Customer Lifetime Value Prediction Model

Introduction

Customer Lifetime Value (CLTV) is a crucial metric used by businesses to estimate the total revenue a company can expect from a customer over the course of their relationship. This project aims to build a predictive model to estimate CLTV based on historical purchasing behaviour, helping businesses to identify and retain high-value customers through targeted marketing strategies.

Abstract

The goal of this project is to predict the CLTV of customers using regression techniques. We used an online retail dataset containing transaction data between 2009 and 2011. After preprocessing the data, we performed RFM (Recency, Frequency, Monetary) analysis, followed by feature engineering and training of regression models such as Random Forest and XGBoost. The performance of each model was evaluated using metrics like MAE, RMSE, and R² score. The model was then used to predict CLTV for each customer, enabling segmentation based on predicted value.

Steps Involved in Building the Project

1.Data Preprocessing

Loaded and cleaned online\_retail\_II.csv

Removed null values and invalid transactions (like negative quantity)

Converted Invoice Date to datetime format

2. Feature Engineering

Calculated RFM features for each customer

Derived Average Order Value (AOV)

3. RFM Segmentation

Scored customers based on Recency, Frequency, and Monetary value

Combined scores for customer segmentation

4. Model Building

Split data into train and test sets

Trained two models: Random Forest and XGBoost

Applied GridSearchCV for hyperparameter tuning

5. Model Evaluation

Evaluated using MAE, RMSE, R² Score

XGBoost performed better with:

MAE: 29.74

RMSE: 556.22

R² Score: 0.69

6. Output

Predicted CLTV for each customer

Exported results to CSV

Created visualizations for model performance and customer segments

Conclusion

This project successfully demonstrates how regression models can be used to predict Customer Lifetime Value. The use of RFM analysis, effective feature engineering, and model comparison enhanced the reliability of our predictions. XGBoost emerged as the best model. The predicted CLTV scores can now be used for personalized marketing, improving customer retention, and maximizing profitability.