Customer Lifetime Value Prediction Model

Introduction

Customer Lifetime Value (CLV) is a fundamental metric used to assess the net profit attributed to the entire future relationship with a customer. Understanding CLV helps organizations tailor their acquisition and retention strategies, focusing on customers who deliver the highest long-term value. This project aims to predict the CLV using historical purchase data and machine learning techniques, providing insights that can guide marketing and sales decisions.

Abstract

In this project, we analyze two years of online retail transaction data from the UK and Europe. The dataset comprises invoices, product codes, quantities, prices, and customer identifiers. After cleaning and transforming the data, we engineered key metrics such as Recency, Frequency, Total Quantity, and Average Order Value. These were then used to train a regression model (XGBoost) to predict the monetary value each customer contributes. This predicted value, treated as CLV, enables us to classify customers into value-based segments. The model achieved good performance on test data and provides a foundation for data-driven customer engagement strategies.

Tools Used

- **Programming Language**: Python
- **Libraries**: Pandas, NumPy, Scikit-learn, XGBoost, Seaborn, Matplotlib, Joblib
- **Notebook Interface**: Jupyter Notebook
- **File Format**: Excel (Online Retail II dataset), CSV
- **Model Type**: Regression (XGBoost Regressor)
- **Evaluation Metrics**: MAE (Mean Absolute Error), RMSE (Root Mean Squared Error)

Steps Involved in Building the Project

1. **Data Integration**:
    - Combined transaction records from 2009–2010 and 2010–2011 Excel sheets.
2. **Data Cleaning**:
    - Removed missing customer IDs and invalid transactions (`Quantity ≤ 0` , `Price ≤ 0`)
    - Created a `TotalPrice` feature by multiplying quantity and price.
3. **Feature Engineering**:
    - Calculated:
        - **Recency**: Days since the most recent transaction
        - **Frequency**: Number of unique invoices per customer
        - **Total Quantity**: Sum of all items purchased
        - **Monetary Value**: Total amount spent
        - **Average Order Value**: Monetary Value / Frequency

4. **Model Building**:
    - Chose `XGBRegressor` for its robustness and accuracy
    - Split data into 80% training and 20% test sets
    - Trained model using default hyperparameters
    - Evaluated performance using MAE and RMSE
5. **Prediction & Segmentation**:
    - Predicted CLV for each customer
    - Used `qcut` to divide customers into 4 segments:
        - **High**, **Mid-High**, **Mid-Low**, and **Low**
    - Saved results to `predicted_clv.csv` for business insights
6. **Model Export**:
    - Exported trained model to `clv_model.pkl` using Joblib for reuse or deployment.

Conclusion

This project provides a clear, scalable approach for CLV prediction using transaction-level data. By predicting which customers are likely to be most valuable, businesses can allocate resources more efficiently, design better loyalty programs, and improve customer retention. Future work can include integrating customer demographics, applying time-series modeling, or testing other algorithms for improved accuracy. The current implementation can be extended to other industries such as e-commerce, banking, or telecom with similar datasets.