Customer Purchase Behavior Analysis & Prediction for Amazon

Problem Statement:

Amazon, a global leader in e-commerce, wants to optimize its **customer segmentation, revenue forecasting, and churn prediction** to enhance customer retention and increase revenue. With millions of customers and transactions daily, Amazon collects **demographic details, purchase history, and transaction data** but faces the following challenges:

- ✅ Identifying **high-value customers** for targeted marketing.
- ✅ Predicting **Customer Lifetime Value (CLV)** to improve revenue forecasting.
- ✅ Understanding **customer churn risks** and improving retention strategies.
- ✅ Grouping customers into **actionable segments** based on behavior patterns.

The goal of this project is to **develop Machine Learning models** to **segment customers, predict their future spending, and classify them as potential churners or active customers**. However, before building ML models, we need to **clean and preprocess the data** to ensure accuracy.

📌 Step 1: Data Cleansing & Preprocessing

Before applying ML models, it is crucial to ensure **data quality** by performing the following steps:

✅ **Handling Missing Values**

- Identify missing values in **Age, Purchase Amount, Rating, and Customer Lifetime Value (CLV)**.
- Apply **mean/median imputation** for numerical fields.
- Apply **mode imputation** for categorical fields like Payment Method.

✅ **Removing Duplicates**

- Remove duplicate entries based on **Customer_ID and Purchase_Date**.

✅ **Data Formatting & Type Correction**

- Convert **Purchase_Date** to datetime format.
- Standardize categorical values (e.g., **Gender: Male, Female, Other**).
- Ensure **consistent data types** (integers for numeric fields, categorical encoding for non-numeric).

✅ **Handling Outliers**

- Identify outliers in **Purchase Amount & CLV** using **boxplots & z-score analysis**.
- Apply **winsorization** or remove extreme outliers.

✅ **Feature Engineering (Adding New Columns)**

To make the dataset more useful for machine learning, we add the following new columns:

1. **Customer_Lifetime_Value (CLV):** Projected future revenue per customer.
2. **Loyalty Score:** Score based on purchase frequency and total spending.
3. **Discount Applied:** Whether the purchase was made with a discount (Yes/No).
4. **Return Status:** Indicates if the item was returned (Yes/No).
5. **Customer Segment:** Categorized as **New, Regular, VIP** based on loyalty.
6. **Preferred Shopping Channel:** Where the customer shops (Online, In-store, Both).

📌 Step 2: Machine Learning Tasks

After data cleaning and feature engineering, we apply **Machine Learning models** to derive insights.

1️⃣ Customer Segmentation (Clustering - K-Means)

📌 **Objective:**

- Categorize Amazon customers into **distinct groups** based on spending patterns, purchase frequency, and loyalty scores.
- Identify **high-value, occasional, and low-value customers** for targeted promotions.

📌 **Method:**

- Use **K-Means Clustering** to segment customers into groups based on:
  - **Total purchase amount**
  - **Number of orders**
  - **Loyalty score**

📌 **Industry Application:**

- Helps Amazon **personalize recommendations and promotions** for different customer segments.
- Enables **dynamic pricing strategies** based on customer type.

2️⃣ Predicting Customer Lifetime Value (Regression - Linear Regression)

📌 **Objective:**

- Estimate the **future revenue** Amazon can generate from each customer.
- Identify **high-CLV customers** and offer exclusive deals to increase retention.

📌 **Method:**

- Train a **Linear Regression model** to predict **CLV** based on:
  - **Age, past purchases, discount usage, payment method, and loyalty score**.

📌 **Industry Application:**

- Helps Amazon in **predictive marketing and resource allocation**.
- Enables **cost-efficient retention strategies**.

🚀 **Expected Deliverables**

✓ **Cleaned dataset with new features (CLV, Loyalty Score, etc.).**
✓ **K-Means Clustering for customer segmentation.**
✓ **Linear Regression model for CLV prediction.**
✓ **Logistic Regression model for churn prediction.**
✓ **Power BI dashboard for visualizing insights.**
✓ **Jupyter Notebook with all models & findings.**