Customer Lifetime Value Prediction Using Machine Learning

Karan Choksey, Vedant Avhad, Sandip T. Shingade

Department of Computer Engineering and Information Technology, Veermata Jijabai Technological Institute, Mumbai, India
{kdchoksey, vaavhad}_b21@it.vjti.ac.in, {stshingade}@it.vjti.ac.in

Abstract. This research focuses on predicting customer lifetime value (CLV) using machine learning techniques, a crucial task for businesses to optimize their marketing strategies and resource allocation. The study aims to address the challenge of accurately forecasting CLV by applying various machine learning algorithms. We utilize Random Forest, Support Vector Machine, AdaBoost, Linear Regression and Gradient Boosting Decision Trees to assess their performance in predicting CLV. The Online Retail Dataset, curated by Ulrik Thyge Pedersen, is used to test these models. By evaluating the accuracy of each algorithm, the research identifies the most suitable approach for predicting customer lifetime value in this particular dataset, offering a more reliable method for businesses compared to previous approaches. This comparison of multiple algorithms enhances the decision-making process for optimal CLV prediction.

Keywords: Customer Lifetime Value, Machine Learning, Classification, Clustering, Predictive Modeling, Regression

1 Introduction

Customer lifetime value (CLV) prediction is a key challenge for businesses aiming to optimize their marketing efforts and improve customer relationship management. CLV refers to the predicted net profit a customer will bring to a business over the duration of their relationship. Predicting CLV accurately allows businesses to allocate resources more effectively, personalize marketing campaigns, and retain high-value customers. Given the rapid growth of e-commerce and online retail, the ability to predict CLV with precision has become an increasingly critical area of research within the field of machine learning and data analytics.

Problem Illustration: Consider an online retail business that spends significant resources on customer acquisition. However, without an accurate prediction of CLV, it becomes difficult to determine how much investment is justifiable for retaining or acquiring specific customers. For instance, a business may focus on acquiring customers who, in the long run, provide minimal returns. On the other hand, businesses might overlook valuable customers whose potential is not immediately obvious.

Problem Identification: The problem of predicting CLV has been recognized within both academia and industry for years, as businesses continue to struggle with effectively forecasting customer behavior. While traditional methods focus on historical data and simple calculations, these approaches often fail to provide actionable insights. During this research, we found that while the problem is well-known, there is still a gap in identifying the most accurate and suitable machine learning model for CLV prediction within the context of large and diverse datasets such as the Online Retail Dataset.

Need for Solution: The need for a reliable solution to predict CLV has become more pressing as businesses face increasing competition and data complexity. An accurate CLV prediction model helps businesses in resource allocation, targeting the right customer segments, and improving customer retention strategies. Addressing this challenge through machine learning techniques contributes to the field of predictive analytics by offering scalable solutions that improve upon traditional approaches.

Review of Related Work: Base Papers Review: Recent research has explored a variety of machine learning algorithms for CLV prediction. For instance, studies have utilized decision trees, neural networks, and regression models to predict customer behavior. Some papers have proposed novel hypotheses on feature selection and customer segmentation, suggesting that a multi-step approach improves prediction accuracy.

Limitations and Improvements: While these studies have made significant contributions, many have limitations, such as ignoring complex customer behaviors, not considering temporal changes in customer activity, or

relying on oversimplified assumptions about customer value. Additional parameters, such as seasonality, product preferences, and socio-demographic data, could enhance the predictive accuracy of CLV models. Moreover, some models are computationally expensive and lack generalizability across diverse datasets.

Future Work: Future research in CLV prediction includes the exploration of hybrid models combining various algorithms, such as ensemble methods, as well as the integration of deep learning approaches for capturing complex patterns in customer data. Other research directions also suggest using real-time data to improve prediction accuracy and applying reinforcement learning for dynamic CLV forecasting.

Innovative Ideas: Our approach introduces a comparison of multiple machine learning algorithms, each tested for their suitability in predicting CLV, offering a robust methodology to identify the most effective model. Moreover, the study includes an evaluation of customer segmentation methods that can be combined with CLV prediction to tailor marketing strategies.

Proposed Solution: In this study, we apply and compare six machine learning algorithms—Naive Bayes, Random Forest, Support Vector Machine, K-Nearest Neighbors, AdaBoost, and Gradient Boosting Decision Trees—on the Online Retail Dataset. By evaluating each algorithm's accuracy and performance, we aim to identify the most suitable approach for CLV prediction, enhancing the ability to forecast customer behavior and optimize business strategies.

Research Objectives:

- To compare the performance of various machine learning algorithms in predicting customer lifetime value.
- To identify the optimal algorithm for CLV prediction within the Online Retail Dataset and suggest improvements for business strategy optimization.

2 Related Work

A Study on Customer Lifetime Value Prediction Using Machine Learning:
This paper investigates the application of machine learning techniques, such as regression analysis, random forest, and gradient boosting, to predict customer lifetime value (CLV). It emphasizes the critical role of accurate CLV prediction in helping businesses allocate resources effectively, personalize customer experiences, and enhance profitability. The study demonstrates that machine learning models can significantly outperform traditional statistical methods but highlights the need for substantial amounts of clean and structured data for optimal results.

Customer Lifetime Value Prediction: A Comparative Analysis of Machine Learning Techniques
This research provides a comprehensive comparison of several machine learning algorithms, including decision trees, neural networks, and support vector machines (SVM), in the context of CLV prediction. It stresses the importance of data preprocessing and feature engineering in improving the predictive performance of machine learning models. While the study offers valuable insights into the comparative strengths and weaknesses of these algorithms, it acknowledges the challenges posed by varying datasets and the need for extensive experimentation to achieve reliable results.

Improving Customer Lifetime Value Prediction Using Hybrid Machine Learning Models
This paper proposes the use of hybrid machine learning models, such as stacking and blending, to enhance the accuracy and robustness of CLV predictions. By combining the strengths of multiple algorithms, the study demonstrates the potential for improved performance across diverse datasets. However, it also highlights the increased complexity in model development, tuning, and computational requirements as a trade-off for achieving better results.

Gap Analysis
Despite significant advancements in CLV prediction, several gaps remain in the existing literature:

- **Limited exploration of hybrid models:**
  While some studies have experimented with hybrid machine learning models, a detailed exploration of the most effective combinations of techniques is still lacking. Addressing this gap could lead to better-performing models for CLV prediction.

Literature Comparison Table

Paper	Aim	Methodology	Accuracy	Limitations
Paper 1	Predict CLV using machine learning	Regression, Random Forest, Gradient Boosting	High accuracy	Requires large, clean, and structured datasets
Paper 2	Compare machine learning techniques for CLV prediction	Decision Trees, Neural Networks, SVM	Varies depending on the dataset	Requires extensive feature engineering and preprocessing
Paper 3	Improve CLV prediction using hybrid models	Hybrid models (e.g., stacking, blending)	Higher accuracy	Complex model development and tuning

Fig. 1. Literature comparision table

- **Lack of focus on interpretability:**
  Most studies prioritize accuracy over interpretability, especially when using complex models like neural networks or hybrid systems. This makes it difficult for businesses to understand the driving factors behind predictions, limiting practical usability.

- **Insufficient attention to dynamic customer behavior:**
  Current models often assume static customer behavior, overlooking the reality that customer preferences and purchasing patterns evolve over time. Developing models that account for this dynamism could improve the relevance and accuracy of predictions.

Justification for Addressing the Gap

Addressing the identified gaps is essential for advancing the field of CLV prediction and improving its practical applications:

- **Improved accuracy:**
  Combining different machine learning techniques in hybrid models can leverage their unique strengths, resulting in enhanced predictive accuracy and robustness. This ensures businesses receive more reliable predictions for strategic decision-making.

- **Enhanced interpretability:**
  Incorporating techniques like feature importance analysis and SHAP values into hybrid models can make predictions more transparent. This allows businesses to understand the factors influencing CLV, fostering trust in the models and enabling more actionable insights.

- **Adaptability to dynamic behavior:**
  Developing models that account for evolving customer behavior ensures predictions remain relevant over time. By aligning with real-world dynamics, these models can provide more timely and accurate forecasts, helping businesses adapt to market changes effectively.

Research Objectives

This study aims to address the identified gaps and advance customer lifetime value prediction through the following objectives:

– **Developing a hybrid machine learning model:**
To combine the strengths of various machine learning techniques, such as ensemble and hybrid approaches, for improving the accuracy and robustness of CLV predictions. This includes exploring effective combinations of models like random forest, gradient boosting, and support vector machines.

– **Enhancing model interpretability:**
To integrate interpretability methods, such as feature importance analysis and SHAP values, enabling businesses to understand the key factors influencing predictions and make informed decisions based on these insights.

– **Adapting to dynamic customer behavior:**
To design models that account for evolving customer preferences and purchasing patterns, ensuring predictions remain relevant and accurate over time, even in dynamic market environments.

3 Methodology

Data preprocessing is an essential step in transforming raw datasets into a structured and analyzable format, especially for tasks like Customer Lifetime Value (CLV) prediction. It involves several techniques to ensure data quality and reliability, ultimately leading to more accurate and meaningful insights.

In this process, missing values were handled by removing rows with missing customer IDs and imputing relevant values where necessary. Duplicate records were eliminated to maintain data integrity, and invoice dates were converted into a datetime format for analysis. Metrics such as total price, recency, frequency, monetary value, and total quantity purchased were computed by aggregating transaction data. Additionally, the customer lifespan was estimated based on purchase frequency and recency.

To prepare the data for modeling, numerical features like monetary value and frequency were normalized using z-score standardization, ensuring consistency and scalability. The CLV metric was then calculated by integrating these normalized features with customer lifespan. Finally, the dataset was split into training and testing subsets to facilitate model development and evaluation.

These preprocessing steps transformed raw data into a structured, clean, and consistent format, ensuring the reliability and accuracy of subsequent predictive models.

4 Proposed Improved approch algorithm for your problem

The goal of this study is to predict Customer Lifetime Value (CLV) using historical customer data. CLV represents the total monetary value a business can expect from a customer over the course of their relationship. Accurate CLV predictions are crucial for resource allocation, personalized marketing, and customer retention strategies. The problem is formulated as a regression task, where the target variable (CLV) is predicted based on input features such as purchase behavior, recency of transactions, and customer lifespan.

4.1 Mathematical Formulation

The prediction of CLV involves learning a function $f : X \to y$, where:

– X: Feature matrix containing predictors (*Recency*, *FrequencyLog*, *MonetaryLog*, *TotalQuantity*, *AvgOrderValue*, and *CustomerLifespan*).
– y: Target variable representing *CLV*.

The objective is to minimize the error between the actual CLV (y_i) and the predicted CLV ($\hat{y}_i$) using a loss function.

Loss Function: The Mean Squared Error (MSE) is given by:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

where n is the number of samples.

4.2 Gradient Boosting Regressor Architecture

The **Gradient Boosting Regressor (GBR)** is selected as the proposed approach due to its strong predictive performance. The key components of GBR are:

- **Base Learner:** Decision trees are used as weak learners. Each tree is shallow, typically having a limited depth to avoid overfitting.
- **Boosting Mechanism:** Trees are built sequentially, with each subsequent tree correcting the errors of the prior ones.
- **Learning Process:** The model minimizes the loss function by computing gradients (errors) at each iteration and fitting a new tree to these residuals:

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \nu \cdot h_m(x_i) \tag{2}$$

where $\hat{y}_i^{(m)}$ is the prediction at iteration m, ν is the learning rate, and $h_m(x_i)$ is the weak learner fitted to residuals.

4.3 System Architecture

The proposed system architecture for solving this regression problem consists of the following components:

- **Input Layer:** Preprocessed features derived from the dataset.
  - Handle missing values, outliers, and scaling for uniformity.
- **Gradient Boosting Model:** Sequential decision trees with hyperparameters tuned for the best performance.
  - Key hyperparameters include the number of estimators, maximum tree depth, and learning rate.
- **Output Layer:** Predicted CLV for each customer.

4.4 Problem to Be Solved

Traditional methods like linear regression fail to capture the non-linear relationships between customer behavior and CLV. This results in suboptimal predictions, impacting decision-making.

The use of Gradient Boosting addresses these challenges by:

- Capturing complex patterns in data.
- Correcting errors iteratively.
- Reducing overfitting through regularization techniques such as learning rate and tree depth control.

5 Experimental Results

5.1 Dataset

The dataset used in this study contains transactional data from an online retail store. It provides comprehensive information about customer purchases, including details such as invoice numbers, stock codes, product descriptions, quantities purchased, transaction dates, unit prices, customer identifiers, and the country of the customer.

The dataset includes the following key attributes:

- **InvoiceNo**: A unique identifier for each transaction.
- **StockCode**: A unique code assigned to each product.
- **Description**: A textual description of the purchased product.
- **Quantity**: The number of units of the product purchased in each transaction.
- **InvoiceDate**: The timestamp of the transaction, indicating the date and time of purchase.
- **UnitPrice**: The price per unit of the product at the time of purchase.
- **CustomerID**: A unique identifier for each customer.
- **Country**: The country where the customer resides or made the purchase.

The dataset comprises transactional records spanning multiple countries, with the majority of the data originating from the United Kingdom. Each record represents a single transaction, offering granular insights into customer purchasing behavior. The inclusion of customer-specific identifiers enables the analysis of recurring purchase patterns and customer segmentation.

This dataset serves as a foundation for calculating key metrics such as Recency, Frequency, Monetary value, and Customer Lifetime Value (CLV). Its structured format and rich transactional data make it suitable for understanding customer behavior and developing predictive models for CLV. Furthermore, the availability of time-based attributes (e.g., InvoiceDate) allows temporal analysis to study purchasing trends over time.

5.2 Statistical Procedures Used

To test the hypotheses, five regression models were applied: **Gradient Boosting Decision Tree**, **AdaBoost Regressor**, **Support Vector Machine (SVM) Regressor**, **Random Forest**, and **Linear Regression**. The models were evaluated using the following metrics:

- **Root Mean Squared Error (RMSE):** Measures the magnitude of prediction error in the same units as the target variable.
- **Mean Squared Error (MSE):** Captures the squared differences between actual and predicted values (used in SVM).
- R^2**-Score:** Determines the proportion of variance in the target variable (Customer Lifetime Value, CLV) explained by the model.

The dataset was split into training (70%) and testing (30%) subsets to evaluate model performance and generalization ability.

5.3 Outcomes of the Procedures

The results of the models are summarized below:

Gradient Boosting Decision Tree:

- Train RMSE: 289.29
- Test RMSE: 1346.98
- R^2-Score (Test): 0.9898

This model achieved the best performance, with the lowest RMSE on the test set and the highest R^2-Score, indicating excellent predictive accuracy and generalization.

AdaBoost Regressor:

- Train RMSE: 1571.30
- Test RMSE: 2557.89
- R^2-Score (Test): 0.9634

While AdaBoost performed well, it was less accurate than Gradient Boosting, with higher RMSE on both the training and test sets.

Support Vector Machine (SVM) Regressor:

- Train RMSE: 0.67
- Test RMSE: 0.69
- Train MSE: 0.4543
- Test MSE: 0.4707
- R^2-Score (Test): 0.4590

SVM achieved very low RMSE values but had the lowest R^2-Score among all models, indicating limited generalization and explanatory power.

Random Forest:

- Train RMSE: 383.84
- Test RMSE: 1462.84
- R^2-Score (Test): 0.9880

Random Forest performed slightly worse than Gradient Boosting but still exhibited strong predictive power and generalization.

Model	Train RMSE	Test RMSE	R^2-Score (Test)
Gradient Boosting Decision Tree	289.29	1346.98	0.9898
AdaBoost Regressor	1571.30	2557.89	0.9634
Support Vector Regressor (SVM)	0.67	0.69	0.4590
Random Forest	383.84	1462.84	0.9880
Linear Regression	2911.72	3191.26	0.9430

Table 1. Performance comparison of different regression models for predicting Customer Lifetime Value (CLV).

Linear Regression:

- Train RMSE: 2911.72
- Test RMSE: 3191.26
- R^2-Score (Test): 0.9430

Linear Regression showed the poorest performance, with the highest prediction errors and the lowest ability to model non-linear relationships.

5.4 Subsidiary Findings

- **Feature Importance:** Gradient Boosting and AdaBoost identified *MonetaryLog*, *FrequencyLog*, and *Customer-Lifespan* as the most influential predictors of CLV, while *Recency* and *TotalQuantity* had less impact.
- **Preprocessing Impact:** Proper data preprocessing, including handling missing values, outlier removal, and normalization, was crucial. Without preprocessing, both AdaBoost and SVM showed degraded performance.
- **Algorithm-Specific Insights:**
  - Gradient Boosting iteratively reduced errors, making it the most robust and accurate model for predicting CLV.
  - AdaBoost effectively handled noisy data but struggled to achieve the accuracy of Gradient Boosting.
  - SVM was computationally expensive and less effective for larger datasets, highlighting the importance of proper scaling and hyperparameter tuning.

Overall, Gradient Boosting emerged as the best model for predicting CLV, achieving a balance of accuracy, generalization, and robustness, followed by Random Forest and AdaBoost as viable alternatives for moderately complex tasks.

6 Conclusion

The results of this study highlight the superiority of ensemble-based models, particularly the Gradient Boosting Regressor, in predicting Customer Lifetime Value (CLV). This performance can be attributed to the model's ability to iteratively minimize errors by correcting residuals from previous iterations, thereby effectively capturing the non-linear and complex relationships between the features and CLV. Furthermore, preprocessing steps such as normalization and outlier handling enhanced model accuracy by ensuring a consistent and clean dataset, which was crucial for ensemble methods like Gradient Boosting and AdaBoost to perform optimally. In contrast, simpler models like Linear Regression struggled due to their inability to model non-linear patterns, and Support Vector Regression was sensitive to feature scaling and hyperparameter tuning, leading to less generalizable predictions.

The results signify that advanced machine learning techniques such as Gradient Boosting and AdaBoost are highly effective tools for predicting CLV, providing accurate and generalizable models. Gradient Boosting achieved the lowest test RMSE and the highest R^2-score, indicating that it was the most reliable model in explaining the variance in CLV. These findings suggest that businesses can leverage such models to make informed decisions regarding customer segmentation, resource allocation, and personalized marketing strategies. Additionally, the strong predictive performance of ensemble models demonstrates their robustness in handling complex datasets with multiple features and varying levels of importance.

This research builds upon prior studies on customer analytics and lifetime value prediction by demonstrating the effectiveness of ensemble-based models in this domain. Previous research has often relied on simpler techniques like Linear Regression or rule-based approaches, which were limited in their ability to capture complex interactions between features. By incorporating Gradient Boosting and AdaBoost, this study bridges the gap between traditional statistical methods and modern machine learning algorithms. The findings corroborate earlier research emphasizing the importance of Recency, Frequency, and Monetary (RFM) metrics in CLV prediction, while extending this understanding by showing how additional derived features, such as 'CustomerLifespan', and advanced preprocessing techniques further enhance predictive accuracy.

In conclusion, this study underscores the value of adopting ensemble machine learning methods for CLV prediction. The results not only validate the efficacy of Gradient Boosting as the leading model but also highlight the importance of thorough data preprocessing and feature engineering. Future research can further refine these methods by exploring deep learning architectures or hybrid approaches for enhanced prediction in more diverse datasets.

Acknowledgements. The second author was in part supported by a research grant from Google.
[2]. [1]

References

1. S. Filippou, A. Tsiartas, P. Hadjineophytou, S. Christofides, K. Malialis, and Christos Panayiotou. Improving customer experience in call centers with intelligent customer-agent pairing, 05 2023.
2. Ankit Kumar, Kamred Udham Singh, Gaurav Kumar, Tanupriya Choudhury, and Ketan Kotecha. Customer lifetime value prediction: Using machine learning to forecast clv and enhance customer relationship management. In *2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–7, 2023.