Customer Lifetime Value (LTV) Prediction Project

Prepared by: VINODINI R

Abstract

Customer Lifetime Value (LTV) prediction plays a vital role in e-commerce business strategies, enabling companies to identify their most profitable customers and create personalized marketing campaigns. This project aims to build a **machine learning model** that predicts the expected revenue a customer will generate throughout their relationship with the company. Using **historical transaction data** and **customer behavior data**, we implemented an **XGBoost regression model** to predict LTV and segmented customers into **High**, **Medium**, and **Low** value groups. The project focuses on **data preprocessing, feature engineering, model training, evaluation, and prediction**, delivering actionable insights for targeted marketing strategies.

Introduction

Customer retention and targeted marketing have become key business priorities in today's competitive e-commerce environment. Identifying valuable customers early allows companies to optimize their marketing efforts and maximize profits.

This project uses historical purchase behavior, transaction frequency, recency, and other customer attributes to predict **Customer Lifetime Value**. The model helps businesses classify customers intodifferent segments and focus their marketing efforts on high-value customers, improving overall return on investment (ROI).

Tools & Technologies Used:

Python	Core programming language
Pandas & NumPy	Data loading, cleaning, and manipulation
Matplotlib & Seaborn	Data visualization
Scikit-learn	Preprocessing, evaluation metrics, model validation
XGBoost	Machine learning model for regression
Joblib	Saving & loading trained models
VS Code	Development environment
Excel	Data storage & exploration

Steps Involved in Building the Project

1. Data Collection & Understanding

- Two datasets were used:
    - **Transactions Dataset** → Contains customer purchase history
    - **Customer Behavior Dataset** → Includes demographics, ratings, membership details, etc.
- Checked for missing values, duplicates, and data inconsistencies.

2. Data Preprocessing & ETL

- Converted dates into proper formats.
- Merged both datasets using **Customer ID**.
- Handled missing values and encoded categorical columns.

3. Feature Engineering

•Created new features such as:

- `frequency` → Number of purchases per customer
- `recency_days` → Days since last purchase
- `product_diversity` → Unique product categories purchased
- Encoded **membership type**, **satisfaction level**, and **discount usage**.

4. Model Building

•Chose **XGBoost Regressor** for better accuracy and performance.

•Trained the model using the engineered features and optimized hyperparameters.

•Evaluated performance using **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)**.

5. Prediction & Customer Segmentation

•Predicted **Customer Lifetime Value (LTV)** for all customers.

•Segmented customers into **High**, **Medium**, and **Low** groups using quantiles.

•Exported predictions to `LTV_predictions.csv` for business insights.

Conclusion

This project successfully developed a **machine learning pipeline** to predict customer lifetime value using real-world transaction and behavior data. By implementing **ETL, feature engineering, and XGBoost regression**, we built an accurate and efficient model to help businesses identify high-value customers and design personalized marketing campaigns.

The results can significantly improve **customer retention**, **business strategy**, and **revenue forecasting**. Future enhancements can include **hyperparameter tuning with Optuna**, integrating **deep learning models**, and deploying the solution using **Streamlit** or **Flask** for interactive dashboards.