CUSTOMER LIFETIME VALUE (LTV) PREDICTION USING MACHINE LEARNING

Pratyush Rawat
Pratyushrawat8846@gmail.com

Abstract

This project focuses on predicting the Customer Lifetime Value (LTV) using historical purchase behaviour from an online retail dataset. LTV prediction is crucial in targeted marketing, customer retention, and resource allocation. The project involves extensive data preprocessing, feature engineering, and the application of multiple regression models (Linear Regression, Decision Tree, Random Forest, XGBoost, and Gradient Boosting) to identify high-value customers. The final output includes customer segmentation and visualizations for business insights.

Introduction

In today's competitive e-commerce environment, understanding the long-term value of a customer is vital for sustainable growth. LTV estimates the total revenue a business can expect from a customer throughout their relationship. By using machine learning models, we aim to automate LTV prediction, enabling personalized campaigns and strategic decisions.

The dataset contains historical transaction records, including invoice numbers, item descriptions, quantities, prices, dates, and customer information. The goal is to train models that predict LTV based on derived behavioural features like Recency, Frequency, and Average Order Value (AOV).

Tools Used

Languages & Libraries	**Algorithms**
Python, NumPy, Pandas, Seaborn, Matplotlib Scikit-learn (Linear Regression, Decision Tree, Random Forest, Gradient Boosting) XGBoost (optional, for gradient-boosted trees) **IDE**: Jupyter Notebook / VS Code	Supervised Learning (Regression) Customer Segmentation via Quantile Binning (qcut)

Methods

Import Libraries

Essential libraries for data handling, visualization, model training, and evaluation were imported.

Data Preprocessing
Loaded the dataset and removed rows with missing customer IDs.
Removed canceled orders (invoices starting with 'C').
Converted InvoiceDate to datetime and calculated TotalPrice.

Feature Engineering
Derived customer-level features:
Recency: Days since the last purchase.
Frequency: Number of unique invoices.
Monetary (LTV): Total spending.
AOV: Average Order Value.

Train-Test Split
Split the dataset into training and test sets (80%-20%) using train_test_split.
Model Building (Trained and evaluated the following models)
Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, XGBoost Regressor (optional)
Model Evaluation (Models were evaluated using below parameters)
Mean Absolute Error (MAE)
Root Mean Squared Error (RMSE)

Results

Model	MAE	RMSE
Linear Regression	1971.32	6741.71
Decision Tree	1932.61	11418.25
Random Forest	1393.97	7160.36
XGBoost	1620.88	8380.43
Gradient Boosting	1492.19	10040.20

Random Forest had the lowest MAE, making it the preferred model.

Prediction & Segmentation
Predicted LTV for all customers using the best model.
Segmented customers into "Low", "Medium", and "High" based on predicted LTV using quantiles.

Conclusion

This project demonstrates how machine learning can be effectively applied to predict Customer Lifetime Value using structured purchase data. The approach helps businesses: *Identify high-value customers, personalize marketing efforts, Improve customer retention strategies.* With the Random Forest model achieving the best accuracy, the solution provides both predictive power and actionable segmentation for strategic decision-making.