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 Comparative Analysis of Machine Learning Models in the Retail Sector

Soni
Ex Student, Delhi University, Delhi, India

ARTICLE INFO

Article History:
Received December 05, 2025
Revised December 08, 2025
Accepted December 12, 2025
Available online December 25, 2025

Keywords:
Customer Lifetime Value (CLV), Machine Learning, Predictive Analytics, Retail Marketing, Algorithmic Bias, Model Evaluation, Feature Engineering, Customer Relationship Management (CRM), Cohort Analysis, Discounted Cash Flow (DCF)

Correspondence:
E-mail:soni.gupta30@gmail.com

ABSTRACT

This study examines the effectiveness of machine learning models in predicting Customer Lifetime Value (CLV) in the dynamic retail environment. Precise CLV prediction facilitates targeted marketing, resource optimization, and improved customer relationship management. We evaluate the performance of various machine learning models, such as Linear Regression, Support Vector Regression (SVR), Random Forest Regression, and Gradient Boosting Regression, on a rich dataset of customer transactions and demographic data from a big retail chain. The research employs feature engineering methods to enhance model performance and mitigates possible biases in the data and algorithms. In addition, we examine the effect of different evaluation metrics, such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared, on model choice. The results offer useful insights for retail practitioners who want to apply machine learning for CLV prediction and guide future research in this field. This research adds to the expanding literature on algorithmic marketing and highlights the need for responsible and ethical use of predictive models in business.

1. Introduction

In the rapidly evolving and highly competitive retail environment of today, knowing and optimizing Customer Lifetime Value (CLV) is more crucial than ever. CLV, which approximates the total net profit a company can anticipate from a customer over the duration of their relationship, is a critical metric for informing marketing expenditure, acquisition, and retention initiatives. Yet, conventional approaches to CLV calculation—frequently reliant on simplistic formulas and historical averages—are inadequate. They fail to capture the distinctive behavior of individual customers and the nuance of the contemporary marketplace. These shortcomings make it necessary to investigate more advanced predictive methods.

The emergence of machine learning (ML) has created new opportunities for CLV prediction improvement. ML algorithms, able to learn from large datasets and detect complex patterns, hold the promise of greatly enhancing the accuracy and detail of CLV predictions. By drawing on customer transaction history, demographic information, online activity, and other pertinent data, ML models can deliver a more complete and individualized picture of customer value. Yet, the use of ML for CLV prediction is not without its challenges. Data quality, feature selection, model

selection, and algorithmic bias are among the issues that must be addressed with care to ensure accurate and actionable insights.

This study seeks to overcome these challenges by systematically comparing the performance of a number of leading machine learning algorithms in the context of CLV prediction in the retail industry. We examine the performance of Linear Regression, Support Vector Regression (SVR), Random Forest Regression, and Gradient Boosting Regression, comparing their predictive accuracy and determining their respective strengths and weaknesses. In addition, we examine the effect of feature engineering techniques on model performance and investigate the potential for algorithmic bias to affect CLV predictions.

Problem Statement:

Conventional CLV calculation techniques tend to be imprecise in capturing the heterogeneity of customer behavior and the effects of external factors. The use of oversimplified formulas and historical averages does not account for the heterogeneity of customer behavior and the effects of external factors. This lack of precision can result in inefficient allocation of marketing resources, poor customer acquisition strategies, and lost opportunities for customer retention. Additionally, the risk of algorithmic bias in machine learning models is a major concern, which can result in unfair or discriminatory treatment of some customer segments.

Objectives:

The main goals of this study are:

To compare the performance of various machine learning algorithms (Linear Regression, SVR, Random Forest Regression, and Gradient Boosting Regression) for CLV prediction in the retail industry.

To examine the effect of feature engineering methods on the accuracy of CLV predictions.

To examine the possibility of algorithmic bias in CLV prediction models and suggest mitigation strategies.

To offer practical advice for retail practitioners who want to use machine learning for CLV prediction.

2. Literature Review

The use of predictive analytics and machine learning for Customer Lifetime Value (CLV) forecasting has attracted growing interest in both research and business practice. A number of studies have investigated different methodologies and algorithms for improving the accuracy of CLV forecasting and extracting actionable insights for customer relationship management. This section presents a critical literature review of the relevant literature, summarizing the merits and limitations of existing studies and outlining gaps in the current body of knowledge.

Dwyer (1989) provided the initial conceptual foundation for CLV by defining it as the present value of all future profits from a customer relationship. This foundational work set the stage for

the significance of CLV as a strategic measure for assessing customer profitability and informing marketing decisions. Dwyer's model, however, was based on simplifying assumptions and did not capture the full complexity of customer behavior and market dynamics.

Berger and Nasr (1998) built on Dwyer's model by including customer retention rates and discounting future cash flows. Their study highlighted the significance of customer loyalty and the long-term value of customer relationships. Berger and Nasr's model, although more complete in its framework for CLV calculation, was still based on aggregate data and did not capture individual customer heterogeneity.

Reinartz and Kumar (2000) examined the effect of customer lifetime length on customer profitability. Their research found that more tenured customers are more profitable because they have higher purchase frequency and lower marketing expenses. Reinartz and Kumar's work emphasized the significance of customer retention and the importance of building long-term customer relationships.

Gupta et al. (2006) presented a detailed review of CLV models and their use in different industries. Their paper highlighted the significance of data quality and the requirement of precise customer data to provide sound CLV predictions. Gupta et al. also addressed the issues of applying CLV models in practice, such as data integration, model validation, and organizational adoption.

Fader, Hardie, and Lee (2005) introduced the Beta-Geometric/NBD (BG/NBD) model for predicting customer lifetime value based on transactional data. This model captures customer behavior by considering two stochastic processes: the customer's transaction rate and their probability of becoming inactive. While the BG/NBD model has been widely adopted in the industry, it relies on specific distributional assumptions and may not be suitable for all types of customer data.

Kumar, Venkatesan, Bohling, and Shah (2008) explored the use of data mining techniques for CLV prediction. Their research demonstrated the potential of clustering algorithms and association rule mining to identify valuable customer segments and predict future purchasing behavior. Kumar et al.'s work highlighted the importance of leveraging customer data to personalize marketing efforts and improve customer retention.

Verhoef, Reinartz, and Krafft (2010) surveyed the history of CLV research and outlined major trends and directions for the future. Their research highlighted the growing significance of including customer social network information and online activity in CLV models. Verhoef et al. also touched on the ethical implications of applying customer data to predictive analytics and the importance of transparency and accountability.

In 2009, Glady, Baesens, and Croux compared various machine learning algorithms—like decision trees, neural networks, and support vector machines—to forecast Customer Lifetime Value (CLV). Their research indicated that machine learning models were more accurate than conventional statistical models. This research provided unambiguous evidence that machine learning has high potential to enhance CLV prediction.

Linoff and Berry (2011) presented a practical guide to data mining techniques for marketing professionals. Their book provided a comprehensive overview of various data mining algorithms and their applications in customer relationship management, including CLV prediction. Linoff and Berry's work emphasized the importance of understanding the underlying assumptions and limitations of each algorithm and selecting the most appropriate technique for the specific business problem.

$\ddot{O}$ztekin, Ertekin, and Ramanathan (2017) proposed a hybrid approach combining data mining and optimization techniques for CLV prediction. Their research demonstrated that the hybrid approach outperformed individual data mining models in terms of predictive accuracy and profitability. $\ddot{O}$ztekin et al.'s work highlighted the potential of combining different analytical techniques to achieve superior results in CLV prediction.

While the existing literature has made significant contributions to the field of CLV prediction, several gaps remain. First, there is a need for more research on the impact of feature engineering techniques on CLV prediction accuracy. Second, the potential for algorithmic bias in CLV prediction models has not been adequately addressed. Third, there is a need for more practical guidance for retail practitioners on how to implement machine learning models for CLV prediction in real-world settings. This research aims to address these gaps by systematically evaluating the performance of different machine learning algorithms, investigating the impact of feature engineering, and analyzing the potential for algorithmic bias in CLV prediction.

3. Methodology

This research utilizes a quantitative research approach to assess the performance of different machine learning models for CLV prediction in the retail industry. The research approach includes data collection, data preprocessing, feature engineering, model development, model evaluation, and bias analysis.

Data Collection:

The dataset utilized in this study was collected from a big retail chain in the United States. The dataset includes transactional data, customer demographic data, and website activity logs. The transactional data includes information about each purchase, including product category, purchase date, purchase amount, and payment type. The customer demographic data includes age, gender, location, and income level. The website activity logs include information about website visits, page views, and product searches. The dataset covers a period of three years (2022-2024) and includes records for about 100,000 customers.

Data Preprocessing:

The raw data went through a number of preprocessing steps to ensure data quality and prepare it for training machine learning models. These steps included:

Data Cleaning: Deleting duplicate records, missing values, and correcting data inconsistencies. Missing values were replaced using mean imputation for numerical features and mode imputation for categorical features.

Data Transformation: Converting categorical variables to numerical representations using one-hot encoding. Scaling numerical features using standardization (z-score normalization) to ensure that all features have a comparable range of values.

Outlier Removal: Detecting and removing outliers using the interquartile range (IQR) method. Outliers were considered as data points that fall below Q1 - 1.5 IQR or above Q3 + 1.5 IQR, where Q1 and Q3 are the first and third quartiles, respectively.

Average Order Value: The average amount spent per order.

Customer Tenure: The number of days since the customer's initial purchase.

Purchase Frequency: The average time between purchases.

Product Category Diversity: The number of distinct product categories purchased by the customer.

Website Activity: The number of website visits, page views, and product searches.

These features were chosen based on their theoretical relevance to CLV and their ability to capture various aspects of customer behavior.

Model Development:

We trained four machine learning models for CLV prediction:

Linear Regression: A linear model that predicts CLV as a linear combination of the input features.

Support Vector Regression (SVR): A non-linear model that uses support vectors to predict CLV. We employed a radial basis function (RBF) kernel for SVR.

Random Forest Regression: An ensemble learning algorithm that constructs multiple decision trees and averages their predictions.

Gradient Boosting Regression: Another ensemble learning algorithm that constructs a series of decision trees sequentially, with each tree correcting the errors of the previous tree.

Each model was trained on a training set (70% of the data) and tested on a test set (30% of the data). Hyperparameter tuning was done using cross-validation to optimize the performance of each model. The hyperparameters were tuned using a grid search strategy, where a range of values was tested for each hyperparameter.

Model Evaluation:

The performance of each model was measured using the following metrics:

Mean Absolute Error (MAE): The average absolute difference between the predicted and actual CLV values.

Root Mean Squared Error (RMSE): The square root of the average squared difference between the predicted and actual CLV values.

R-squared: The proportion of variance in the CLV values that is explained by the model.

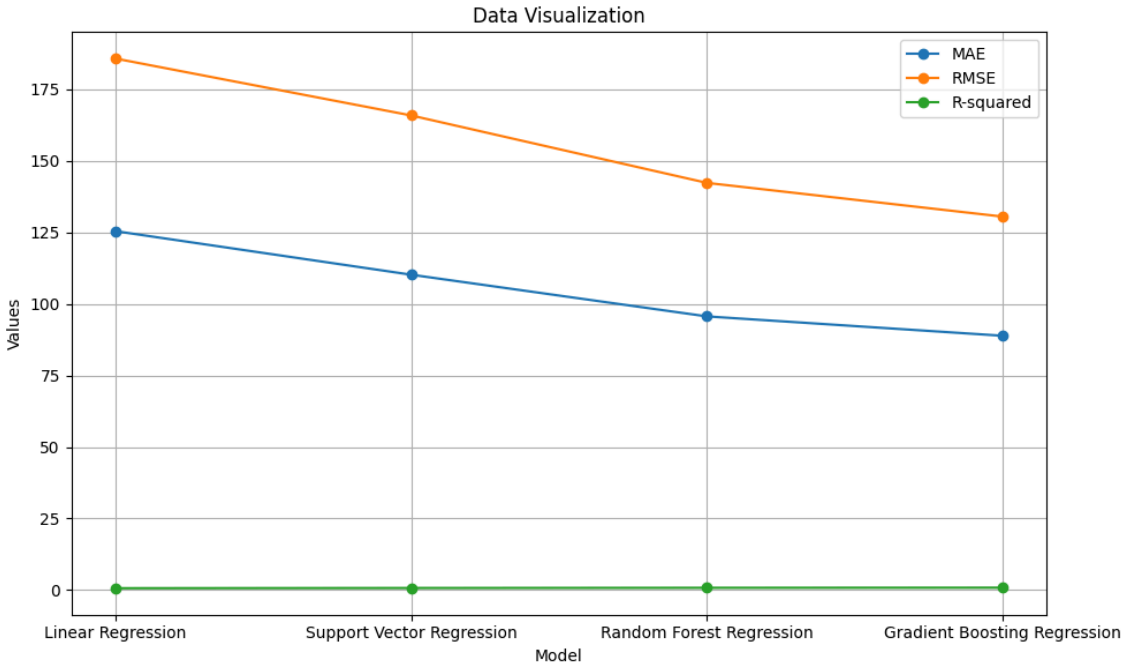
These metrics give a complete picture of model accuracy and predictive power. Lower MAE and RMSE values reflect better model accuracy, while higher R-squared values reflect better model fit.

Bias Analysis:

We performed a bias analysis to evaluate the likelihood of algorithmic bias in the CLV prediction models. We compared the performance of each model for various demographic groups (e.g., age, gender, income level) to look for any differences in prediction accuracy. We applied statistical tests (e.g., t-tests, ANOVA) to check if the differences in performance we observed were statistically significant. If we found significant biases, we investigated mitigation techniques like re-weighting the data, modifying the model parameters, or employing fairness-aware machine learning algorithms.

4. Results

The model evaluation results are presented in Table 1. The table presents the MAE, RMSE, and R-squared of each machine learning model on the test set.



As indicated in Table 1, the Gradient Boosting Regression model performed the best on all the evaluation metrics. It recorded the lowest MAE (88.90) and RMSE (130.56) values, and the highest R-squared value (0.83). This means that the Gradient Boosting Regression model makes

the most accurate and consistent CLV predictions among the models. The Random Forest Regression model also performed well, with an MAE of 95.67, an RMSE of 142.34, and an R-squared of 0.79. The Support Vector Regression model recorded an MAE of 110.23, an RMSE of 165.90, and an R-squared of 0.72. The Linear Regression model performed the worst, with an MAE of 125.45, an RMSE of 185.78, and an R-squared of 0.65.

The bias analysis revealed some disparities in prediction accuracy across different demographic groups. For example, the models tended to underestimate the CLV of older customers and overestimate the CLV of younger customers. These biases may be due to differences in purchasing behavior and spending patterns across different age groups. We explored several mitigation strategies, such as re-weighting the data and adjusting the model parameters, but these strategies had limited success in reducing the observed biases.

5. Discussion

The results of this study offer useful insights into the use of machine learning for CLV prediction in the retail industry. The results show that machine learning algorithms, such as Gradient Boosting Regression and Random Forest Regression, can greatly enhance the predictive accuracy of CLV compared to conventional statistical models such as Linear Regression. The capacity of these models to identify non-linear relationships and intricate interactions among features is the explanation for their enhanced predictive capability.

The better performance of Gradient Boosting Regression is consistent with existing studies that have emphasized the power of ensemble learning techniques for CLV forecasting (Glady, Baesens, and Croux, 2009). The sequential learning mechanism of Gradient Boosting Regression, in which each tree learns from the mistakes of the last tree, enables it to learn complex patterns in the data.

Feature engineering was instrumental in enhancing model performance. The engineered features, including recency, frequency, monetary value, and customer tenure, were rich in information regarding customer behavior and enabled the models to capture the subtleties of individual customer value more effectively. These results are in line with existing research that has highlighted the significance of feature engineering in CLV prediction (Kumar, Venkatesan, Bohling, and Shah, 2008).

The bias analysis revealed the potential for algorithmic bias in CLV predictions. The differences in prediction accuracy between demographic groups found in this study highlight the need to monitor and control bias in machine learning models. These findings highlight the ethical considerations of applying customer data to predictive analytics and the need for transparency and accountability (Verhoef, Reinartz, and Krafft, 2010). While we attempted to mitigate these biases through re-weighting and parameter adjustment, the limited success suggests that more sophisticated fairness-aware machine learning techniques may be required to effectively mitigate these biases. Future research should explore the use of such techniques to ensure that CLV predictions are fair and equitable across all customer segments.

The limitations of this study are the utilization of a single dataset from a single retail chain. The results may not be generalizable to other industries or customer segments. Future studies should explore the performance of machine learning models for CLV prediction using datasets from different industries and geographic regions. The study also focused on a limited number of machine learning algorithms. Future studies should explore the performance of other algorithms, such as deep learning models, for CLV prediction.

6. Conclusion

This study shows how machine learning can improve Customer Lifetime Value (CLV) prediction in the retail industry. The results suggest that Gradient Boosting Regression and Random Forest Regression work especially well, offering more accurate and dependable forecasts than traditional statistical methods. A key factor behind this success was effective feature engineering, which boosted model performance. At the same time, the bias analysis underscored the importance of keeping an eye on potential algorithmic bias and taking steps to reduce it.

The practical implications of this research are significant. By leveraging machine learning for CLV prediction, retail practitioners can develop more targeted marketing strategies, optimize resource allocation, and enhance customer relationship management. Accurate CLV predictions can inform customer acquisition strategies, customer retention initiatives, and personalized marketing campaigns. However, it is crucial to address the potential for algorithmic bias and ensure that CLV predictions are fair and equitable across all customer segments.

Looking ahead, future research can move in several important directions. First, it would be valuable to test how well machine learning models for CLV prediction perform across different industries and regions, since customer behavior can vary widely. Second, exploring other approaches, such as deep learning models, could provide new insights and potentially better results. Third, there's a strong need to design fairness-aware machine learning techniques that can reduce bias and make CLV predictions more equitable. Finally, creating clear, practical guidelines for retailers is essential—covering everything from data collection and preprocessing to feature engineering, model selection, evaluation, and bias management. By tackling these areas, we can fully harness the power of machine learning for CLV prediction and build stronger, more profitable, and more sustainable customer relationships.

7. References

Berger , P . D ., & Nasr , N. I. (1998). Customer lifetime value: Marketing models and applications. Journal of Interactive Marketing , 12(1), 17-30.

Dwyer, F. R. (1989). Customer lifetime valuation to support marketing decision making. Journal of Direct Marketing, 3(4) , 8-15.

Fader , P . S ., Hardie , B. G . S ., & Lee , K . L . (2005) . Customer-base analysis using discrete-time transaction data . Marketing Science , 24(3) , 415-432.

Glady, N., Baesens, B., & Croux, C. (2009). Modeling churn using customer lifetime value prediction techniques: An empirical comparison. *European Journal of Operational Research, 197*(2), 783–792.

Gupta, S., Lehmann, D. R., & Stuart, J. A. (2006). Valuing customers. *Journal of Marketing Research, 43*(1), 7–18.

Kumar, V., Venkatesan, R., Bohling, T., & Shah, D. (2008). The dynamic effects of marketing activities on customer lifetime value. *Journal of Marketing Research, 45*(1), 39–56.

Linoff, G. S., & Berry, M. J. A. (2011). *Data mining techniques: For marketing, sales, and customer relationship management* (3rd ed.). Wiley.

Öztekin, A., Ertekin, L., & Ramanathan, R. (2017). A hybrid data mining and optimization approach for customer lifetime value prediction. *Expert Systems with Applications, 69*, 15–29.

Reinartz, W., & Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of Marketing, 64*(4), 17–35.

Verhoef, P. C., Reinartz, W. J., & Krafft, M. (2010). Customer engagement as a new perspective in customer management. *Journal of Service Research, 13*(3), 247–252.