Customer Lifetime Value Prediction Model

Abstract

This project focuses on predicting Customer Lifetime Value (CLTV) using advanced machine learning techniques. By analyzing historical transaction data, the goal is to identify high-value customers and support targeted marketing and retention strategies. The project employs feature engineering on RFM metrics (Recency, Frequency, Monetary), along with additional behavioural features. Two models were trained — Random Forest (achieving $R^2 = 0.9986$) and XGBoost ($R^2 = 0.9753$), enabling highly accurate LTV predictions. The output supports intelligent customer segmentation and business decision-making.

Introduction

CLTV is a core metric in customer analytics, measuring the expected revenue from a customer over their lifetime. Businesses use it to:

- Allocate resources efficiently
- Focus on high-value customers
- Drive personalized marketing

This project aims to build a **predictive pipeline** for LTV using transaction-level data. Starting with RFM feature engineering, the model incorporates statistical analysis, machine learning, and visual storytelling. The final output segments customers and offers a clear path to data-driven strategic decisions.

Tools Used

- **Programming Language**: Python
- **Libraries**:
    - **NumPy** – numerical operations
    - **Pandas** – data manipulation
    - **Scikit-learn** – ML algorithms (Random Forest)
    - **XGBoost** – gradient boosting
    - **Matplotlib, Seaborn** – data visualization
    -

Steps Involved in Building the Project

1. **Data Preprocessing**
    - Removed null and duplicate entries
    - Formatted invoice dates, created total transaction value
    - Filtered canceled transactions
2. **Feature Engineering**
    - RFM features (Recency, Frequency, Monetary)

- Customer tenure
- Average order value (AOV)
- Purchase frequency
- Product variety and transaction variability

3. **LTV Calculation**
    - Used frequency × AOV × lifespan × profit margin (20%)
    - Adjusted lifespan using customer tenure
    - Normalized for prediction scale
4. **Model Development**
    - Trained Random Forest and XGBoost regressors
    - Performed hyperparameter tuning
    - Evaluated using R^2 and MAE metrics
5. **Customer Segmentation**
    - Used quantile-based segmentation:
        - **Platinum**: Top 25%
        - **Gold**: 25–50%
        - **Silver**: 50–75%
        - **Bronze**: Bottom 25%
6. **Visualization**
    - CLTV distribution plots
    - Top 10 customers by predicted value
    - Feature importance graphs
    - Segment-wise breakdown

Conclusion

This project delivers a scalable and accurate model for Customer Lifetime Value prediction. The results demonstrate high predictive accuracy ($R^2 \approx 0.99$), enabling real business impact through:

- Strategic customer segmentation
- Targeted marketing investment
- Retention optimization
- Improved customer service prioritization

With future improvements, the model can be integrated into **CRM systems**, support **real-time scoring**, and evolve into a **live customer analytics dashboard**.