Customer Lifetime Value (CLV) Prediction Report

1. Introduction

Customer Lifetime Value (CLV) is a crucial metric that helps businesses understand the long-term value of each customer. It enables data-driven decisions for customer acquisition, retention, and marketing investment. In this project, we use customer behaviour data (e.g., time on app, session duration, membership length) to build a predictive model that estimates CLV. The objective is to identify high-value customers and support personalized marketing strategies.

2. Abstract

This project uses machine learning to predict CLV for an e-commerce business. We use regression modelling to estimate the *Yearly Amount Spent* by each customer — a practical proxy for CLV. The dataset contains features such as average session length, time on app, time on website, and membership duration. We train an **XGBoost Regressor** to learn from these features and produce accurate predictions. The final output segments customers into **High**, **Medium**, and **Low-value** categories to support business decisions.

3. Tools and Technologies Used

- **Programming Language:** Python
- **Libraries:** Pandas, NumPy, Scikit-learn, XGBoost, Seaborn, Matplotlib
- **IDE:** Jupyter Notebook
- **Others:** Microsoft Excel (for basic exploration)

4. Steps Involved in Building the Project

Step 1: Data Collection & Exploration

- Loaded the Ecommerce Customers.csv dataset.
- Conducted initial exploration: null values check, summary stats, correlations.

Step 2: Feature Selection

- Selected key features:
    - Average Session Length
    - Time on App
    - Time on Website
    - Length of Membership

- Target Variable: *Yearly Amount Spent* (used as proxy for CLV)

Step 3: Data Preprocessing

- Handled data formatting and scaling.
- Split into training and test sets (80%-20%).

Step 4: Model Building

- Used **XGBoost Regressor** due to its performance and handling of feature interactions.
- Trained the model and optimized it using basic parameter tuning.

Step 5: Model Evaluation

- **Mean Absolute Error (MAE):** 12.36
- **Root Mean Squared Error (RMSE):** 16.19
- Found 'Length of Membership' to be the most impactful feature.

Step 6: Customer Segmentation

- Based on predicted CLV, customers were grouped as:
    - **High Value:** Top 20%
    - **Medium Value:** Next 30%
    - **Low Value:** Bottom 50%
- Saved the segmented output as a CSV file.

Step 7: Visualization & Insights

- Used feature importance and correlation plots to explain model behaviour.

5. Conclusion

This project demonstrates a successful implementation of a machine learning-based approach for predicting Customer Lifetime Value. By identifying high-CLV customers, businesses can design focused retention strategies and improve marketing ROI. The model achieved a good balance of accuracy and interpretability, making it useful for real-world deployment. Future improvements can include integrating more customer attributes and deploying the model for real-time use.

Report By: *Sion Rani Kondepudi*

Date: *26-06-2025*