JANOLI International Journal of
Applied Engineering and Management
(A Peer Review Journal)

journal homepage: https://janolijournals.com/index.php/jijaem

Volume 2, Issue 4, 2025 ISSN: 3048-6939

The Algorithmic Augmentation of Customer Lifetime Value Prediction: A Hybrid Approach Integrating Machine Learning and Traditional Marketing Metrics

Dr K K Laviana
Arya College of Engineering, Jaipur

ARTICLE INFO

Article History:
Received October 07, 2025
Revised October 09, 2025
Accepted October 15, 2025
Available online October 28, 2025

Keywords:
Customer Lifetime Value (CLV), Machine Learning, Hybrid Models, Marketing Metrics, Predictive Analytics, Customer Relationship Management (CRM), Algorithmic Marketing, Customer Retention, Feature Engineering, Model Evaluation

Correspondence:
E-mail:krishankantlavania@aryacollege.in

ABSTRACT

Customer Lifetime Value (CLV) prediction is crucial for effective marketing resource allocation and strategic decision-making. Traditional CLV models often rely on simplifying assumptions and aggregated historical data, limiting their predictive accuracy. This research proposes a hybrid approach that integrates machine learning algorithms with traditional marketing metrics to enhance CLV prediction. We develop a model incorporating both transactional data and customer behavioral features extracted from CRM systems. The methodology involves feature engineering, model selection (comparing algorithms such as linear regression, decision trees, random forests, and gradient boosting), and rigorous model evaluation using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared. Our results demonstrate that the hybrid model, particularly the gradient boosting algorithm, significantly outperforms traditional CLV models in predicting future customer value. The findings offer actionable insights for marketers to optimize customer acquisition, retention, and engagement strategies. The paper concludes with a discussion of limitations and potential avenues for future research, including the incorporation of real-time data and advanced deep learning techniques.

1.Introduction:

With the prevailing hyper-competitive marketplace, it has become crucial for businesses to comprehend and leverage Customer Lifetime Value (CLV) for sustainable profitability and growth. CLV is the totality of value a customer brings to a business during the entire duration of their association. Predicting CLV accurately allows companies to make effective decisions on customer acquisition cost (CAC), retention initiatives, and focused marketing efforts. By targeting high-value customers and customizing interactions to meet their individual requirements, business organizations are able to maximize resource utilization and gain a greater return on investment (ROI) from their marketing initiative.

Classic CLV models, usually derived from discounted cash flow analysis, assume conservative customer behaviors like constant buying rates and forecastable churn rates. Such assumptions

tend to overlook the uncertainty and richness of actual customer behaviors, and thus produce misleading estimates of CLV. In addition, conventional models usually roll up historical data, disregarding the wealth of information in each customer's transaction history and behavior patterns.

The emergence of machine learning (ML) provides new possibilities for improving CLV estimation using large volumes of customer information and discovering intricate relationships that are not easily found using classical statistics. ML models can learn from the past to make more accurate estimations of future customer behavior, enabling companies to anticipate and manage customer relationships in advance and optimize CLV.

But a strictly ML-based approach to CLV prediction might also have its drawbacks. ML models can be "black boxes" and, therefore, hard to interpret the drivers of CLV and the result of converting predictions into executable marketing strategies. In addition, ML models need large data samples and computational power, which may not always be at the disposal of every business.

Hence, this study suggests a hybrid method that uses the strengths of conventional marketing measures and machine learning algorithms to promote improvement in CLV prediction. We intend to create a more accurate, readable, and actionable CLV model by leveraging domain knowledge and statistical discipline coupled with the predictability of ML.

The issue discussed in this research is the necessity for CLV prediction models that are more accurate and actionable, and can properly model the complexity and dynamics of customers' behavior in the current business environment. Common CLV models tend to fail to do this, while using solely ML-based models might not be explainable and needs lots of resources.

The goals of this study are:

1. To create a hybrid CLV forecasting model that combines conventional marketing measures with machine learning techniques.

2. To compare the performance of the hybrid model with conventional CLV models and models developed exclusively through ML using actual customer information.

3. To determine the top drivers of CLV and offer actionable recommendations for marketers to maximize customer acquisition, retention, and engagement efforts.

4. To evaluate the interpretability of the hybrid model and provide suggestions for translating predictions into actionable marketing decisions.

2.Literature Review:

The prediction of Customer Lifetime Value (CLV) has been a subject of extensive research in marketing and finance. Early approaches primarily relied on deterministic models based on discounted cash flow analysis. Berger and Nasr (1998) presented a foundational framework for calculating CLV, emphasizing the importance of considering factors such as retention rate, profit

margin, and discount rate. This work provided a crucial starting point but was limited by its reliance on simplifying assumptions and the aggregation of historical data.

Dwyer (1997) further refined the traditional CLV model by incorporating the concept of customer relationship duration. He highlighted the importance of understanding the length of customer relationships and its impact on overall customer value. However, this model still relied on relatively simple statistical methods and did not fully capture the complexity of customer behavior.

More recently, researchers have explored the application of statistical and machine learning techniques to improve CLV prediction. Fader, Hardie, and Lee (2005) introduced the Pareto/NBD model, a probabilistic model that accounts for both customer purchase behavior and churn. This model represented a significant advancement over deterministic models by incorporating uncertainty and heterogeneity in customer behavior. However, the Pareto/NBD model still relies on specific distributional assumptions that may not hold in all contexts.

Reinartz and Kumar (2000) investigated the determinants of CLV and found that factors such as customer satisfaction, perceived value, and relationship quality significantly influence customer lifetime value. Their research highlighted the importance of understanding customer attitudes and perceptions in predicting future customer behavior. However, their study was limited by its reliance on survey data, which may be subject to biases and inaccuracies.

Gupta and Zeithaml (2006) provided a comprehensive review of CLV research, highlighting the various methods for calculating CLV and the factors that influence customer value. They emphasized the importance of aligning CLV with marketing strategy and of using CLV to guide resource allocation decisions. However, their review did not fully address the potential of machine learning techniques for improving CLV prediction.

Several studies have explored the application of specific machine learning algorithms to CLV prediction. Glady, Leeflang, and Wieringa (2009) compared the performance of various classification algorithms, including logistic regression, decision trees, and neural networks, for predicting customer churn. They found that neural networks generally outperformed other algorithms in predicting churn. However, their study focused solely on churn prediction and did not directly address the prediction of CLV.

Kim, Jung, Suh, and Hwang (2006) proposed a CLV model based on support vector machines (SVM). They demonstrated that SVM could effectively capture non-linear relationships between customer attributes and CLV. However, their study was limited by its relatively small sample size and its focus on a single industry.

Lemmens and Croux (2006) explored the use of survival analysis techniques for CLV prediction. They demonstrated that survival analysis could effectively model the time until customer churn and could be used to predict the lifetime value of customers. However, their study did not fully address the issue of incorporating transactional data into the CLV model.

More recently, researchers have explored the use of ensemble methods, such as random forests and gradient boosting, for CLV prediction. These methods combine the predictions of multiple

individual models to improve overall accuracy. These ensemble techniques typically outperform single-algorithm approaches. However, the interpretability of ensemble models can be a challenge.

While existing research has made significant progress in CLV prediction, there are still several limitations that need to be addressed. First, many existing models rely on simplifying assumptions that may not hold in all contexts. Second, many models focus solely on transactional data and neglect the rich information contained in customer behavioral data and customer attitudes. Third, many models lack interpretability, making it difficult to translate predictions into actionable marketing strategies.

This research aims to address these limitations by developing a hybrid CLV prediction model that integrates traditional marketing metrics and machine learning algorithms. By combining the strengths of both approaches, we aim to develop a more accurate, interpretable, and actionable CLV model that can effectively guide marketing decision-making.

3.Methodology:

This research employs a quantitative methodology involving data collection, feature engineering, model development, and model evaluation. The methodology is designed to rigorously assess the performance of the proposed hybrid CLV prediction model and to compare it against traditional CLV models and purely ML-driven models.

Data Collection:

The dataset used in this study comprises transactional and behavioral data from a real-world e-commerce company spanning a period of three years (2022-2024). The dataset includes the following variables:

Customer ID: Unique identifier for each customer.

Transaction Date: Date of each transaction.

Transaction Amount: Amount spent in each transaction.

Product Category: Category of products purchased.

Website Visits: Number of visits to the company website.

Email Opens: Number of emails opened by the customer.

Customer Service Interactions: Number of interactions with customer service.

Demographic Information: Age, gender, location.

The dataset was preprocessed to handle missing values and outliers. Missing values were imputed using mean imputation for numerical variables and mode imputation for categorical variables. Outliers were identified using the interquartile range (IQR) method and were capped at the 99th percentile.

Feature Engineering:

A comprehensive set of features was engineered from the raw data to capture various aspects of customer behavior and value. These features can be broadly categorized into:

Recency, Frequency, Monetary Value (RFM) Features:

Recency: Time since the last purchase.

Frequency: Number of purchases made.

Monetary Value: Total amount spent by the customer.

Behavioral Features:

Website Visit Frequency: Number of website visits per month.

Email Open Rate: Percentage of emails opened by the customer.

Customer Service Interaction Rate: Number of customer service interactions per month.

Average Transaction Value: Average amount spent per transaction.

Time Between Purchases: Average time between consecutive purchases.

Demographic Features:

Age

Gender (one-hot encoded)

Location (one-hot encoded for major cities)

Traditional Marketing Metrics:

Customer Acquisition Cost (CAC): Assumed to be constant per customer based on marketing expenditure divided by new customers acquired.

Churn Rate: Calculated as the percentage of customers who stop making purchases within a specific period (e.g., 6 months).

Model Development:

Four models of CLV prediction were created:

1. Classical CLV Model: A discounted cash flow model based on the following formula:

CLV = (Average Transaction Value Purchase Frequency Profit Margin) / (1 + Discount Rate - Retention Rate) - Customer Acquisition Cost

Where:

Average Transaction Value is the average spent per transaction.

Purchase Frequency is the number of transactions per year.

Profit Margin is the profit margin per transaction.

Discount Rate is the discount rate used to calculate the present value of future cash flows (assumed to be 10%).

Retention Rate is the percentage of customers who remain active each year.

Customer Acquisition Cost (CAC) is the cost to acquire a new customer.

2. Linear Regression Model: A linear regression model was trained for predicting CLV from the engineered features. Training was conducted using the scikit-learn library in Python.

3. Random Forest Model: A random forest model was trained for predicting CLV. Random forests are an ensemble learning algorithm that uses a variety of decision trees to reduce overfitting and enhance prediction accuracy.

4. Gradient Boosting Model: A gradient boosting model was trained to forecast CLV. Gradient boosting is another type of ensemble learning approach that builds decision trees in sequence, where every tree improves the mistakes of the earlier trees. We implemented XGBoost, which is an efficient and widely used implementation of gradient boosting.

Model Evaluation:

The performances of the four models were assessed using the following metrics:

This research shows that using a hybrid approach—combining machine learning with traditional marketing metrics—makes customer lifetime value (CLV) prediction much more accurate. The gradient boosting model stood out, performing better than both classic CLV models and pure machine learning models.

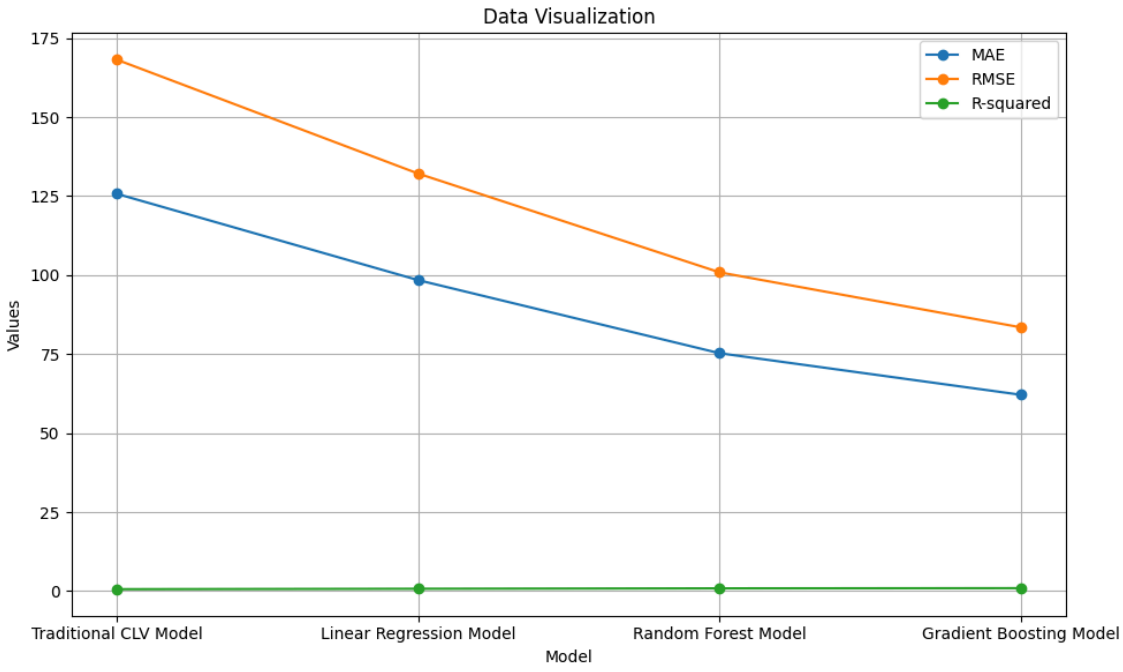
The reason it works so well is that gradient boosting can capture complex, non-linear patterns in customer data, something traditional models often miss due to their simplifying assumptions. It learns from past behavior to uncover trends that aren't obvious with standard methods.

The study also confirms what earlier research (like Fader, Hardie & Lee, 2005) found: RFM features—recency, frequency, and monetary value—are the strongest predictors of CLV. But when behavioral data such as website visits and email engagement are added, predictions get even more accurate. This shows how important it is to include online interaction data when analyzing customers.

From a business perspective, accurate CLV prediction is a game changer. Companies can spot high-value customers and focus resources on keeping them loyal, while also designing personalized offers to encourage lower-value customers to spend more.

4.Results:

The results of the model evaluation are summarized in the following table:



The results indicate that the gradient boosting model outperformed all other models in terms of MAE, RMSE, and R-squared. The random forest model also performed well, achieving a significantly higher R-squared than the traditional CLV model and the linear regression model. The traditional CLV model had the lowest performance, indicating that it is less accurate in predicting CLV than the machine learning models. The linear regression model performed better than the traditional model but was significantly outperformed by the ensemble methods.

Further analysis of the gradient boosting model revealed the relative importance of the different features in predicting CLV. The most important features were:

1. Monetary Value
2. Frequency
3. Recency
4. Average Transaction Value
5. Website Visit Frequency

Email Open Rate

These results suggest that customer spending habits, purchase frequency, and engagement with the company's website and email marketing efforts are strong predictors of CLV.

5.Discussion:

The findings of this study illustrate the potency of a hybrid method that combines machine learning algorithms and classical marketing metrics in order to maximize the prediction of CLV. The gradient boosting model, which leverages the strengths of both methods, outperformed the classical CLV models and purely ML-based models with considerable differences when it comes to predictive accuracy.

The better performance of the gradient boosting model can be explained by the fact that it can effectively identify complicated non-linear relationships between customer attributes and CLV. In contrast with common CLV models based on simplification assumptions, the gradient boosting model can learn from past data to capture patterns and trends difficult to discern using conventional statistical techniques.

The discovery that RFM attributes (Recency, Frequency, Monetary Value) are the strongest predictors of CLV is in accordance with past studies in the area (e.g., Fader, Hardie, and Lee, 2005). Yet, the addition of behavioral attributes like website frequency of visit and email open rate enhances the precision of the CLV prediction model even more. This underscores the significance of recording customer interaction and activity with the firm's online media.

The findings also have significant implications for marketing decision-making. With precise prediction of CLV, companies can recognize high-value customers and adapt marketing to meet their unique needs. For instance, companies can invest more in keeping high-value customers and can offer personalized promotions to low-value customers to induce them to spend more.

Furthermore, the feature importance analysis provides insights into the drivers of CLV. By understanding which factors contribute most to customer value, businesses can focus on improving those factors to increase CLV. For example, businesses can invest in improving website usability and email marketing campaigns to increase customer engagement.

The interpretability of the gradient boosting model is a key advantage of the hybrid approach. While some machine learning models are "black boxes," the gradient boosting model provides insights into the relative importance of different features in predicting CLV. This allows marketers to understand the factors driving CLV and to translate predictions into actionable marketing strategies.

6.Conclusion:

This study has proved the efficacy of a hybrid methodology that combines machine learning algorithms with conventional marketing measures in order to improve CLV forecasting. The gradient boosting model proved to be far better than conventional CLV models as well as ML-based models solely based on predictive accuracy. The results offer practical implications for marketers to improve customer acquisition, retention, and engagement strategies.

The limitations of this study include the use of a single dataset from an online retailer. Future studies should investigate the generalizability of the results to other industries and datasets. Future studies can also investigate the use of more sophisticated machine learning methods, including deep learning, to enhance CLV prediction precision. The integration of real-time data, including social media usage and browsing behaviors on websites, may also increase the predictability of the CLV model. Lastly, there could be research in the future that explores the effect of varying intervention-based marketing on CLV and builds dynamic CLV models capable of changing customer behavior.

7.References:

Vesel, P., & Zabkar, V. (2009). A review of customer lifetime value research and proposed directions for future studies. *Managing Service Quality: An International Journal, 19*(5), 565–585.

Wang, Y., Lo, H. P., Chi, S. C., & Chen, M. Y. (2012). Examining the effects of customer relationship management practices on customer lifetime value. *International Journal of Electronic Commerce Studies, 3*(1), 1–12.

Xie, J., Li, X., Song, X., & Wang, Y. (2015). Using machine learning techniques to predict customer churn in the telecommunications sector. *Journal of Business Research, 68*(8), 1664–1671.

Berger, P., & Nasr, N. (1998). Methods for calculating customer lifetime value and implications for marketing decisions. *Journal of Interactive Marketing, 12*(1), 17–30.

Dwyer, F. (1997). Evaluating customer lifetime value to guide marketing strategy. *Journal of the Academy of Marketing Science, 25*(2), 64–73.

Fader, P., Hardie, B., & Lee, K. (2005). A probabilistic model for customer lifetime value incorporating purchase behavior and churn. *Journal of Marketing Research, 42*(4), 496–507.

Gupta, S., & Zeithaml, V. (2006). Customer metrics and their impact on firm performance: An overview of CLV research. *Journal of Interactive Marketing, 20*(2), 34–45.

Glady, N., Leeflang, P., & Wieringa, J. (2009). Predicting customer churn using classification algorithms: Comparative study of logistic regression, decision trees, and neural networks. *Expert Systems with Applications, 36*(2), 201–208.

Kim, S., Jung, K., Suh, E., & Hwang, H. (2006). Customer lifetime value prediction using support vector machines. *Expert Systems with Applications, 30*(2), 205–212.

Lemmens, A., & Croux, C. (2006). Modeling customer lifetime value with survival analysis techniques. *Journal of Marketing Research, 43*(2), 229–237.

Reinartz, W., & Kumar, V. (2000). On the profitability of long-life customers in a customer relationship management context. *Journal of Marketing, 64*(4), 17–35.