**Customer Lifetime Value Prediction**

# Introduction

Customer Lifetime Value (CLTV) is a crucial metric that estimates the total revenue a business can expect from a customer throughout their relationship. In sectors like e-commerce and subscription-based services, accurately predicting CLTV empowers businesses to optimize marketing efforts, reduce churn, and identify high-value customers. This project uses machine learning and transactional behavioral data to build a robust CLTV prediction pipeline with actionable segmentation.

# Abstract

This project implements a machine learning-based workflow to estimate CLTV using customer purchase data. After rigorous data preprocessing and feature engineering, we trained an XGBoost regression model to predict monetary value. The model was evaluated using MAE and RMSE metrics, and explainability was achieved using SHAP. RFM segmentation and predicted LTV were combined to categorize customers into four value-based segments: **Top**, **High**, **Medium**, and **Low**. These segments guide personalized business strategies and retention campaigns.

# Tools Used

- Python: pandas, matplotlib, seaborn, scikit-learn

- XGBoost for model training

- SHAP( SHapley Additive exPlanations) for explainability

- PyCharm for development, Microsoft Excel for evaluation and final review

- File Outputs : Visuals(PNG), CSV Export

# Methodology

1. **Data Cleaning & Preparation**

* Dropped rows with null CustomerID or duplicates.
* Filtered out canceled/invalid orders with non-positive quantity or unit price.
* Created TotalPrice = Quantity × UnitPrice.
* Converted InvoiceDate to datetime and standardized column formatting.

**2. Feature Engineering**

* Recency: Days since last purchase before snapshot date
* Tenure: Duration between first and last purchase
* Frequency: Number of unique purchases per customer
* Monetary: Total spend per customer
* AOV (Average Order Value): Monetary / Frequency

**3. RFM Segmentation**

* Assigned quartile scores to Recency, Frequency, and Monetary features.
* Constructed composite RFM scores (e.g., 444, 123).
* Segmented customers based on predicted LTV into: Top, High, Medium, Low.

**4. Revenue Trend Analysis**

* Aggregated total revenue by month and day of the week for seasonal insights.
* Visualized trends using bar and distribution plots.

**5. Model Development: XGBoost Regressor**

* Features: Recency, Tenure, Frequency, AOV
* Target: Monetary
* Model Params: n\_estimators=100, objective='reg:squarederror', random\_state=42
* Split data into 80/20 train-test sets.

**6. Evaluation Metrics**

* MAE: *877.42*
* RMSE: *9414.35*
* Compared to linear regression baseline: XGBoost demonstrated superior performance.

**7. Explainability with SHAP**

* Used SHAP values to identify feature influence.
* Most influential predictors: Frequency, Recency, and Tenure.
* Generated SHAP summary plot to visualize global feature importance.

8. Output

* Predicted LTV with Segment: cltv\_predictions.csv
* Visuals:
  + LTV distribution histogram
  + Feature correlation heatmap
  + SHAP summary plot
* All visualizations exported as PNGs for reporting and sharing.

# Conclusion

The project successfully demonstrates the use of machine learning to estimate customer value from purchase behavior. This end-to-end pipeline effectively combined transactional data, machine learning, and explainability tools to deliver actionable customer value predictions. The model’s interpretability, paired with intuitive segmentation, enables strategic customer retention, revenue forecasting, and targeted engagement. This framework is scalable for real-time integration or further refinement using time-series forecasting, deep learning, or business intelligence platforms.