Customer Lifetime Value (LTV) Prediction Project

Abstract

Customer Lifetime Value (LTV) prediction is one of the most crucial strategies for businesses in the modern digital era. By estimating the future revenue that a customer will bring to a company, managers can make smarter decisions in marketing, customer retention, and profitability.
In this project, we used **machine learning models** to analyze e-commerce customer purchase behavior and predict their expected lifetime value. A structured workflow of **data preprocessing, feature engineering, model training, evaluation, and customer segmentation** was followed. The outcomes of this project enable businesses to identify **high-value customers**, reduce customer churn, and optimize targeted marketing campaigns.

Introduction

The growth of e-commerce has led to the availability of vast amounts of customer data. Every purchase, transaction, and interaction provides meaningful information about customer behavior. However, raw data alone cannot directly inform marketing strategies.

Customer Lifetime Value (LTV) serves as a key metric, estimating how much revenue a customer will generate over their entire relationship with a company. Predicting this value allows businesses to:

- Focus marketing budgets on the most valuable customers.
- Design loyalty programs to retain medium and high-value customers.
- Identify customers at risk of churn and take proactive steps.

In this project, a **synthetic e-commerce dataset** was used to simulate real-world purchase patterns. With the help of **XGBoost regression models**, predictions were generated for future customer lifetime values. The final deliverables include a Python notebook, a trained ML model, and a CSV file with customer-level predictions.

Tools Used

- **Python Libraries**:
    - Pandas & NumPy – data handling and feature generation
    - Matplotlib & Seaborn – data visualization
    - Scikit-learn – preprocessing and evaluation metrics
    - XGBoost – regression model for LTV prediction
    - Joblib – saving the trained model

- **Platforms**:
    - Google Colab / Jupyter Notebook for development
    - Excel for quick visualization of final predictions

Steps Involved in Building the Project

1. **Data Collection**
    - Kaggle dataset: *E-commerce Customer Behavior and Purchase Dataset*.
2. **Data Preprocessing**
    - Removed null or inconsistent values.
    - Standardized column names.
    - Converted purchase dates into datetime format.
3. **Feature Engineering**
    - **Recency**: Number of days since the last purchase.
    - **Frequency**: Total number of purchases by a customer.
    - **Monetary Value**: Total purchase amount.
    - **Average Order Value (AOV)**: Ratio of monetary value to frequency.
4. **Model Training**
    - Target variable: log-transformed total purchase value.
    - Algorithm: **XGBoost Regressor**.
    - Dataset split: Training (70%), Validation (15%), Test (15%).
5. **Model Evaluation**
    - Metrics:
        - MAE (Mean Absolute Error)
        - RMSE (Root Mean Squared Error)
        - R^2 Score (Goodness of Fit)
    - Visualization: scatter plots of predicted vs actual LTV.
6. **Customer Segmentation**
    - Customers grouped into **High, Medium, and Low LTV segments**.
    - Useful for marketing and retention strategies.

7. **Deliverables**
    - Jupyter Notebook (.ipynb) with all code and visualizations.
    - Trained Model (.joblib) saved using Joblib.
    - CSV file containing predicted LTV for each customer.

Conclusion

This project successfully developed a machine learning pipeline to predict **Customer Lifetime Value (LTV)** using e-commerce transaction data. The results provide businesses with actionable insights:

- **High-value customers** can be prioritized with exclusive offers and loyalty programs.
- **Medium-value customers** can be encouraged with personalized marketing.
- **At-risk customers** can be retained using targeted retention strategies.

By combining predictive modeling with segmentation, companies can **maximize revenue, minimize churn, and allocate resources more effectively**.

Future Scope:

- Implementing **deep learning models** for more accurate predictions.
- Using **time-series models** to capture customer purchase patterns over time.
- Integrating predictions directly into **CRM platforms** for real-time marketing.