Customer Lifetime Value Prediction – Project Report

Objective

- Goal: Predict Customer Lifetime Value (CLV) from transactional data and segment customers into Low, Medium, and High value groups to inform retention and marketing strategies.

Data

- Source: transactions.csv
- Key fields used: CustomerID, InvoiceDate, Amount
- Preprocessing: Parsed InvoiceDate to datetime and aggregated transactions at the CustomerID level.

Feature Engineering

- Computed per-customer features: Recency (days since last purchase), Frequency (number of transactions), AOV (average order value).
- Target: TotalSpend (sum of Amount per customer).

Modeling Approach

- Train/Test Split: 80/20 split with random_state=42.
- Models: RandomForestRegressor (n_estimators=200) and XGBRegressor (n_estimators=200, learning_rate=0.1).
- Features used: Recency, Frequency, AOV.

Evaluation

- Metrics: MAE and RMSE on the test set.
- Selected model for final predictions: XGBoost.

Results

- Customers scored: 200
- Predicted_CLV summary: min 617.96, mean 3,144.10, median 3,168.83, max 7,210.37, std 1,344.30
- Top 5 customers by predicted CLV: C0075: 7,210.37; C0072: 6,202.51; C0131: 6,025.98; C0060: 5,941.51; C0082: 5,582.23
- Segment distribution: High: 68 (34.0%), Medium: 66 (33.0%), Low: 66 (33.0%)

Predictions and Segmentation

- Predicted CLV for all customers using XGBoost.

- Created CLV segments via predicted CLV quantiles: Low (<=33rd), Medium (33rd–66th), High (>66th).
- Output file: Predicted_CLV.csv with CustomerID, Recency, Frequency, AOV, TotalSpend, Predicted_CLV, CLV_Segment.

Visualization

- Histogram with KDE of Predicted_CLV distribution.

Key Insights

- Recency, Frequency, and AOV collectively explain a significant portion of variance in Total Spend.
- Segmentation helps prioritize High-value customers and nurture Medium/Low segments.

Limitations

- Target is historical spend (proxy for CLV), not discounted cash flows.
- No time-based validation; temporal split recommended.
- Potential data leakage from features/target derived from same period.
- Churn probability and marketing response not modeled.

Recommendations and Next Steps

- Use temporal split (train on earlier periods, test on later).
- Add richer features: trends, intervals, product mix, tenure.
- Incorporate survival/churn and uplift modeling.
- Calibrate CLV for a business horizon with discounting.
- Monitor performance drift and retrain periodically.

Environment and Reproducibility

- Python libraries: pandas, numpy, scikit-learn, seaborn, matplotlib, xgboost.
- Install: pip install pandas numpy scikit-learn seaborn matplotlib xgboost reportlab
- Run: Execute data.py, then this script to produce CLV_Project_Report.pdf.

Deliverables

- Predicted_CLV.csv: Per-customer predictions and segments.
- Console logs: MAE and RMSE for both models.
- CLV distribution plot shown at runtime.

Business Use

- High: Loyalty perks, VIP support, early access.
- Medium: Targeted cross-sell/upsell, personalized offers.
- Low: Cost-effective retention nudges, reactivation campaigns.