Customer Lifetime Value Prediction Report

1. Introduction

Customer Lifetime Value (LTV) is a key metric used in marketing and customer relationship management. It estimates the total revenue a business can reasonably expect from a customer over the entire span of their relationship. Predicting LTV allows businesses to optimize marketing strategies, personalize customer engagement, and allocate resources efficiently.

2. Abstract

This project aims to predict the Lifetime Value (LTV) of customers using historical transaction data. The dataset used originates from the UCI Machine Learning Repository's **Online Retail** dataset, which contains transactional data from a UK-based online store. The project involves cleaning the data, engineering behavioral features, training a machine learning model to predict average order value (AOV), and calculating LTV. Customers are then segmented into Low, Medium, and High value categories based on predicted LTV.

3. Tools Used

- **Python** (pandas, numpy, matplotlib, seaborn, sklearn, xgboost)
- **Jupyter Notebook**

4. Steps Involved in Building the Project

Data Cleaning

- Removed null and duplicate values
- Filtered out negative quantities and missing customer IDs
- Converted date fields and computed total purchase amounts

Data Preprocessing

- Merged transaction data to create a **customer-level dataset**.

Outlier Removal

- Removed extreme outliers based on the IQR (Interquartile Range) method for AOV, Monetary, and Avg. Order Gap to improve model stability.

Feature Engineering

- Recency: Days since last purchase
- Frequency: Number of purchases
- Monetary: Total money spent
- Avg Order Gap: Average time between purchases
- AOV: Average Order Value

Model Building with RandomForest Regressor

- Trained a Random Forest model initially.
- Got decent but not very strong results (MAE ~ 350+, RMSE ~ 1600+).
- R^2 score was low (Train: 0.36, Test: -0.19).

Model Building with XGBoost Regressor

- Used XGBoost, a more powerful model.
- After feature improvements and cleaning, achieved excellent results:

  MAE: 7.17   **RMSE**: 24.88   **Train R^2**: 1.00   **Test R^2**: 0.99
- R^2 score showed the model explains nearly 99% of the variance.

LTV Prediction & Segmentation

- LTV estimated using: Predicted AOV × Frequency
- Customers segmented as:

  Low Value (< ₹50)   **Medium Value** (₹50–₹150)   **High Value** (> ₹150)

Output Generation

- Final dataset with predicted LTV and segment saved as

  'segmented_customers_predicted_LTV.csv'

5. Conclusion

This project demonstrates a simple yet effective method for estimating Customer Lifetime Value using transactional data and machine learning. Businesses can use the resulting customer segments to personalize marketing efforts, improve retention strategies, and prioritize high-value customers. The model is scalable and can be integrated into real-time analytics systems or dashboards for ongoing insights.