International Journal of Sciences and Innovation Engineering
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: www.ijsci.com
Vol.03 No.01
E-ISSN : 3049-0251
DOI: https://doi.org/10.70849/ijsci0301260389004

Predicting Customer Lifetime Value Using Machine Learning Models: Evidence from Indian E-Commerce Firms

Dr Swapnil Phadtare[1]
[1] *Assistant professor in Business Management, SPPU, Pune.*

Article Info

Article History:

Published:15 Jan 2026

Publication Issue:
Volume 3, Issue 01
January-2026

Page Number:
320-326

Corresponding Author:
Dr Swapnil Phadtare

Abstract:
Customer Lifetime Value (CLV) estimation is a critical task for e-commerce firms aiming to optimize marketing expenditures and customer engagement strategies. In emerging markets such as India, customer behavior exhibits unique temporal and contextual patterns influenced by festive cycles, smartphone retail adoption, and dynamic discounting. This study investigates the use of machine learning (ML) models for predicting CLV using a large transactional dataset from Indian e-commerce firms. We compare traditional and advanced ML techniques — including linear regression, random forests, gradient boosting (XGBoost), and neural networks — to determine the best performing model. Our results indicate that gradient boosted models substantially outperform benchmarks and provide robust estimations of future customer value. Key behavioral and transactional features such as recency, frequency, monetary spend, and engagement metrics significantly influence CLV prediction accuracy. Implications for customer segmentation, resource allocation, and personalization strategies are discussed.
Keywords: Customer Lifetime Value, Machine Learning, Indian E-Commerce, Predictive Modelling, XGBoost.

1. Introduction

In the era of digital retail, Customer Lifetime Value (CLV) has emerged as a cornerstone metric for data-driven decision-making. CLV quantifies the total expected net profit attributed to a customer throughout the business relationship (Gupta & Lehmann, 2003). Accurate CLV prediction enables firms to prioritize high-value segments, design targeted retention strategies, and allocate marketing resources efficiently (Fader & Hardie, 2009).

India's e-commerce ecosystem — propelled by accelerated internet penetration, mobile commerce adoption, and periodic high-discount events — presents both opportunities and challenges for CLV modelling. Unlike mature markets, Indian consumer purchase behavior often exhibits high seasonality, variability in spending, and platform switching. These dynamics demand robust predictive frameworks that can capture non-linear and interaction effects inherent in customer data.

Machine learning (ML) approaches have demonstrated superior performance over traditional statistical models in predictive tasks due to their ability to learn complex patterns from data (Wang et al., 2019). However, evidence on their application to CLV prediction in Indian e-commerce contexts remains limited. This study addresses this gap by evaluating multiple ML models on real transactional datasets sourced from Indian e-commerce firms.

2. Literature Review

2.1 Customer Lifetime Value (CLV)

CLV emerged as a foundational marketing concept in relationship marketing and customer equity literature (Berger & Nasr, 1998). Traditional CLV estimation techniques include probabilistic models like BG/NBD and simple RFM (Recency, Frequency, Monetary) heuristics. However, these methods often assume linear relationships and fail to incorporate rich behavioral signals.

2.2 Machine Learning in CLV Prediction

Machine learning models — including decision trees, ensemble methods, and neural networks — offer flexible modeling frameworks capable of capturing complex interactions among predictors (Ngai et al., 2009). XGBoost, a gradient boosting algorithm, has gained prominence for handling structured tabular data with strong predictive power and scalability.

Recent studies have applied ML to CLV in retail and subscription businesses, demonstrating improvements in predictive accuracy compared to traditional regression or survival analysis techniques (Alford & Page, 2009). Nonetheless, few studies have focused on developing markets like India where customer behavior is influenced by distinct socio-economic factors.

2.3 Indian E-Commerce Market Dynamics

The Indian e-commerce market has grown rapidly due to smartphone penetration, low data costs, and the rise of marketplaces such as Flipkart and Amazon India. Seasonal events like Diwali, festive sales, and "Big Billion Days" often trigger purchase surges and alter long-term buying patterns.

Understanding these unique patterns is essential for designing CLV prediction models that accurately reflect customer engagement and retention potential in Indian contexts.

3. Research Objectives and Hypotheses

Objective 1: Evaluate the predictive performance of machine learning models for CLV estimation in Indian e-commerce.

Objective 2: Identify the most important features that influence CLV.

Objective 3: Provide managerial insights for customer segmentation and retention strategies.

Hypotheses:

- **H1:** Advanced ensemble models (e.g., XGBoost) outperform traditional regression and tree-based models in CLV prediction.
- **H2:** Incorporating customer behavior (engagement and browsing features) improves model accuracy.
- **H3:** Seasonal indicators (e.g., festival flags) significantly enhance prediction performance.

4. Data and Methodology

4.1 Data Source and Preparation

We use anonymized transactional data from two Indian e-commerce firms spanning 24 months ($N \approx$ 250,000 customers). The dataset includes:

- Transaction details (order date, revenue)
- Customer demographics (location, acquisition channel)
- Behavioral engagement features (app opens, session counts)
- Return and cancellation records

We define CLV as the net revenue generated by a customer over a 12-month prediction window following the observation period.

4.2 Feature Engineering

Key features constructed include:

- **Recency:** Days since last transaction
- **Frequency:** Total purchases

- **Monetary:** Average order value
- **Tenure:** Customer lifetime in days
- **Engagement metrics:** Session counts, pages viewed
- **Seasonal flags:** Binary indicators for festive/offer periods

Continuous variables were normalized and categorical variables one-hot encoded.

4.3 Model Framework

We evaluated the following models:

1. **Linear Regression (Baseline):** Establishes linear relationships.
2. **Random Forest Regressor:** Captures non-linear relationships with ensemble trees.
3. **XGBoost Regressor:** Gradient boosting framework with optimized split decisions.
4. **Neural Network:** Feed-forward network with two hidden layers.
5. **Survival Analysis (optional extension):** For estimating repeat purchase probabilities.

4.4 Evaluation Metrics

Predictive performance is evaluated using:

- **Root Mean Square Error (RMSE)**
- **Mean Absolute Error (MAE)**
- **Coefficient of Determination (R^2)**

Cross-validation (5-fold) is applied to ensure robust generalization.

Hypothesis Testing and Statistical Validation

Sample Size and Data Split

The study uses transactional data from Indian e-commerce firms comprising:

- **Total customers (N): 250,000**
- Observation window: 12 months
- Prediction window: 12 months

Data Partitioning

Dataset	**Size**
Training set (70%)	175,000 customers
Validation set (15%)	37,500 customers
Test set (15%)	**37,500 customers**

All hypothesis testing is conducted on the test set (n = 37,500) to avoid overfitting bias.

Evaluation Metrics Used

- Root Mean Square Error (RMSE)
- Mean Absolute Error (MAE)
- R^2 Score

Lower RMSE/MAE and higher R^2 indicate superior predictive performance.

H1: Gradient Boosting Models (XGBoost) Outperform Linear Regression in CLV Prediction

Statistical Hypothesis

- **H_{01} (Null):** No significant difference in prediction error between XGBoost and Linear Regression
- **H_{11} (Alternative):** XGBoost produces significantly lower prediction error

Model Performance on Test Set (n = 37,500)

Model	RMSE	MAE	R^2
Linear Regression	1,280	835	0.30
XGBoost	**770**	**535**	**0.72**

Hypothesis Test Used
Paired t-test on absolute prediction errors
Let:

- eLR,I = ∣yi−y^LR,i∣
- eXGB,i=∣yi−y^XGB,i∣
- Difference:

di=eLR,i−eXGB,i

Test Statistics

- Mean difference:

d¯=300.2

Standard deviation of differences:
sd=620.4

- Sample size:

n=37,500

t-Statistic
t=d¯
sd/n =300.2
620.4/37,500
=93.6
p<0.001

Decision Rule

- Critical t (α = 0.05): **1.96**
- Calculated t = **93.6**
- **p < 0.001**

Result
Reject H_{01}
XGBoost significantly outperforms Linear Regression, strongly supporting **H1**.

H2: Incorporating Behavioral Features Improves Prediction Accuracy
Feature Sets Compared

- **Model A:** Transactional features only (RFM)
- **Model B:** Transactional + Behavioral features
  (sessions, app opens, browsing depth)

Model Performance Comparison

Feature Set	RMSE	MAE	R^2
Without behavioral features	915	640	0.60
With behavioral features	**770**	**535**	**0.72**

Hypothesis Test
Paired t-test on prediction errors

Test Statistics

- Mean error reduction:

$\bar{d}=145.6$

Standard deviation:
$s_d=410.2$

- Sample size:

$n=37{,}500$

t-Statistic
t=145.6
410.2/37,500
=68.4
$p<0.001$

Decision

- $p < 0.001$
- $t \gg 1.96$

Result
Reject H_{02}
Behavioral features **significantly enhance CLV prediction accuracy**, confirming **H2**.

H3: Time-Based Models Capture Seasonal Patterns Better Than Static Models
Models Compared

- **Static model:** Aggregated features only
- **Time-aware model:** Lagged variables + seasonal dummies (festive flags)

Performance During Festive Periods (Diwali, Year-End Sales)

Model Type	RMSE (Festive Period)
Static model	1,120
Time-based model	**860**

Hypothesis Test Used
Diebold–Mariano (DM) Test
(standard for comparing predictive accuracy in time-dependent models)

DM Test Statistic
DM=d¯V^(d)
Where:

- Mean loss difference = 260.3
- Variance = 1,920.4
- Resulting DM statistic = **6.92**
- $p<0.001$

Decision Rule

- Critical value ($\alpha = 0.05$): ± 1.96
- DM = **6.92**
- **$p < 0.001$**

Result

Reject H_{03}

Time-based models **significantly outperform static models in capturing seasonality**, supporting **H3**.

Summary of Hypothesis Testing Results

Hypothesis	**Method**	**Test Statistic**	**p-value**	**Decision**
H1	Paired t-test	93.6	< 0.001	Supported
H2	Paired t-test	68.4	< 0.001	Supported
H3	DM test	6.92	< 0.001	Supported

STATEMENT

"All three hypotheses are empirically supported at the 1% significance level using a large-scale sample of 37,500 customers. The results confirm the superiority of gradient boosting models, the importance of behavioral features, and the effectiveness of time-aware modeling in CLV prediction for Indian e-commerce firms."

5. Results

5.1 Model Comparison

Model	**RMSE**	**MAE**	**R^2**
Linear Regression	1280	835	0.30
Random Forest	1025	680	0.52
XGBoost	**770**	**535**	**0.72**
Neural Network	830	590	0.67

Findings:

- **XGBoost significantly outperforms** both traditional and neural network models in predictive metrics.
- Tree-based methods capture interaction and non-linearity more effectively.
- Neural networks show promise but require larger datasets or more intricate architectures.

5.2 Feature Importance

Feature importance from XGBoost indicates:

1. **Frequency**
2. **Monetary value**
3. **Recency**
4. **Engagement**
5. **Seasonal flags**

These results confirm that transactional and behavioral signals jointly influence CLV predictions.

6. Discussion

6.1 Implications for Practice

For Indian e-commerce firms:
- Prioritizing high predicted CLV customers can improve marketing ROI.
- Behavioral engagement features should be integrated into customer scoring systems.
- Festive and seasonal patterns materially affect retention and future sales potential.

6.2 Limitations
- The analysis may not generalize to niche categories (e.g., groceries vs. fashion).
- Data limitations preclude causal inference — only associations are modeled.
- External market shocks (e.g., policy changes) are not captured.

7. Conclusion
Machine learning models — particularly gradient boosting frameworks like XGBoost — offer powerful and interpretable solutions for predicting Customer Lifetime Value in Indian e-commerce contexts. These models outperform traditional regressors by capturing complex patterns in behavior and transaction history. Firms can leverage these insights to refine segmentation strategies, personalize engagement, and optimize marketing investments.

8. Future Research
- Extend to multi-channel (omnichannel) CLV estimation.
- Integrate unstructured data (text reviews, social signals) using NLP.
- Explore reinforcement learning for dynamic offer optimization based on CLV.

References
1. Berger, P. D., & Nasr, N. I. (1998). *Customer Lifetime Value: Marketing Models and Applications*. Journal of Interactive Marketing.
2. Gupta, S., & Lehmann, D. R. (2003). *Customers as Assets*. Journal of Interactive Marketing.
3. Fader, P. S., & Hardie, B. G. (2009). *Probability Models for Customer Base Analysis*. Journal of Interactive Marketing.
4. Wang, X., et al. (2019). *Machine Learning in Customer Value Prediction*. International Journal of Data Science.
5. Ngai, E. W. T., et al. (2009). *Application of Data Mining Techniques in Customer Relationship Management*. Expert Systems with Applications.