CUSTOMER LIFETIME VALUE PREDICTION MODEL

INTRODUCTION

In today's competitive market, knowing a customer's long-term value is vital. Customer Lifetime Value (CLTV) estimates the total revenue a customer can bring over time. Predicting CLTV helps businesses personalize marketing, improve retention, and focus on high-value customers.

ABSTRACT

This project aims to predict the Customer Lifetime Value (CLTV) using historical purchase behavior data. The model leverages customer transaction data to compute key behavioral metrics—Recency, Frequency, Monetary (RFM)—and uses machine learning to predict the future value of each customer. Additionally, customers are segmented into Low, Mid, and High Value groups for actionable insights. The implementation uses Python libraries such as Pandas, XGBoost, and Seaborn, with data sourced from a retail invoice dataset.

TOOLS USED

- **Programming Language:** Python
- **Libraries & Frameworks:**
  - Pandas, NumPy – Data manipulation
  - Matplotlib, Seaborn – Visualization
  - XGBoost, Sklearn – Machine Learning and Evaluation
- **Data Source:** Excel file containing invoice-level transaction data
- **IDE:** Jupyter Notebook or VS Code
- **Output File:** CLTV_Customer_Segments.csv

STEPS INVOLVED IN BUILDING THE PROJECT

Step 1: Data Preprocessing

- Loaded the Excel data file containing customer invoice details.
- Calculated TotalAmount for each invoice as Quantity × UnitPrice.
- Extracted Recency, Frequency, Monetary (RFM) features:
  - **Recency:** Days since last purchase
  - **Frequency:** Number of purchases (Invoice count)
  - **Monetary:** Total spending

Step 2: Feature Engineering

- Calculated AOV (Average Order Value) = Monetary / Frequency.
- Assigned the target variable CLTV as the total monetary value.

- Extracted categorical features:

  Most frequentProductCategory , Dominant PaymentMethod , Most used SalesChannel
- Applied one-hot encoding on categorical features to prepare for modeling.

Step 3: Model Training

- Split the dataset into training and testing sets (80/20 split).
- Used XGBoost Regressor with 100 trees to fit the training data.
- Predicted CLTV on the test set and evaluated using:
  - MAE (Mean Absolute Error): 16.29
  - RMSE (Root Mean Squared Error): 23.37

Step 4: Segmentation and Visualization

- Based on predicted CLTV, customers were divided using quantiles:
  - **Low Value:** ≤ 25th percentile
  - **Mid Value:** 25th–75th percentile
  - **High Value:** ≥ 75th percentile
- Created a new column: Segment with one of the three categories.
- Visualized customer count in each segment using a count plot:

Step 5: Output

- Saved the segmented customer dataset as CLTV_Customer_Segments.csv
- The chart showed that:
  - Most customers are in the Mid Value segment.
  - A balanced number are in Low and High Value groups.

CONCLUSION

This CLTV prediction model effectively segments customers based on their historical purchase data, enabling smarter business decisions. With a low MAE and RMSE, the XGBoost model offers strong predictive performance. Segmenting customers empowers the business to:

- Target high-value customers for loyalty rewards,
- Re-engage mid-value customers with offers,
- Understand and address the needs of low-value customers.

In future versions, additional features like customer churn probability or campaign response scores could be integrated for even richer insights.