Customer Lifetime Value Prediction Report

April 27, 2025

Abstract

This report outlines the development of a Customer Lifetime Value (CLV) prediction model to enhance targeted marketing strategies. Utilizing customer purchase data, the model leverages machine learning techniques to forecast CLV based on purchase behavior. The project involves data preprocessing, feature engineering, model training, and customer segmentation. Python libraries such as Scikit-learn and XGBoost were employed, with results validated using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The final output segments customers into low, medium, and high-value groups, enabling optimized marketing efforts.

Introduction

Customer Lifetime Value (CLV) is a critical metric for businesses aiming to optimize marketing strategies and maximize profitability. By predicting the future value of customers based on their purchase behavior, companies can allocate resources effectively, targeting high-value customers for retention and personalized campaigns. This project develops a predictive model to estimate CLV using historical customer data, including purchase history, tenure, and total spent. The objective is to create a robust model that segments customers based on predicted CLV, facilitating data-driven marketing decisions.

Tools Used

The project was built using the following tools:

- **Python**: Core programming language for data processing and modeling.
- **Scikit-learn**: For implementing Random Forest regression and model evaluation metrics (MAE, RMSE).
- **XGBoost**: For advanced gradient boosting regression to enhance prediction accuracy.
- **Pandas and NumPy**: For data manipulation and preprocessing.
- **Excel**: For storing and exporting the final segmented customer data.

Steps Involved in Building the Project

The development of the CLV prediction model followed a structured approach:

1. **Data Preprocessing**: Loaded customer data from a CSV file containing customer IDs, purchase history, tenure, total spent, and CLV. Missing or invalid values (e.g., negative total spent or zero tenure) were removed to ensure data quality.

2. **Feature Engineering**: Created key features to capture customer behavior:

- *Recency*: Derived from tenure (in months).
- *Frequency*: Based on the number of purchases.
- *Average Order Value (AOV)*: Calculated as total spent divided by purchase frequency.
- Additional features: Spend per month and purchases per month to reflect spending patterns over time.

3. **Model Training**: Two regression models were trained:

- Random Forest Regressor: A robust ensemble method for capturing non-linear relationships.
- XGBoost Regressor: A gradient boosting algorithm for improved accuracy.

The dataset was split into 80% training and 20% testing sets. Features included recency, frequency, AOV, spend per month, and purchases per month, with CLV as the target variable.

4. **Model Evaluation**: Model performance was assessed using MAE and RMSE. Both models showed strong predictive capabilities, with XGBoost slightly outperforming Random Forest due to its ability to handle complex feature interactions.

5. **Customer Segmentation**: The trained XGBoost model predicted CLV for all customers. Customers were segmented into three groups (Low, Medium, High) using quartile-based thresholds on predicted CLV. Each segment was assigned a unique identifier for tracking.

6. **Output Generation**: The final dataset, including predicted CLV and segment labels, was exported to an Excel file for use in marketing strategies.

Conclusion

The CLV prediction model successfully estimates customer lifetime value and segments customers based on their predicted value. By leveraging features such as recency, frequency, and AOV, the model provides actionable insights for targeted marketing. The use of XGBoost ensured high predictive accuracy, as validated by low MAE and RMSE scores. The segmentation output enables businesses to prioritize high-value customers, optimize resource allocation, and enhance customer retention strategies. Future improvements could include incorporating additional features (e.g., product categories) or exploring deep learning models for even greater accuracy.