Customer Lifetime Value (CLV) Prediction – Project Report

Introduction
Customer Lifetime Value (CLV) is a key business metric that estimates the total revenue a company can expect from a customer throughout their relationship. Identifying high-value customers enables businesses to optimize marketing efforts, improve retention, and increase profitability.

In this project, the Online Retail dataset was used to build a machine learning pipeline to predict CLV. The dataset includes transactional details such as invoice number, customer ID, product, quantity, price, and date. The focus was on building predictive models and segmenting customers into meaningful groups.

Abstract
This project aims to predict Customer Lifetime Value (CLV) by applying data preprocessing, feature engineering, and machine learning models on transactional data. The Recency, Frequency, and Monetary (RFM) model along with Average Order Value (AOV) was used to derive predictive features.

Regression models, including RandomForest and XGBoost, were trained to estimate CLV. Model performance was evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Finally, customers were segmented into Low, Medium, and High value groups based on predicted CLV, providing actionable insights for targeted marketing.

Tools Used
- Python: Core programming language
- Pandas, NumPy: Data manipulation and preprocessing
- Matplotlib, Seaborn: Data visualization
- Scikit-learn: Machine learning model building and evaluation
- XGBoost: Gradient boosting regression model
- Jupyter Notebook: Interactive development environment

Steps Involved in Building the Project
1. Data Loading – Imported the Online Retail dataset into Pandas.
2. Data Cleaning – Removed missing values and invalid records.
3. Feature Engineering – Created Recency, Frequency, Monetary, and AOV.
4. Model Training – Trained RandomForest and XGBoost regression models.
5. Model Evaluation – Measured performance using MAE and RMSE.
6. Feature Importance – Identified key drivers of CLV with XGBoost.
7. Customer Segmentation – Divided customers into Low, Medium, High LTV.
8. Result Saving – Exported predictions with customer IDs and segments to CSV.

Conclusion
The project demonstrates the value of combining transactional data with machine learning models to predict CLV.

Key insights:
- High-value customers form a smaller proportion but contribute significantly to revenue.
- Medium-value customers can be nurtured with loyalty and cross-selling programs.
- Low-value customers may represent churn risk and need tailored retention strategies.

This predictive approach equips businesses with actionable insights to optimize marketing, improve retention, and maximize profitability. Future extensions could include adding customer behavior features, time-series models, or testing deep learning architectures for more accurate predictions.