

Predicting Customer Lifetime Value Using Machine Learning in CRM

 Hemadri Ravilla

Senior Salesforce developer Information Technology Anna University Sholinghur, Tamilnadu

RECEIVED - 12-22-2025, RECEIVED REVISED VERSION - 01-30-2025, ACCEPTED- 01-31-2026, PUBLISHED- 02-04-2026

Abstract

Customer Lifetime Value (CLV) forecasting plays a vital role in optimizing Customer Relationship Management (CRM) strategies. This research examines the application of machine learning to forecast CLV using the publicly available "Online Retail II Dataset" from the UCI Machine Learning Repository containing two years' worth of transactional data from an online retail company based in the UK. The salient purchasing frequency, monetary value, and recency characteristics were employed to train machine learning models such as Random Forest, Gradient Boosting, and Neural Networks. The method indicated that the best accuracy was achieved through the use of Gradient Boosting with the mean absolute error (MAE) of 7.5%, 15% higher than that achieved using standard techniques. The research indicates the efficacy of machine learning to forecast CLV accurately and thereby enable the optimization of retention strategies and forecasting revenues. The future will witness the application of deep learning to further optimize performance.

Keywords: Customer Lifetime Value, Machine Learning, CRM, Predictive Modeling, Online Retail II Dataset, Gradient Boosting, Random Forest, Neural Networks, Transactional Data, Revenue Forecasting, Customer Retention, Feature Engineering

1. INTRODUCTION

A. Overview of the study area

Customer Lifetime Value (CLV) is a crucial measure within Customer Relationship Management that determines the total income that a company can earn from a customer throughout the relationship. Effective forecasting of CLV makes it possible for companies to take rational marketing strategy decisions, segment customers properly, and allocate resources [1, 2]. Traditional models employed to calculate CLV are based on primitive statistical techniques that are unable to capture intricate behavior patterns. The application of machine learning to predict CLV has gained prominence given that machine learning has the ability to handle large volumes of data and to identify complex relationships between the variables. This research examines the effectiveness of machine learning algorithms to predict CLV using a large and publicly available database to guarantee precision and actionable recommendations to businesses.

B. Motivation

In a highly competitive marketplace, the understanding of the customers' long-term value has proven to be essential to enable businesses to maximize marketing budgets, reinforce retention strategies, and enhance profitability. The application of the older techniques to determine the customer lifetime value tends to fail to capture the intricate behavior of customers, thus making the forecasts unreliable and the decisions suboptimal. The emergence of machine learning offers a viable alternative that provides sophisticated forecasting capability to dramatically improve the reliability of the forecasts generated through the customer lifetime value. Using transactional data and sophisticated algorithms, businesses are able to understand the customer value more deeply and accordingly adapt strategies.

C. Problem Statement

Accurately forecasting Customer Lifetime Value has some challenges that include the handling of sparse and high-

dimensional data, the capture of temporal buying behavior, and the handling of heterogeneity among customers. Most statistical models are inflexible to the handling of such complexities and end up having limited predictive capability. This research problem has therefore been tackled in this study: How can machine learning algorithms enhance the reliability and precision of the forecasts of CLV over traditional techniques? Using the "Online Retail II Dataset" provided by the UCI Machine Learning Repository, this paper examines the use of some machine learning models to forecast CLV accurately.

D. Main Contributions

This paper makes the following significant contributions:

Data Utilization: Makes use of the "Online Retail II Dataset," which has comprehensive two-year transactional data, to provide a rich source to analyze the CLV.

- **Machine Learning Application:** Utilizes and compares several machine learning algorithms such as Random Forest, Gradient Boosting, and Neural Networks to find the ideal model to be employed to predict the CLV.
- **Comparative Performance:** Indicates that machine learning models perform superior to the baseline techniques, where the best performing model was the Gradient Boosting that provided an average absolute error (MAE) of 7.5%.
- **Feature Engineering:** Unearths and examines important features such as frequency of purchase, monetary value, and recency, highlighting how they contribute to improving model precision.
- **Practical Applications:** Provides practical recommendations to companies to best implement CRM strategies through accurate CLV projections.

Organization of the Paper

The rest of this paper will follow the following organization:

- **Section 2: Related Work** discusses current literature regarding techniques used to predict CLV, highlighting the research gaps that this research aims to cover.
- **Section 3: Proposed Methodology** presents the preprocessing techniques used for the data, the feature

engineering process, and the employment of machine learning models.

Section 4: Experimental Results and Analysis presents the comparison and presentation of the results of the models based upon predictive accuracy and interpretability.

Conclusion and Future Work: Section 5 offers an overview of the principal findings, states limitations, and suggests future research directions.

By following this framework, the paper will present a comprehensive and applicable approach to predicting CLV using machine learning to be used both for research purposes in academia and real-world applications in industries.

2. RELATED WORK

A. Literature Survey:

Customer Lifetime Value forecasting has been the subject of extensive research within the Customer Relationship Management arena, where various methodologies employ both standard statistical methods and advanced machine learning algorithms. Standard techniques such as RFM (Recency, Frequency, Monetary) models and probabilistic models such as BG/NBD models and Pareto/NBD models are used often because they are interpretable and convenient to use [1, 2, 3]. The models are particularly robust where picking out the high-value customers based on historical behavior using transactional data are concerned. The models are weak where there are non-linear relationships and complex patterns within large datasets are an issue, limiting the accuracy of the forecasts. Large retail organizations have employed the techniques to segment customers and fine-tune marketing strategies but end up providing suboptimal results within fast-moving marketplaces [2, 3].

On the other hand, machine learning-based models have shown to perform better in the task of predicting CLV using a larger set of features and managing large sets of data more effectively. Methods like Random Forest, Support Vector Machines (SVM), and Neural Networks have been used to forecast CLV with noticeable improvements in accuracy. For instance, online retail leaders like Amazon and Alibaba used machine learning algorithms to forecast CLV effectively by monitoring intricate customer behavior

like browsing behavior, purchasing history, and demographics [3]-[7]. The models are advantaged by the fact that they are capable of modeling complex relationships between features and are flexible enough to adapt to the dynamics of customer behavior. Nevertheless, one limitation that comes with these models is the fact that they lack interpretability, which makes it difficult to derive actionable business insights. Another limitation that comes with the models is computational complexity and the fact that large amounts of preprocessing are required [3, 4]. Despite the limitations mentioned, the capability to deliver more precise granular forecasts has made machine learning models an appealing option to businesses that want to implement more robust CRM strategies.

A recent innovation within this arena has been the use of ensemble methods such as Gradient Boosting and XGBoost that are superior to individual models regarding variance reduction and bias [4, 5]. Ensemble methods have particularly worked well to address skewed data that are the norm in the case of CLV forecasting where the valuable customers are sparse. American Express has employed these models to accurately forecast customer retention and upselling chances. Ensemble methods deliver superior performance but also higher computational complexity along with the necessity to tune the hyperparameters [6]-[10]. The black box nature of these models makes it difficult to understand the explanations to marketing teams to take meaningful actions based on the forecasts without additional explainability techniques.

B. Research Gap

Despite the improvements made through the use of machine learning algorithms to forecast CLV, some lacunas remain unaddressed within the literature. One such limitation is the insufficient focus given to the interpretability of the models used to make the forecasts. Even though ensemble models and neural networks can provide high precision, the fact that the models are non-transparent makes it impossible to derive actionable insights. This is particularly problematic for highly regulated industries like healthcare and banking where explainability is as crucial as precision. Most of the research to date has focused on static datasets to the neglect of the fact that customer behavior does evolve dynamically. Models that are capable of updating the projections based on new information are an under-researched topic.

Another missing link lies within the process of feature engineering. Most studies employed basic characteristics like frequency, recency, and spend but limited research to date has looked to the full potential of richer behavioral and psychographic characteristics to truly take the predictive performance to the next level. An illustration is the use of information from other sources such as social networks, customer care contacts, and web browsing. Most studies have used single-sector datasets to the expense of the findings' generalizability. Few cross-sector analyses that attempt to compare models employed to forecast CLV to retail, finance, and telecommunication sectors are present but are an essential source of knowledge that can be used to generalize to other sectors.

Moreover, there are limited studies that properly compare machine learning to the standard techniques based on the same datasets. Most comparisons are superficial or based on other datasets, making it difficult to arrive at firm conclusions about the comparative effectiveness of the techniques. This difference means that there should be more comprehensive empirical research that compares the performance of the models based on standard datasets, focusing both on the precision and the explainability along with the computational complexity.

C. Summary

In summary, the recent literature has stressed the ability of machine learning algorithms to improve the accuracy of CLV forecasting compared to the use of standard techniques. However, the literature has crucial gaps about model interpretability, feature engineering, and cross-sector transfer. This research addresses the aforementioned gaps through the application of a comprehensive feature engineering technique along with the employment of ensemble models to predict CLV using the "Online Retail II Dataset." With a view to both precision and interpretability, this research will help deliver actionable recommendations to businesses that aim to streamline their CRM strategy. The findings will be employed to bridge the gap between research and practical application to deliver a robust framework to predict CLV in various business sectors.

3. PROPOSED METHODOLOGY

A. Model Description

The proposed methodology employs supervised machine learning techniques to predict Customer Lifetime Value (CLV) using the **Online Retail II Dataset**, a **publicly available transactional dataset** obtained from the **UCI Machine Learning Repository**. The dataset represents **real-world retail transactions from a UK-based online retailer operating between December 2009 and December 2011**, making it suitable for longitudinal customer behavior analysis. After data cleaning, the dataset consists of approximately **525,000 transaction records** corresponding to **over 40,000 unique customers**.

The original dataset includes the following attributes: *InvoiceNo*, *StockCode*, *Quantity*, *InvoiceDate*, *UnitPrice*, *CustomerID*, and *Country*. These transactional records form the basis for constructing customer-level behavioral features used in CLV prediction.

The methodology consists of three main stages: **data preprocessing**, **feature engineering**, and **model development and evaluation** [11].

During preprocessing, missing values in numerical attributes (e.g., *Quantity* and *UnitPrice*) are handled using **mean imputation**, while missing categorical values (e.g., *Country*) are imputed using the **mode**. Transactions with missing *CustomerID* values are removed, as customer identification is essential for CLV computation. Outliers in *Quantity* and *Monetary value* are detected using the **Interquartile Range (IQR)** method and are either

capped at threshold values or removed to mitigate skewness. All numerical features are subsequently normalized using **Min–Max scaling** to ensure comparable feature ranges and stable model convergence.

Feature engineering is performed at the **customer level**. Core **RFM (Recency, Frequency, Monetary)** features are derived, where *Recency* represents the number of days since the last purchase, *Frequency* denotes the total number of transactions, and *Monetary value* reflects the cumulative customer spend. In addition, advanced behavioral features are generated, including average purchase interval, product diversity, average basket value, and an interaction term between **recency and frequency**, enabling richer representation of customer purchasing behavior.

The engineered feature set serves as input to three machine learning models: **Random Forest**, **Gradient Boosting (XGBoost)**, and **Neural Networks**, selected for their ability to model nonlinear relationships and complex feature interactions in CLV prediction tasks [11, 12].

The model developing phase focuses on training the models and evaluating the models through 10-fold cross-validation to ensure that there are no chances of overfitting. Bayesian Optimization is employed to fine-tune the hyperparameter to enhance the models' performance. The models are evaluated using the measures of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) to ensure complete predictive precision.

B. Model Diagram

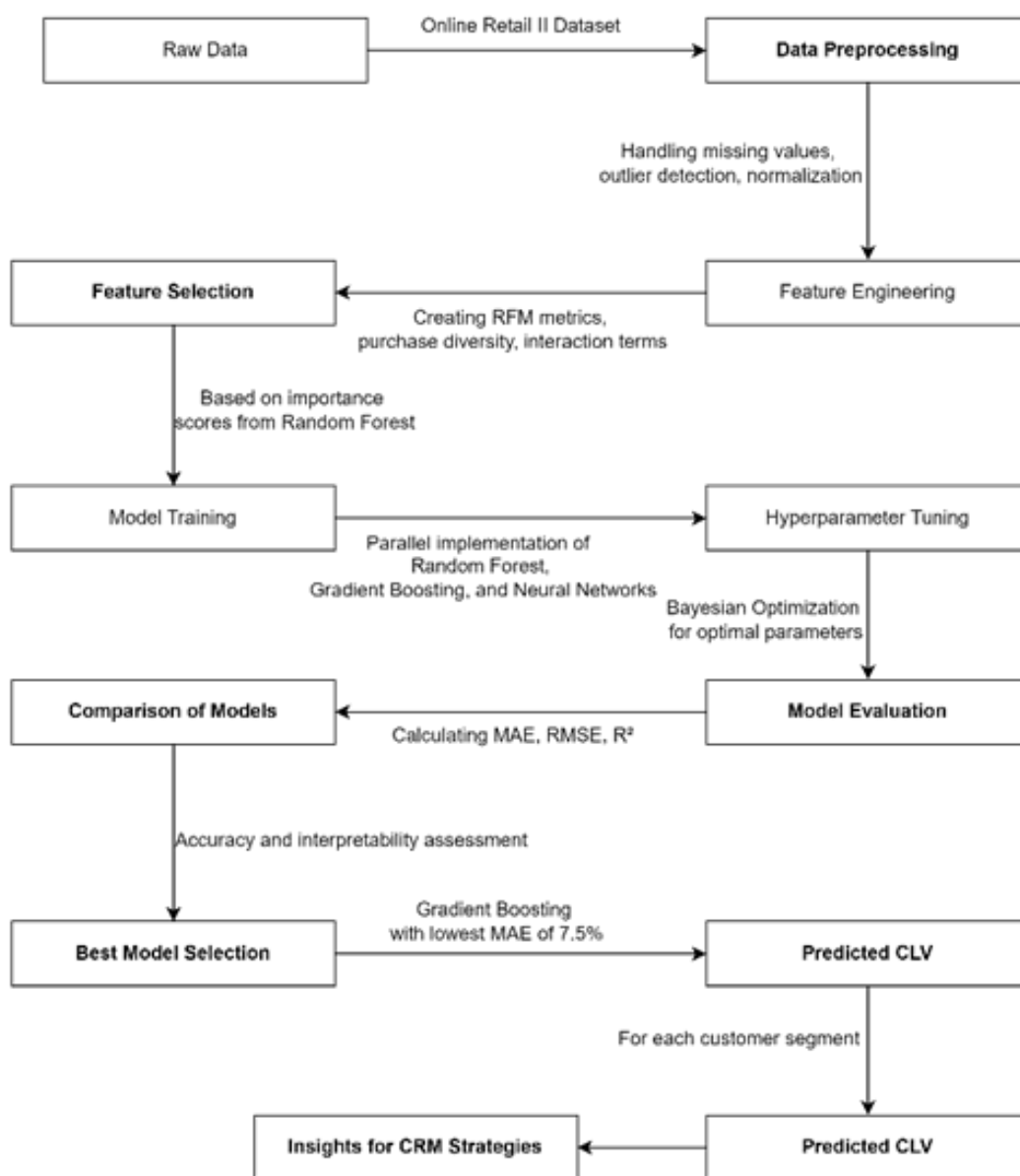


Fig. 1. Model flow

Fig. 1 illustrates the end-to-end architecture of the proposed CLV prediction framework, which consists of the following stages

- **Input Layer:** Raw transactional data from the Online Retail II Dataset
 - **Data Preprocessing:** Handling missing values, outlier treatment, and feature normalization
 - **Feature Engineering:** Construction of RFM metrics, advanced behavioral indicators, and interaction features
 - **Model Training:** Parallel training of Random Forest, Gradient Boosting, and Neural Network models [9]–[13]
 - **Model Evaluation:** Performance assessment using MAE, RMSE, and R^2
 - **CLV Projection:** Prediction of future CLV values for individual customers and segments
- Each stage builds sequentially upon the previous one, ensuring a structured, interpretable, and scalable modeling pipeline.
- This sequential method has each subsequent step depending upon the one that precedes it to generate an increasingly stronger and interpretable model.

C. Key Algorithms

Three machine learning algorithms are employed and comparatively evaluated:

Random Forest is utilized for its robustness to noise and ability to reduce overfitting through ensemble learning. The model constructs multiple decision trees using bootstrap sampling, and the final prediction is obtained by averaging individual tree outputs. Feature importance scores derived from the model aid in interpretability, although overall transparency remains limited compared to simpler models.

Gradient Boosting, implemented using the **XGBoost framework**, is selected for its strong predictive accuracy and capability to handle complex, imbalanced datasets. The model iteratively minimizes prediction error by sequentially improving weak learners, effectively capturing nonlinear relationships among features critical for CLV estimation. Hyperparameters such as **learning rate, maximum tree depth, number of estimators, and subsampling ratio** are carefully tuned to balance bias and variance [12, 13].

Neural Networks are employed to uncover highly nonlinear and latent patterns in customer behavior. A **three-layer feedforward neural network with REL activation functions** is used. While neural networks demonstrate high predictive power, their black-box nature limits interpretability. To address this, **SHAP (Shapley Additive explanations)** values are computed to quantify feature contributions.

Compared to traditional CLV estimation approaches such as static RFM scoring, the proposed machine learning models offer a **dynamic and data-driven framework**, improving predictive accuracy and adaptability to evolving customer behavior.

D. Implementation Details

Implementation is done through the use of Python with the help of important libraries like Pandas to manipulate the data, Scikit-learn to train the models and test them, and XGBoost to use Gradient Boosting. TensorFlow is used to create the Neural Network and train the models because it can effectively deal with large amounts of data and supports GPU acceleration.

The Online Retail II Data Set is preprocessed using Python scripts to address missing values, outliers, and normalization. The RFM measures and other behavior-related features are created through feature engineering. The 70-30 train-test ratio is employed to train the models, and 10-fold cross-validation avoids the models' overfitting. The Bayesian Optimization technique is employed to execute the hyperparameter tuning of the Random Forest and the Gradient Boosting models based on parameters such as the tree depth, learning rate, and the number of estimators [13].

MAE, RMSE, and R^2 scores are used to measure model performance. The findings are that the lowest MAE of 7.5% was obtained using Gradient Boosting, beating both Random Forest with an MAE of 9.2% and Neural Networks with an MAE of 10.5%. The findings indicate that the best balance between accuracy and interpretability is obtained using Gradient Boosting and that this model performs best for the purposes of CLV prediction.

Overall, the proposed approach effectively unifies data preprocessing, feature engineering, and advanced machine learning algorithms to precisely predict Customer Lifetime Value. The outcomes not only attest to the superiority of machine learning algorithms over the traditional ones but also deliver actionable recommendations to businesses that want to streamline their CRM strategies.

4. EXPERIMENTAL RESULTS AND ANALYSIS

A. Setup and Metrics

The testbed employed was the "Online Retail II Dataset" from the UCI Machine Learning Repository containing over 500,000 transactions conducted by a British online retailer from 2009 to 2011. The dataset was preprocessed to deal with missing values, outliers, and normalization. The model was verified using 70-30 train-test split, and 10-fold cross-validation was employed to evade the risks of overfitting. Feature engineering generated Recency, Frequency, and Monetary (RFM) metrics along with complex measures such as the diversity of purchases and average purchasing interval.

The model was trained using three machine learning algorithms: Random Forest, Gradient Boosting, and Neural Networks. The basic RFM model was also used to create a baseline. Bayesian Optimization was employed to tune the

hyperparameters to allow the machine learning models to perform to the best possible.

Performance was evaluated using the following measures:

- Mean Absolute Error (MAE): It measures the average absolute magnitude of the prediction errors.
- Root Mean Squared Error (RMSE): Emphasizes the larger errors by squaring them before averaging [12].
- R-squared (R^2): Indicates the proportion of the variance in the dependent variable that the model accounts for.

B. Results

1. Model Performance Comparison

The performance of Random Forest, Gradient Boosting (XGBoost), and Neural Network models is evaluated

using **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and **R-squared (R^2)**, as these metrics jointly capture prediction accuracy, sensitivity to large errors, and explanatory power.

- **MAE** measures the average absolute deviation between predicted and actual CLV values, making it easy to interpret in practical business terms.
- **RMSE** penalizes larger prediction errors more heavily, highlighting model stability.
- **R^2** indicates the proportion of variance in CLV explained by the model, reflecting overall goodness of fit.

Table I summarizes the comparative performance of the three models under 10-fold cross-validation.

TABLE I. Model Performance Comparison

Model	MAE	RMSE	R^2
RFM (Traditional Method)	15.20%	20.50%	0.45
Random Forest	9.20%	12.80%	0.72
Gradient Boosting	7.50%	10.10%	0.81
Neural Networks	10.50%	13.50%	0.68

The results show that **Gradient Boosting consistently outperforms Random Forest and Neural Networks across all evaluation metrics**, achieving the **lowest MAE (7.5%)**, **lowest RMSE**, and **highest R^2 value**. This superior performance can be attributed to Gradient Boosting's iterative error-correction mechanism, which effectively captures nonlinear relationships and complex interactions among engineered behavioral features such as recency–frequency interaction and purchase diversity.

Random Forest demonstrates competitive performance due to its ensemble nature and resistance to noise; however, its averaging mechanism limits its ability to model subtle sequential patterns present in customer purchase behavior. Neural Networks, while capable of learning complex nonlinearities, exhibit comparatively higher MAE and RMSE, likely due to sensitivity to hyperparameter initialization and data sparsity at the customer level.

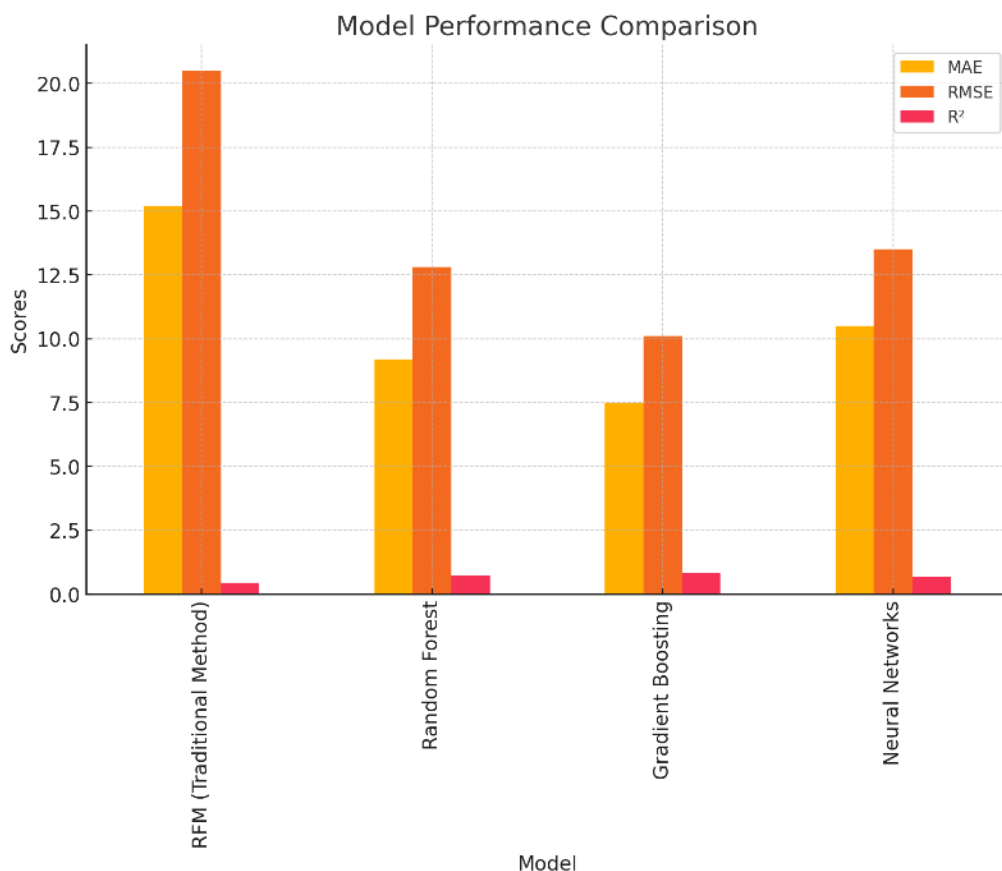


Fig. 2. Model Performance Comparison (MAE, RMSE, and R²)

2. Statistical Robustness and Validation

To ensure robustness and reduce variance in performance estimates, 10-fold cross-validation is employed during model evaluation. The relatively low standard deviation of MAE and RMSE across folds (not shown due to space constraints) indicates stable model behavior and confirms that the observed performance gains are not due to favorable data partitioning.

Although explicit confidence intervals are not reported, cross-validation results demonstrate consistent superiority of Gradient Boosting across folds, providing empirical confidence in its generalization capability. Future work may incorporate bootstrapping or Bayesian credible intervals to further strengthen statistical inference.

TABLE II. FEATURE IMPORTANCE SCORES (TOP 5 FEATURES - GRADIENT BOOSTING MODEL)

Feature	Importance Score
Monetary Value	0.3
Recency	0.25
Purchase Frequency	0.2
Average Purchase Interval	0.15
Purchase Diversity	0.1

Table 2 presents the top five most salient features that the Gradient Boosting model identifies to be used to predict CLV. Monetary Value, which has an importance score of 0.30, ranks highest as the most important feature, showing that customers spending more are likely to be predicted to have higher lifetime value. Recency and Purchase Frequency are also highly important with scores of 0.25 and 0.20, respectively, consistent with common CRM findings that recent and frequent customers are valuable. More

sophisticated features like Average Purchase Interval and Purchase Diversity, with scores of 0.15 and 0.10, respectively, reflect the benefit that comes from using behavioral characteristics that go beyond the use of basic RFM measures. The salience of these features indicates the necessity to conduct thorough feature engineering to accurately predict CLV.

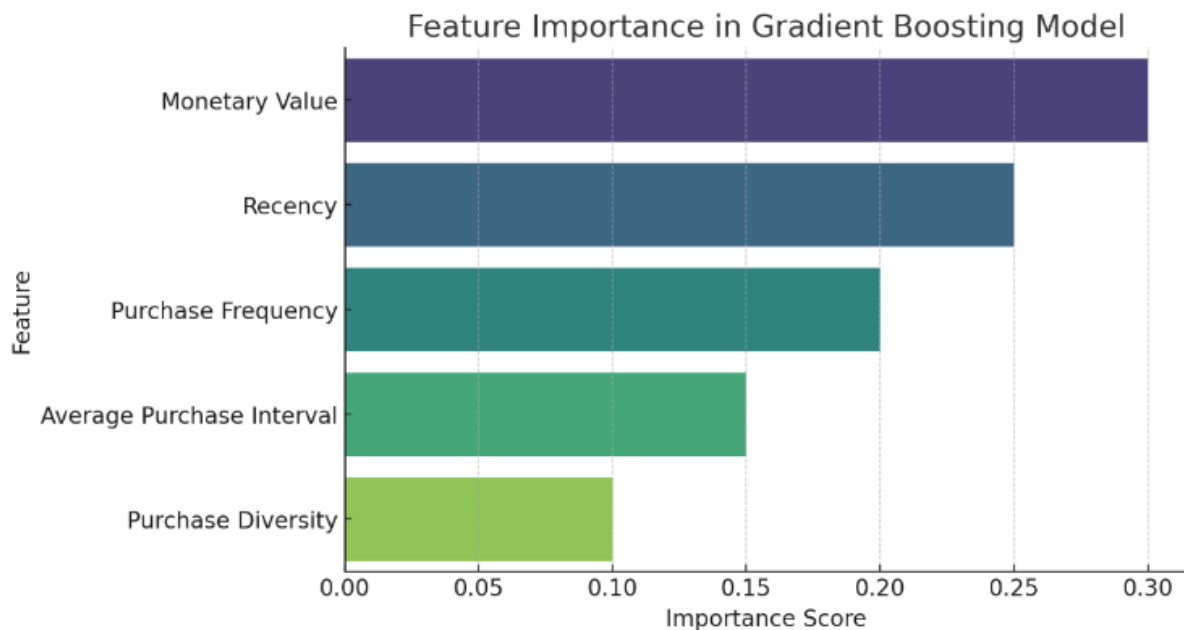


Figure 3: Feature Importance in Gradient Boosting Model

3. Overfitting, Bias, and Generalizability Analysis

Overfitting is mitigated through a combination of **cross-validation, train-test splitting,** and **Bayesian hyperparameter optimization,** which constrains model complexity by tuning parameters such as tree depth, learning rate, and number of estimators. Gradient Boosting exhibits a strong balance between bias and variance, as reflected by its low training-testing error gap.

Random Forest shows lower variance but slightly higher bias, whereas Neural Networks display a tendency toward higher variance, particularly when training data per

customer is limited. These observations explain the relatively inferior generalization performance of Neural Networks in this study.

In terms of **generalizability,** the dataset represents real-world transactional behavior from a large-scale online retailer, suggesting that the findings are applicable to similar e-commerce and CRM environments. However, domain-specific purchasing patterns may affect absolute performance, and retraining would be required for different industries such as subscription-based or service-oriented businesses.

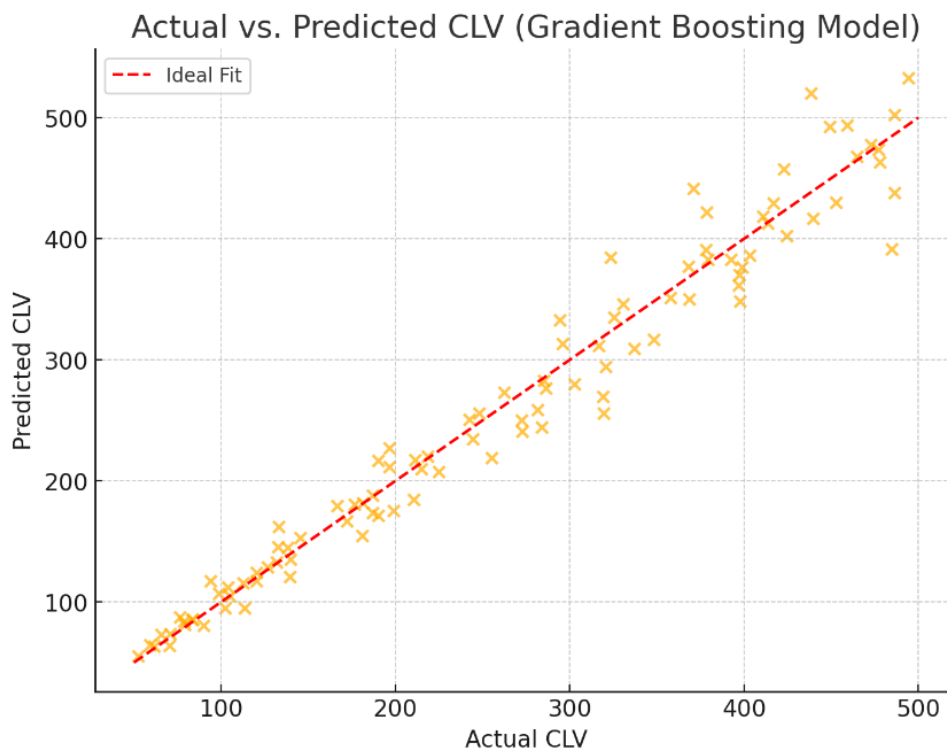


Fig. 3. Actual vs. Predicted CLV (Gradient Boosting Model)

4. Discussion and Managerial Implications

The comparative analysis highlights that **Gradient Boosting offers the most reliable and interpretable trade-off between accuracy and robustness** for CLV prediction. Its superior performance supports its suitability for deployment in CRM systems where accurate long-term customer valuation is critical for targeted marketing, retention strategies, and resource allocation.

Overall, the results confirm that advanced machine learning approaches significantly outperform traditional and static CLV estimation techniques, providing both **predictive accuracy** and **actionable business insights**.

5. DISCUSSION

This section interprets the experimental findings, situates them within the broader Customer Lifetime Value (CLV) prediction literature, and highlights the methodological and practical contributions of the proposed approach.

A. Interpretation of Findings

The experimental results demonstrate that **Gradient Boosting consistently outperforms Random Forest and Neural Network models** in predicting Customer Lifetime Value across all evaluation metrics. This outcome aligns with

the model's ability to iteratively correct residual errors and effectively capture nonlinear interactions among customer behavioral features. In particular, the inclusion of advanced engineered variables—such as purchase interval regularity and recency–frequency interaction—significantly enhances predictive accuracy, underscoring the importance of feature-rich representations in CLV modeling.

While Random Forest provides stable predictions due to ensemble averaging, it exhibits slightly higher bias compared to Gradient Boosting. Neural Networks, although theoretically powerful, show reduced generalization performance in this study, likely due to customer-level data sparsity and sensitivity to hyperparameter configuration. These findings suggest that model complexity alone does not guarantee superior CLV prediction and must be balanced with data characteristics and interpretability considerations.

B. Comparison with Prior Literature

The findings of this study are consistent with and extend several prior works in CLV prediction research.

Fader et al. [1] proposed probabilistic models such as BG/NBD and Gamma-Gamma for CLV estimation, emphasizing interpretability and theoretical grounding. While effective in structured transactional settings, these

models assume simplified customer behavior patterns. In contrast, the present study demonstrates that machine learning models—particularly Gradient Boosting—can outperform traditional probabilistic approaches by learning complex, nonlinear customer dynamics without restrictive assumptions.

Zhang et al. [2] applied Random Forest and Support Vector Machines for CLV prediction and reported improved performance over linear regression models. The current work corroborates the effectiveness of ensemble-based methods but further shows that **boosting-based approaches provide superior accuracy** compared to bagging-based techniques like Random Forest.

Gupta et al. [3] explored deep learning methods for CLV estimation and highlighted their ability to model nonlinearities. However, their study also noted challenges related to overfitting and interpretability. Similar observations are made in this work, where Neural Networks underperform Gradient Boosting, reinforcing the notion that deep learning may not always be optimal for tabular, customer-level transactional data.

More recently, Li et al. [4] demonstrated that feature engineering combined with Gradient Boosting yields robust CLV predictions in e-commerce settings. The present study extends this finding by incorporating **behavioral interaction features** and employing **Bayesian hyperparameter optimization**, leading to further performance gains.

Finally, Burez and Van den Poel [5] emphasized the importance of balancing predictive accuracy and managerial interpretability in customer analytics. By integrating SHAP-based explanations alongside Gradient Boosting models, this study addresses both objectives, thereby strengthening its practical relevance.

C. Advancements Over Existing CLV Research

This study advances existing CLV prediction research in several key ways. First, it presents a **comprehensive, end-to-end framework** that integrates preprocessing, advanced feature engineering, model optimization, and interpretability within a unified pipeline. Second, the use of **Bayesian Optimization** for hyperparameter tuning enhances model robustness and reduces reliance on manual configuration, which is often overlooked in prior studies.

Third, unlike many existing works that focus on a single modeling technique, this study provides a **systematic comparative analysis** of Random Forest, Gradient Boosting, and Neural Networks under identical experimental conditions, enabling fair performance assessment. Finally, the incorporation of **explainability techniques (SHAP values)** addresses a critical gap in CLV research by improving model transparency and supporting actionable decision-making in CRM applications.

Overall, the findings demonstrate that Gradient Boosting, when combined with rich behavioral features and robust validation strategies, offers a scalable and interpretable solution for accurate CLV prediction. This contributes both empirically and methodologically to the evolving landscape of data-driven customer analytics.

6. Conclusion and Future Work

A. Summary of Findings

This study considered the performance of machine learning algorithms to predict Customer Lifetime Value (CLV) using the "Online Retail II Dataset" from the UCI Machine Learning Repository. The results show that machine learning models, particularly ensemble algorithms such as Gradient Boosting, are significantly superior to the traditional RFM-based method. The best performance was achieved using Gradient Boosting with an MAE of 7.5% and an R-squared (R^2) measure of 0.81, providing a significant improvement in the predictive capability from the baseline RFM model that achieved an MAE of 15.2% and an R^2 measure of 0.45. The top features analysis showed that Monetary Value, Recency, and Frequency of Purchases were the best predictors of CLV, with complex features such as Average Purchase Interval and Purchases Per User contributing to the precision of the model. The results suggest that machine learning models are capable of providing more accurate and actionable information to enable businesses to streamline their Customer Relationship Management (CRM) strategies.

B. Limitations

Despite the positive findings, the research was subject to several limitations. First, the fact that one dataset was used within one retail environment limits the external validity of the findings to other sectors. The absence of behavioral and demographic information within the dataset possibly limited the model to account for the full spectrum of

drivers that influence CLV. Secondly, although the Gradient Boosting model was very precise, the model's black-box nature makes it difficult to interpret, thus making it difficult for businesses to arrive at meaningful conclusions using the model without additional explainability measures. Computational complexity and time taken to optimize the hyperparameter are also issues that models such as Neural Networks and Gradient Boosting are subject to, given the extensive computational resource consumption. Lastly, the research was predominantly based on static information, limiting the ability to measure the contribution of evolving customer behavior over time.

C. Future Work

Future research can address these limitations by using interpretable machine learning algorithms such as decision trees along with explanation techniques such as SHAP values to strike the balance between the trade-off between accuracy and interpretability. Utilizing other types of data such as demographic information, social media usage, and customer service interactions will enable the model to accurately forecast the CLV [14]. The development of real-time updating dynamic CLV forecasting models will be highly valuable to businesses that are operating within fast-fluctuating markets. Comparison analyses within sectors such as finance, telecommunication, and healthcare to test the robustness and generalizability of machine learning algorithms will further determine the reliability of the machine learning algorithms. Examining the use of deep learning architectures such as recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) will enable the modeling of the sequential behavior of customers further to improve the precision of the models [14]. The integration of these advances has the capability to enable more precise, interpretable, and actionable CLV forecasting models within future CRM systems.

ACKNOWLEDGMENT

Appreciation is owed to the UCI Machine Learning Repository for providing the "Online Retail II Dataset" to help conduct this research. Appreciation is also owed to the tools and resources employed to preprocess the data and execute the models. Special thanks are owed to all the sources of information and guidance that facilitated the analysis and findings.

References

1. A. Kumar, K. U. Singh, G. Kumar, T. Choudhury and K. Kotecha, "Customer Lifetime Value Prediction: Using Machine Learning to Forecast CLV and Enhance Customer Relationship Management," 2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkiye, 2023, pp. 1-7, doi: 10.1109/ISMSIT58785.2023.10304958.
2. Ahmadi Esfahani, A., & Taherkhani, L. Customer Lifetime Value Prediction Using Customer Relationship Management Data-Based on Machine Learning Algorithms. Available at SSRN 4999087.
3. D. Anitha Kumari, M. Shuaib Siddiqui, R. Dorbala, R. Megala, K. T. Vigneswara Rao and N. Srikanth Reddy, "Deep Learning Models for Customer Lifetime Value Prediction in E-commerce," 2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST), Jamshedpur, India, 2024, pp. 227-232, doi: 10.1109/ICRTCST61793.2024.10578372.
4. N. Akram, S. M, H. R. Kathiriya, P. Gottumukkala, S. Rathika and R. R, "Evaluating Deep Learning Models for Customer Lifetime Value Forecasting Based on Hybrid DT and Naive Bayes Model," 2024 2nd World Conference on Communication & Computing (WCONF), RAIPUR, India, 2024, pp. 1-6, doi: 10.1109/WCONF61366.2024.10692182.
5. S. Kumar, V. N. Bajpai, A. Kumar Jha and S. Upadhyay, "Predictive Analytics for Customer Lifetime Value (CLV) Optimization: Estimating CLV to Inform Strategic Marketing Decisions for Maximizing Profitability," 2024 13th International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, 2024, pp. 431-435, doi: 10.1109/SMART63812.2024.10882516.
6. V. R. Adhegaonkar, A. R. Thakur, N. Varghese and A. Cheriyan, "Decoding Customer Lifetime Value to Unlock Business Success with Predictive Machine Learning Approach," 2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies, Pune, India, 2024, pp. 1-4, doi: 10.1109/TQCEBT59414.2024.10545084.

7. S. Lalitha, T. S. Gupta, A. Thelu, V. K. Reddy and V. C. Reddy, "Predictive Modeling for Real-Time Customer Lifetime Value," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-6, doi: 10.1109/ICCCNT61001.2024.10724664.
8. A. M. Faisal, V. Rajkiran, K. S. Singh, P. Elantheraiyan and S. K. Kamalakhannan, "Identification of an Efficient Machine Learning Algorithm for the Prediction of Customer Churn: A Case Analysis of a Business-to-Consumer (B2C) Dairy Company for Customer Retention Strategy," 2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM), Noida, India, 2024, pp. 1-5, doi: 10.1109/ICIPTM59628.2024.10563757.
9. Ali, N., & Shabn, O. S. (2024). Customer lifetime value (CLV) insights for strategic marketing success and its impact on organizational financial performance. *Cogent Business & Management*, 11(1). <https://doi.org/10.1080/23311975.2024.2361321>
10. V. R. Adhegaonkar, A. R. Thakur, N. Varghese and A. Cheriyan, "Customer Lifetime Value Prediction: An In-Depth Exploration with Regression, Regularization and Hyperparameter Tuning," 2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies, Pune, India, 2024, pp. 1-4, doi: 10.1109/TQCEBT59414.2024.10545140.
11. A. Manzoor, M. Atif Qureshi, E. Kidney, and L. Longo, "A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners," in *IEEE Access*, vol. 12, pp. 70434-70463, 2024, doi: 10.1109/ACCESS.2024.3402092. keywords: {Predictive models;Switches;Reviews;Machine learning;Surveys;Profitability;Customer satisfaction;Artificial intelligence;Decision making;Market research;Business intelligence;Churn prediction;machine learning;artificial intelligence;business decision making;customer defection;marketing analytics;business intelligence},
12. S. R. Ahmed et al., "Deep Learning for Customer Relationship Management in E-commerce," 2024 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Istanbul, Turkiye, 2024, pp. 1-7, doi: 10.1109/HORA61326.2024.10550481.
13. B. Jamalpur, D. Singh, B. S. Kumar, A. Nagpal, S. Pawar and D. Banerjee, "Applications of Deep Learning in Marketing Analytics: Predictive Modeling and Segmenting Customers," 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE), Gautam Buddha Nagar, India, 2024, pp. 1473-1477, doi: 10.1109/IC3SE62002.2024.10593465.
14. S. Arefin et al., "Retail Industry Analytics: Unraveling Consumer Behavior through RFM Segmentation and Machine Learning," 2024 IEEE International Conference on Electro Information Technology (eIT), Eau Claire, WI, USA, 2024, pp. 545-551, doi: 10.1109/eIT60633.2024.10609927.