Project Report: Customer Lifetime Value (CLV) Prediction

# Project Title

Predicting Customer Lifetime Value for E-Commerce Using Machine Learning

# Problem Statement

Customer Lifetime Value (CLV) is a critical metric for understanding the long-term value a customer brings to a business. This project aims to build a machine learning model that accurately predicts CLV based on historical transaction data of an e-commerce platform.

# Objectives

- Predict CLV using customer transaction behavior.

- Identify high-value customers for targeted marketing.

- Deploy the model with an interactive dashboard using Streamlit.

# Dataset

Source: Online Retail Dataset from UCI Machine Learning Repository

Description: Contains 541,909 transactions from a UK-based online retailer between 2010 and 2011.

Key Columns: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country

# Tech Stack

Language: Python

Data Handling: Pandas, NumPy

ML Algorithm: XGBoost Regressor

Visualization: Matplotlib, Seaborn

Dashboard: Streamlit

Model Saving: joblib

Deployment: Streamlit Cloud / Localhost

# Project Structure

CLV-Prediction-Project/

├── app/

│ ├── clv\_predictor.py ← Streamlit app

│ └── model.pkl ← Trained XGBoost model

├── notebooks/

│ └── clv\_modeling.ipynb ← EDA, feature engineering, modeling

├── data/

│ └── online\_retail.csv ← Raw dataset

├── requirements.txt ← Python dependencies

├── README.md ← Project overview

# Methodology

1. Data Preprocessing:

- Removed rows with missing CustomerID.

- Removed negative quantities and cancellations.

- Created features like TotalAmount, Recency, Frequency, Monetary (RFM Model).

2. Feature Engineering:

- Aggregated customer-level features.

- Created time-based metrics (average spend, recency in days).

3. Modeling:

- Applied XGBoost Regressor.

- Evaluated using MAE, RMSE, R² Score.

4. Model Export:

- Saved using joblib.

5. Deployment:

- Streamlit UI for input & predictions.

- Used os.path.join for model loading.

# Results

- R² Score: ~0.82

- RMSE: Acceptable for business-level CLV prediction

- Accurately distinguishes between low, medium, and high CLV segments.

# Insights

- Top 20% customers contribute to over 60% revenue (Pareto Principle).

- CLV positively correlates with frequency and monetary value.

- Recency has a strong negative correlation with CLV.

# Conclusion

Demonstrates how predictive analytics can drive real-world business impact. Helps:

- Prioritize marketing

- Identify churn risks

- Increase retention

# Live Demo & Source Code

- Streamlit App: \_[Deploy on Streamlit Cloud or Localhost]\_

- GitHub Repository: \_[Add your repo link here]\_