📄 Customer Lifetime Value Prediction – Project Report

👤 **Author: Keerthana Kothoju**

📅 **Date: 18/05/2025**

🏢 **Project by: Elevate Labs**

🎯 Objective

The aim of this project is to **predict Customer Lifetime Value (CLTV)** using historical purchase behaviour. This predictive model will assist in **targeted marketing** by segmenting customers based on their predicted future value.

🧰 Tools & Technologies

- **Python** (Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn, XGBoost)
- **Jupyter Notebook**
- **Excel** (for data inspection and sanity checks)

📦 Dataset Overview

- Source: Online Retail transactional data
- Fields: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country
- Total records: Rows: 541,909 | Columns: 8
- Missing Values: Found in CustomerID and Description
- Country: Focused on United Kingdom customers only

🔍 Exploratory Data Analysis (EDA)

🧼 Data Cleaning

- Removed null CustomerID values
- Removed negative Quantity and UnitPrice
- Created TotalPrice = Quantity * UnitPrice

📊 Key Insights:

- **Monthly Revenue Trend**: Reveals business performance over time.
- **Top Products Sold**: Identified top 10 bestsellers by quantity.
- **Top Customers**: Customers contributing the highest revenue.
- **Order Value Distribution**: Most orders are of smaller value; right-skewed distribution.

🔧 Feature Engineering

We aggregated the transactional data to generate **customer-level metrics**:

- **Recency**: Days since last purchase
- **Frequency**: Number of unique invoices
- **Average Order Value (AOV)**: TotalSpend / NumInvoices
- **CLTV (Target Variable)**: Total spend by customer

All features are stored in a new DataFrame customer_df.

📈 Advanced Visual Analysis (Post-Feature Engineering)

- **CLTV Distribution**: Right-skewed, suggesting a few high-value customers.
- **Recency Boxplot**: Most customers purchase recently, but long tails exist.
- **Frequency vs CLTV**: More frequent customers often contribute higher value.
- **AOV vs CLTV**: High AOV customers tend to have high lifetime value.

🤖 Model Training

Algorithm: XGBoostRegressor

- **Features Used**: Recency, Frequency, AOV
- **Target**: CLTV

Performance Metrics:

Metric	Score
MAE	305.83
RMSE	2695.84

🧪 Validation & Segmentation

- The model was validated using standard regression metrics.
- Based on predicted CLTV, customers were segmented into:
- **High Value**: Top 20% predicted scores
- **Medium Value**: Next 30%
- **Low Value**: Bottom 50%

This segmentation enables marketing teams to personalize promotions and retain high-value customers.

📌 Conclusion

This CLTV prediction pipeline successfully demonstrates:

- Effective EDA and feature extraction
- Insights into customer behaviour
- A predictive model to estimate lifetime value
- A foundation for marketing segmentation.