segmentation based on purchase history.
* **Designing personalized marketing campaigns** that cater to the specific needs and preferences of each segment.
* **Identifying upselling and cross-selling opportunities** by aligning product recommendations with segment-specific tendencies.

**3. Data Source**

The dataset named Online Retail II contains the sales of a UK-based online retail store between 01/12/2009 - 09/12/2011.

* Source Platform: Kaggle
* Full Citation:

An e-commerce company wants to segment its customers and determine marketing strategies according to these segments.

<https://www.kaggle.com/datasets/sanlian/online-retail-dataset>

**4. Tools & Technologies**

* Programming Language: Python
* Core Libraries:
  + Data Manipulation & Analysis: Pandas, NumPy
  + Machine Learning: Scikit-learn
  + Data Visualization & Storytelling: Matplotlib, Seaborn
* Development Environment: Our team will primarily use Jupyter Notebook
* BI Tools: Tableau or Power BI to create a final dashboard.

**5. Project Workflow**

The project will follow a structured data science lifecycle, visualized as follows:

Data Acquisition → Data Cleaning & Preprocessing → Exploratory Data Analysis (EDA) →

Feature Engineering → Model Building & Training → Model Evaluation → Reporting &

Visualization

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 classification models (e.g., Logistic Regression, Random

Forest, Gradient Boosting).

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 Tableau/PBI to present the key

findings and predictions to stakeholders.

**6. Data Extraction**

" Online Retail II" 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 extractionprocess 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 filesand load the data directly into a Pandas DataFrame, making it immediately available for the next phase of our project.

o Notebook: data\_extraction.ipynb

**7. Schema/Data Dictionary:**

This data dictionary is created after inspecting the dataset.

Excel sheet: Data\_Dictionary\_Cust\_Segmentation.xls