Customer Lifetime Value (LTV) Prediction using XGBoost

Introduction

Customer Lifetime Value prediction plays a crucial role in customer relationship management and strategic business decisions. Knowing the future value of customers helps in optimizing acquisition costs, designing retention strategies, and maximizing long-term profitability. The project aimed to predict LTV using behavioral and transactional data features such as frequency, recency, and monetary value, combining statistical analysis with machine learning for insightful predictions.

Abstract

This project focuses on predicting Customer Lifetime Value (LTV) using data-driven machine learning techniques. LTV represents the total revenue a business can expect from a customer throughout their relationship. The goal was to build a robust model using XGBoost Regression to estimate future customer value and segment them into actionable categories for strategic marketing. The process included data preprocessing, feature selection based on correlation, model training, and customer segmentation.

Tools and Technologies Used

Python: Core programming language for analysis and modeling. • Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, and xgboost. • IDE: Jupyter Notebook and Spyder for development and visualization. • Excel: For initial inspection and understanding of data.

Steps Followed

1. Data Collection and Understanding: The dataset included customer-level purchase behavior and transactional variables, with LTV as the target. 2. Data Cleaning and Preprocessing: Handled missing values, removed duplicates, and encoded categorical variables. 3. Feature Engineering: Derived variables like Recency, Frequency, and Monetary Value; log-transformed skewed variables to normalize data. 4. Feature Selection: Used correlation analysis to retain only relevant features. 5. Model Building: Used XGBoost Regressor with tuned hyperparameters for regression. 6. Evaluation: Model achieved $R^2 = 0.61$, RMSE = 7210.57, and MAE = 1819.51. 7. Prediction: Predicted LTV for the full dataset and stored it in a new column.8. Customer Segmentation: Applied quantile-based and K-Means segmentation to categorize customers as Low, Medium, or High value.

Conclusion

The XGBoost model effectively predicted Customer Lifetime Value with an R^2 of 0.61, providing moderate predictive accuracy. Segmentation using quantiles and K-Means enabled the identification of Low, Medium, and High-value customers. These insights support customer retention strategies, cross-selling opportunities, and improved marketing ROI. The project demonstrates the power of machine learning for actionable business intelligence and strategic decision-making.