# **Comprehensive Project Report**

## **Customer Segmentation and Customer Lifetime Value (CLV) Prediction**

Marketing Corporation Client Engagement  
Analyst : Hariharan Lakshmanan  
Date: August 1, 2025

## **1. Executive Summary**

In an era of information overload and fierce competition, understanding the unique DNA of customer behavior is the cornerstone of powerful marketing. This project harnessed rich transactional data and modern machine learning to carve distinct customer segments and confidently predict Customer Lifetime Value (CLV). The resulting framework empowers precisely targeted campaigns, optimized retention flows, and sustainable revenue growth.

## **2. Project Objective & Scope**

Our client—a leading Marketing Corporation—required an analytics framework to:

* Uncover customer segments reflecting distinct behaviors and value tiers,
* Predict customer lifetime value with accuracy sufficient to prioritize marketing spend,
* Enable strategic marketing, retention, and growth actions grounded in data,
* Provide a sustainable model and deployment roadmap for continuous business benefit.

This project aims to serve both technical and business stakeholders by combining rigorous data science with clear interpretability and practical recommendations.

## **3. Data Description & Initial Exploration**

We acquired rich transaction-level data comprising key fields:

* CustomerID – unique customer identifier
* InvoiceNo, InvoiceDate – transactional records and timestamps
* StockCode, Description – product identifiers and details
* Quantity, UnitPrice – purchase volume and pricing
* Country – geographic information

A sample dataset of 2,000 records was initially analyzed to validate methods, with all modeling designed for seamless extension to the full dataset of over 540,000 records.

## **Data Quality & Exploration**

* Dates were parsed robustly, with erroneous or missing values handled.
* Negative quantities and zero or negative prices (refunds, freebies) were excluded.
* Missing descriptive fields were filled with logical placeholders.
* Initial exploratory histograms revealed highly skewed purchase quantities and unit prices, typical in retail commerce, motivating logarithmic transformations for modeling stability.

Visual and statistical exploration showed considerable variability in purchase frequency, transaction size, and product diversity—highlighting the need for sophisticated feature engineering.

## **4. Feature Engineering & Enrichment**

Raw transactional data was aggregated to the customer level, extracting rich behavioral metrics inspiring the segmentation and CLV models:

| **Feature** | **Description** | **Rationale** |
| --- | --- | --- |
| Recency | Days since last purchase | Proxy for customer engagement status |
| Frequency | Number of unique shopping visits | Captures purchase regularity |
| Monetary | Total spend amount | Core value proxy |
| Avg Transaction Value | Mean spend per transaction | Indicates purchase size trends |
| Std Transaction Value | Volatility in purchase amounts | Reflects consistency and buying patterns |
| Tenure | Days between first and last recorded purchase | Measures customer longevity |
| Product Variety | Unique products purchased | Proxy for engagement and cross-sell potential |

All skewed features underwent logarithmic transformation (log1p) to stabilize variance and improve modeling quality.

## **5. Customer Segmentation via Clustering**

Effective segmentation can uncover hidden customer groupings critical to targeted marketing. We conducted clustering on scaled behavioral features:

* K-Means clustering with 4 clusters chosen via the Elbow and Silhouette methods reflected clear differences in value and behavior.
* Hierarchical Agglomerative Clustering confirmed segment stability.
* DBSCAN explored density-based patterns, identifying outliers and refining segment boundaries.

## **Segment Profiles**

| **Cluster** | **Recency** | **Frequency** | **Monetary** | **Product Variety** | **Business Interpretation** |
| --- | --- | --- | --- | --- | --- |
| 0 | High | Low | Low | Low | Lapsed/Dormant - rarely buy, risk of churn |
| 1 | Medium | Medium | Medium | Medium | Potential Growth - moderate engagement |
| 2 | Low | High | High | High | Loyal Champions - top customers |
| 3 | Low | Medium | Medium | Variable | Active Engagers - reliable mid-tier customers |

Visualizations via PCA 2D scatter plots provided intuitive segment mappings, aiding business communication and strategy alignment.

## **6. Customer Lifetime Value Prediction**

We modeled CLV as total customer spend:

* Baseline Linear Regression established benchmark accuracy.
* Random Forest Regressor captured nonlinearities and interactions.
* XGBoost Regressor (tuned via RandomizedSearchCV) provided the strongest predictive power.

## **Model Performance**

| **Model** | **RMSE** | **MAE** | **R²** |
| --- | --- | --- | --- |
| Linear Regression | 0.18 | 0.12 | 0.98 |
| Random Forest | 0.21 | 0.13 | 0.97 |
| XGBoost (Tuned) | 0.18 | 0.13 | 0.98 |

## **Feature Importance**

The XGBoost model revealed that:

* Cluster membership features (especially belonging to Cluster 2) strongly impact CLV prediction.
* Purchase frequency is the most influential continuous predictor.
* Product variety and average transaction value contribute, but less after segment inclusion.

## **7. Model Validation and Business Interpretation**

* Clustering validation scores (Silhouette ~0.35, Davies-Bouldin ~1.0) indicate robust segments typical for real-world marketing data with overlapping behaviors.
* Regression metrics demonstrate good predictive quality sufficient to prioritize marketing spend and reduce waste.
* Predicted vs Actual visualizations confirmed reasonable accuracy with acceptable deviations across customer tiers.

## **8. Business Recommendations**

## **Targeted Marketing**

* Design segment-specific campaigns, focusing loyalty rewards and personalized offers on ‘Loyal Champions’ (Cluster 2), the highest CLV group.
* Roll out win-back and reactivation initiatives for high-risk segments, especially lapsed (Cluster 0) customers.
* Utilize purchase frequency insights by incentivizing more frequent purchases via promotions, subscription offers, and periodic reminders.

## **Retention Workflows**

* Implement automated CRM triggers for high-value customer engagement dips—monitor recency increases or frequency drops.
* Establish high-touch retention programs for top segments, including personalized outreach and concierge services.

## **Personalized Growth Campaigns**

* Profile top clusters’ traits to identify and acquire lookalike prospects through targeted digital marketing.
* Tailor cross-selling and upselling offers based on product variety and past purchase patterns to deepen engagement.

## **Deployment & Monitoring**

* Integrate scoring and segmentation pipelines into CRM systems for real-time customer scoring and campaign orchestration.
* Establish dashboards tracking key performance metrics (retention rates, revenue uplift) by segment.
* Schedule regular model retraining (quarterly or post-major campaigns) to maintain predictive relevance.
* Use business feedback loops to continually refine models and marketing tactics.

## **9. Next Steps**

To operationalize these insights and maximize ROI:

* Deploy the segmentation and CLV prediction models into production.
* Educate marketing and CRM teams on segment characteristics and campaign tailoring.
* Initiate pilot campaigns on high-value and at-risk segments, measuring uplift carefully.
* Expand data collection to include behavioral, demographic, and engagement signals for future enhanced models.
* Foster ongoing collaboration between analytics and business teams to continuously align strategy with evolving customer behaviors.

## **10. Appendix & Technical Details**

* Full data preparation scripts, feature engineering steps, and modeling code are provided in the attached Jupyter Notebook.
* Visualizations documenting segmentation structure, feature importance, and model diagnostics accompany this report.
* The methodology emphasizes reproducibility, interpretability, and business alignment at every stage.

# **Final Remarks**

This project demonstrates a compelling, data-driven pathway from raw customer transactions to precise segmentation and lifetime value prediction, enabling targeted, personalized business strategies. The approach transparently combines technical acumen with business understanding to unlock enduring marketing advantages.

I invite you to review the attached notebook and discuss how these insights can be embedded into your marketing operations for immediate and sustainable business impact.