Internship Project Report

Project Title: Customer Lifetime Value (LTV) Prediction

Intern Name: ***BHAVESH SANTOSHKUMAR TUPE***

Duration: 2 Weeks

📘 1. Introduction

Customer Lifetime Value (LTV) is one of the most important metrics for customer-centric businesses. It represents the total revenue a business can expect from a customer throughout their relationship. The aim of this project was to build a machine learning model that can predict a customer's LTV based on their historical purchase behavior. The insights derived help in customer segmentation, personalized marketing, and optimizing return on investment.

📄 2. Abstract

This project involved building an LTV prediction model using transactional data from a UK-based online retailer. The dataset contained over a million records including customer IDs, invoice details, product quantities, prices, and timestamps. After preprocessing and feature engineering, an XGBoost regression model was used to predict the LTV for each customer. Post-prediction, customers were segmented into value-based groups (Low, Mid, High, Very High) to guide strategic marketing. The model performance was evaluated using MAE and RMSE, and the results were visualized to validate the segmentation logic.

🛠 3. Tools Used

- **Platform**: Google Colab
- **Language**: Python
- **Libraries**: pandas, numpy, matplotlib, seaborn, scikit-learn, XGBoost
- **Data Format**: Excel/CSV files

🔧 4. Steps Involved in Building the Project

a. Data Cleaning & Preprocessing

- Removed cancelled transactions and records with missing Customer IDs
- Calculated TotalPrice = Quantity × UnitPrice
- Parsed and cleaned the InvoiceDate column

b. Feature Engineering

- **Recency**: Days since last transaction

- **Frequency**: Total number of invoices per customer
- **AOV (Average Order Value)**
- **LTV (Target Variable)**: Total spending per customer

c. Model Building

- Algorithm: **XGBoost Regressor**
- Input features: Recency, Frequency, AOV
- Target: Total LTV
- Evaluation Metrics:
    - MAE: **703.16**
    - RMSE: **1672.68**

d. Customer Segmentation

Customers were classified into 4 LTV-based segments:

- **Low**: Bottom 25%
- **Mid**: 25–50%
- **High**: 50–75%
- **Very High**: Top 25%

e. Visualization

- Histogram and Boxplot showed LTV distribution was right-skewed
- Segment-wise plots highlighted clear separation in value tiers

---

🏁 5. Conclusion

The LTV prediction model provides valuable insights into customer behavior and value contribution. It enables businesses to identify and focus on high-LTV customers for retention and upselling. The segmentation approach can be directly used to implement personalized marketing strategies and optimize resource allocation.