Customer Lifetime Value (CLV) Prediction – Project Report
Author: Shraddha Jharbade

Introduction

Customer Lifetime Value (CLV) is a key metric used by businesses to estimate the total revenue a customer will generate throughout their relationship with the company. This project focuses on predicting CLV using historical transaction data, helping businesses to target high-value customers and improve retention strategies.

Abstract

The objective of this project is to segment customers based on Recency, Frequency, and Monetary value (RFM) and build a supervised learning model to predict their future CLV. The analysis uses a 3-month window of past transactions for feature generation and a 6-month window for validating future revenue. The UK region's customer data was filtered for consistency. KMeans clustering was used for segmentation, and XGBoost regression was applied for predicting revenue.

Tools Used

- Languages: Python
- Libraries: Pandas, NumPy, Matplotlib, Seaborn, scikit-learn, XGBoost
- IDE: Jupyter Notebook
- Visualization: Matplotlib, Seaborn
- Model Serialization: Pickle

Steps Involved in Building the Project

1. Data Preprocessing
    - Loaded data.csv and filtered records from the United Kingdom.
    - Converted InvoiceDate to datetime format.
    - Created two datasets: tx_3m (March–May 2011) for training and tx_6m (June–November 2011) for target CLV.
2. RFM Feature Engineering
    - Recency: Days since last purchase.
    - Frequency: Count of invoices per customer.

    - Monetary: Total revenue (Quantity × UnitPrice).
3. Customer Segmentation
    - Applied KMeans clustering on RFM features individually.
    - Used the Elbow method to determine optimal clusters.
    - Assigned scores to Recency, Frequency, and Monetary clusters to calculate a final CLV score.
4. Predictive Modelling
    - Merged RFM features with actual revenue from tx_6m.
    - Used XGBoost Regressor for CLV prediction.
    - Evaluated using RMSE and cross-validation.
5. Visualization
    - Displayed cluster distributions and revenue segments.
    - Illustrated feature importance from the model.
    - Compared RFM characteristics across clusters.

Conclusion

The project successfully segmented customers and predicted their future value using RFM features and XGBoost modelling. High-value customers were characterized by recent and frequent purchases with high spend. These insights can drive marketing strategies such as personalized promotions, upselling, and churn prevention. The model performed reliably, making it a useful tool for customer value forecasting.

Recommendations

- Retention Campaigns: Focus on recent, high-frequency customers with top CLV scores.
- Upselling Opportunities: Target mid-value customers showing frequent activity but lower spend.
- Churn Risk: Reach out to low-recency, low-frequency segments with incentives.