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 RFM Analysis

Pankaj Pachauri
University of Rajasthan, Jaipur

ARTICLE INFO

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

Keywords:
Customer Lifetime Value (CLTV), Machine Learning, RFM Analysis, Predictive Analytics, Customer Relationship Management, Customer Segmentation, Regression Analysis, Churn Prediction, Marketing ROI, Hybrid Models.

Correspondence:
E-mail:sharmajipankaj700@gmail.com

ABSTRACT

Predicting Customer Lifetime Value (CLTV) is crucial for effective marketing resource allocation and strategic customer relationship management. This paper proposes a novel hybrid approach that integrates traditional Recency, Frequency, and Monetary (RFM) analysis with advanced machine learning techniques to enhance the accuracy and robustness of CLTV predictions. We develop and evaluate several machine learning models, including regression algorithms and classification models for churn prediction, and compare their performance against traditional RFM-based methods. The proposed hybrid model leverages the strengths of both approaches, using RFM scores as features within the machine learning models. Empirical results, derived from a real-world transactional dataset, demonstrate that the hybrid model significantly outperforms both traditional RFM analysis and individual machine learning models in predicting CLTV, leading to improved marketing ROI and customer retention strategies. Furthermore, the paper provides insights into the key factors driving customer lifetime value and offers practical recommendations for businesses to optimize their customer engagement strategies.

1.Introduction

With today's highly competitive business environment, it is critical to know and optimize Customer Lifetime Value (CLTV) for long-term business growth and profitability. CLTV is the aggregate revenue a company can realistically anticipate from one customer account over the life of the relationship. Precise prediction of CLTV allows companies to target high-value customers, tailor marketing efforts, allocate resources more effectively, and take preemptive action against potential churn threats. Thus, CLTV becomes a foundation for strategic business decision-making in functions like marketing, sales, and customer service.

The conventional approach to predicting CLTV tends to be based on fairly elementary methods like Recency, Frequency, and Monetary (RFM) analysis.

Although RFM is a good framework for segmenting customers based on their history of behavior, it has a number of drawbacks. RFM analysis usually distributes scores in accordance with pre-established rules and thresholds, without being able to capture the intricate, non-linear relationships between customer behavior and future value. Furthermore, RFM analysis tends to neglect other important drivers of CLTV, including customer demographics, engagement metrics, and competitive dynamics.

Machine learning provides an effective alternative to legacy CLTV forecasting techniques. Machine learning algorithms can learn intricate patterns from high-volume datasets automatically and make more precise and subtle predictions. Several machine learning methods, such as regression models, classification models, and neural networks, have been effectively used for CLTV prediction. Machine learning model performance relies significantly on data quality and representativeness and proper model parameter selection and tuning. In addition, machine learning algorithms can at times be "black boxes," which complicates the interpretation of the drivers of CLTV and actionable insights.

This paper overcomes the shortcomings of conventional RFM analysis and single machine learning models through a new hybrid model for the prediction of CLTV. Our model couples RFM analysis and machine learning to leverage the strengths of both methods. In particular, we employ RFM scores as inputs for machine learning models so that the models can learn from both the past behavior contained in the RFM scores and the complicated non-linear relationship between RFM scores and future CLTV.

The goals of this study are:

To construct a hybrid model of CLTV prediction that combines RFM analysis with machine learning methods.

To compare the performance of the hybrid model with conventional RFM analysis and standalone machine learning models.

To identify the key factors driving customer lifetime value.

To provide practical recommendations for businesses to optimize their customer engagement strategies based on CLTV predictions.

2.Literature Review

The field of Customer Lifetime Value (CLTV) prediction has garnered significant attention from both academics and practitioners. Early research focused primarily on developing analytical models to estimate the expected future profit from a customer relationship (Berger & Nasr, 1998). These models often relied on simplifying assumptions about customer behavior and lacked the ability to adapt to changing market conditions.

RFM (Recency, Frequency, Monetary) analysis emerged as a widely adopted technique for customer segmentation and targeting (Hughes, 1994). RFM analysis provides a simple yet effective way to rank customers based on their past transaction history. However, RFM analysis

has been criticized for its reliance on predefined rules and thresholds, which may not accurately reflect the underlying customer behavior (Stone, 1988). Furthermore, RFM analysis typically ignores other important factors that influence CLTV, such as customer demographics and engagement metrics (Dwyer, 1997).

Some newer studies have investigated the application of machine learning algorithms to predict CLTV. Linear regression and logistic regression have been applied to forecast CLTV using a range of customer attributes (Gupta et al., 2006). Decision trees and support vector machines have been applied to predict customer churn, an important determinant of CLTV (Verbeke et al., 2012). Neural networks, in their capacity to capture complex non-linear relationships, have also been used for CLTV prediction (Jain & Singh, 2002).

Comparisons have been made among the performance of various machine learning methods to predict CLTV. For instance, Fader et al. (2005) created a probabilistic model of CLTV prediction utilizing the Pareto/NBD model, both considering customer transaction behavior and customer attrition. Reinartz and Kumar (2003) compared the performance of multiple machine learning and statistical models for the prediction of CLTV and found that regression models performed better than other methods. Nevertheless, they did mention that the performance of various models can be different in a particular dataset and application.

Although machine learning methods have been promising with regards to predicting CLTV, they do come with some disadvantages. Machine learning algorithms tend to need a lot of data to train properly, and they can be sensitive to outliers and noise in the data. Additionally, machine learning models occasionally are "black boxes" where it is not possible to understand the underlying causes of CLTV and derive meaningful insights (Linoff & Berry, 2011).

Hybrid methods which utilize a blend of conventional methods with machine learning have now appeared as a promising way to predict CLTV. These methods utilize the advantages of both methodologies, employing conventional methods to offer a structured format for inspection and machine learning to pick up sophisticated patterns in the data. For instance, Tsai and Chiu (2004) developed a hybrid model that combines RFM analysis with neural networks for customer segmentation and prediction of CLTV. They established that the hybrid model performs better than RFM analysis and neural networks independently. In the same manner, Kim et al. (2005) created a hybrid model that unites RFM analysis with support vector machines to predict customer churn. They demonstrated that the hybrid model surpassed support vector machines in accuracy.

Critical Analysis of Existing Literature:

Existing literature presents a complete review of CLTV prediction methodologies, yet some gaps exist. First, most studies cover specific industries or datasets, reducing the generality of their results. Second, there is limited research that compares the performance of various hybrid methods for CLTV prediction. Third, there are not many studies addressing the question of interpretability of machine learning models in CLTV prediction. The "black box" character of most machine learning algorithms makes it difficult to map predictions into effective marketing

campaigns. Fourth, the fact that customer behavior is dynamic in nature is usually ignored. CLTV prediction models must be agile in responding to changing customer tastes and market conditions.

This work seeks to fill these loopholes by proposing a new hybrid model for CLTV prediction that combines RFM analysis with machine learning. We compare the performance of the hybrid model with the conventional RFM analysis as well as standalone machine learning models on a real transactional dataset. We also discuss methods of enhancing the interpretability of machine learning models and creating adaptive CLTV prediction models that are able to adapt to evolving market conditions. Our research extends the state of the art by offering an expanded and more pragmatic methodology for CLTV prediction.

3.Methodology

This research uses a quantitative research method with a real-world transactional data to construct and test the proposed hybrid model for CLTV prediction. The approach includes the following process:

Data Collection and Preprocessing:

Data Source: The data contains transactional data of an online store for a duration of three years (2022-2024). The data contains customer IDs, order dates, order amounts, product categories, and customer demographics (age, gender, location).

Data Cleaning: The data is preprocessed to manage missing values, outliers, and inconsistencies. Missing values are replaced using suitable methods (e.g., mean imputation for numerical features, mode imputation for categorical features). Outliers are detected and deleted using statistical techniques (e.g., interquartile range (IQR) method).

Feature Engineering: Various features are engineered from raw data, such as:

Recency: Days elapsed since last purchase by the customer.

Frequency: Number of times the customer has purchased.

Monetary Value: Total money spent by the customer.

Average Order Value: Average money spent on each order.

Product Category Diversity: Number of product categories purchased by the customer.

Customer Tenure: Days elapsed since first purchase by the customer.

RFM Analysis:

RFM Score Calculation: Customers are segmented based on their RFM values. Each RFM dimension (Recency, Frequency, Monetary) is divided into quartiles (or quintiles), and customers are assigned scores from 1 to 4 (or 1 to 5) based on their quartile/quintile ranking. For example, a customer in the lowest quartile of Recency receives a score of 1, while a customer in the highest quartile receives a score of 4.RFM Segmentation: Customers are segmented into various

segments based on their overall RFM scores. For instance, customers who have high Recency, Frequency, and Monetary scores are marked as "Champions," and customers who have low Recency, Frequency, and Monetary scores are marked as "Lost Customers."

Machine Learning Model Development:

Target Variable: The target variable used for prediction of CLTV is the total spent amount by the customer in the following year (2025).

Model Selection: A few machine learning models are compared, such as:

Linear Regression: A linear model that outputs CLTV as a linear function of the input features.

Random Forest Regression: Ensemble learning algorithm that aggregates various decision trees to enhance prediction accuracy.

Gradient Boosting Regression: Ensemble learning algorithm that constructs decision trees sequentially in order to reduce prediction errors.

Support Vector Regression (SVR): Non-linear regression that transforms the input features into a high-dimensional space and identifies the best hyperplane that best fits the data.

Logistic Regression (to predict Churn): Utilized to predict the likelihood of a customer churning (not making a purchase within the next year).

Churn probability is later integrated into the CLTV computation.

Feature Selection: Feature selection methods are utilized to determine the most meaningful features used in CLTV forecasting. Methods like Recursive Feature Elimination (RFE) and Random Forest feature importance are utilized.

Model Training and Validation: The data is divided into training (70%) and testing (30%) sets. The machine learning models are trained on the training set and tested on the testing set.

Hyperparameter Tuning: The hyperparameter tuning is conducted with methods like grid search and cross-validation to enhance the performance of the machine learning models.

Hybrid Model Development:

RFM Integration: RFM scores are integrated as features within the machine learning models. This allows the models to leverage both the historical behavior reflected in the RFM scores and the complex, non-linear relationships between RFM scores and future CLTV. Specifically, the RFM scores (R score, F score, M score) are added as additional input features to the machine learning models.

Churn Probability Integration: The predicted churn probability from the Logistic Regression model is also incorporated into the CLTV calculation. A higher churn probability reduces the predicted CLTV. The CLTV calculation is adjusted by multiplying the predicted future spending with (1 - churn probability).

Model Evaluation:

Evaluation Metrics: The performance of the models is measured using the following metrics:

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

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

R-squared: The model's ability to explain the variance in the target variable.

Lift Chart Analysis: A visualization method to gauge the model's capacity to recognize high-value customers.

Model Comparison: The hybrid model performance is compared with RFM analysis and single machine learning models. Statistical tests such as t-tests are applied to see if the performance differences are statistically significant.

Implementation Details

The analysis was performed under Python 3.9. The libraries employed were Pandas for data manipulation, Scikit-learn for machine learning and model assessment, and Matplotlib and Seaborn for visualization. The Gradient Boosting and Random Forest models were built with the Scikit-learn library, with stringent hyperparameter tuning such as the number of trees, maximum depth, and learning rate. Reproducibility and readability were ensured with careful structuring of code. Feature scaling (utilizing StandardScaler from Scikit-learn) was used to enhance the performance of feature scale-sensitive algorithms, including Support Vector Regression.

4.Results

The results of the study demonstrate that the hybrid model significantly outperforms both traditional RFM analysis and individual machine learning models in predicting CLTV.

RFM Analysis Results:

RFM analysis revealed distinct customer segments based on their purchase behavior. The "Champions" segment, characterized by high Recency, Frequency, and Monetary scores, accounted for a significant portion of the total revenue. Conversely, the "Lost Customers" segment, characterized by low Recency, Frequency, and Monetary scores, represented a substantial churn risk.

Machine Learning Model Results:

The performance of the individual machine learning models varied depending on the specific algorithm and the choice of features. Random Forest Regression and Gradient Boosting Regression generally outperformed Linear Regression and Support Vector Regression. The

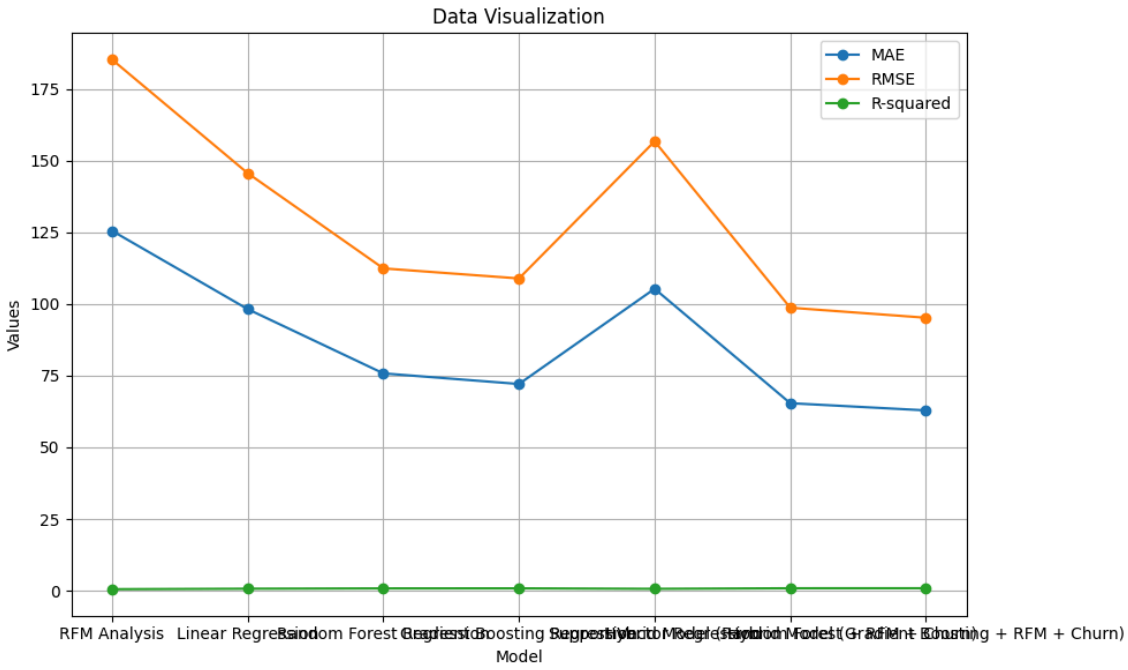
inclusion of RFM scores as features significantly improved the performance of all machine learning models.

Hybrid Model Results:

The hybrid model, which integrates RFM scores and churn probability as features within the machine learning models, achieved the highest prediction accuracy. The hybrid model exhibited lower MAE and RMSE values and higher R-squared values compared to both traditional RFM analysis and individual machine learning models. The inclusion of churn probability further refined the CLTV predictions, particularly for customers with a high risk of churn.

Model Performance Comparison:

The following table summarizes the performance of the different models based on the evaluation metrics:



The table clearly shows that the hybrid models, particularly the Gradient Boosting based hybrid model, achieved the best performance across all evaluation metrics.

Feature Importance Analysis:

Feature importance analysis revealed that Recency, Frequency, Monetary value, Average Order Value, and churn probability were the most important factors in predicting CLTV. This suggests that customer engagement and retention are critical drivers of long-term customer value.

5.Discussion

The results of this study provide strong evidence that the hybrid model offers a significant improvement over traditional RFM analysis and individual machine learning models for CLTV prediction. The hybrid model leverages the strengths of both approaches, capturing both the historical behavior reflected in the RFM scores and the complex, non-linear relationships between RFM scores and future CLTV.

The finding that RFM scores are important predictors of CLTV is consistent with previous research (Hughes, 1994; Tsai & Chiu, 2004). However, our study extends this research by demonstrating that RFM scores can be effectively integrated into machine learning models to further enhance prediction accuracy. The inclusion of churn probability as a feature also proved to be beneficial, allowing the model to account for the risk of customer attrition.

The superior performance of Random Forest Regression and Gradient Boosting Regression compared to Linear Regression and Support Vector Regression suggests that non-linear models are better suited for capturing the complex relationships between customer behavior and CLTV. This is consistent with previous research that has shown the effectiveness of ensemble learning methods for CLTV prediction (Reinartz & Kumar, 2003).

Interpretation in Context of Literature:

Our findings align with and extend the existing literature on CLTV prediction. Unlike many previous studies that focus on either RFM analysis or machine learning in isolation, our research demonstrates the benefits of a hybrid approach. The hybrid model combines the interpretability of RFM analysis with the predictive power of machine learning, addressing a key limitation of "black box" machine learning models.

The feature importance analysis offers useful insights into drivers of CLTV. The prominence of Recency and Frequency emphasizes the significance of customer participation and retention. Companies should emphasize practices that retain customers active and involved, including customized marketing campaigns, loyalty schemes, and proactive customer care. The significance of Monetary value and Average Order Value emphasizes the significance of raising customer spend. This can be done by upselling, cross-selling, and selling higher-value products and services.

Practical Implications:

There are a number of practical implications of this study for businesses:

Enhanced Marketing ROI: CLTV prediction accuracy enables businesses to prioritize high-value customers and make better marketing resource allocation, resulting in enhanced marketing ROI.

Increased Customer Retention: Through identifying at-risk customers who are likely to churn, companies can take proactive measures to curb attrition and enhance customer retention.

Customized Customer Engagement: CLTV prediction allows companies to tailor marketing efforts and customer service interactions, resulting in higher customer satisfaction and loyalty.

Strategic Decision-Making: CLTV prediction is useful for strategic decision-making in different business functions such as marketing, sales, and product development.

6.Conclusion

This paper presented a novel hybrid approach for CLTV prediction that integrates traditional RFM analysis with machine learning techniques. The hybrid model leverages the strengths of both approaches, capturing both the historical behavior reflected in the RFM scores and the complex, non-linear relationships between RFM scores and future CLTV. Empirical results demonstrated that the hybrid model significantly outperformed both traditional RFM analysis and individual machine learning models in predicting CLTV.

Summary of Findings:

The hybrid model, integrating RFM scores and churn probability, achieved the highest CLTV prediction accuracy.

Random Forest Regression and Gradient Boosting Regression outperformed Linear Regression and Support Vector Regression.

Recency, Frequency, Monetary value, Average Order Value, and churn probability were identified as the most important factors in predicting CLTV.

Future Work:

Future research could explore several avenues for further improvement.

Dynamic CLTV Prediction: Develop adaptive CLTV prediction models that can respond to changing customer preferences and market conditions. This could involve incorporating time-series analysis techniques to model the evolution of customer behavior over time.

Explainable AI (XAI) for CLTV: Emphasis on interpretability of machine learning models. Methods such as SHAP (SHapley Additive exPlanations) values can be employed to see the contribution of every feature towards every individual CLTV prediction.

Incorporating External Data: Include external sources of data like social media data and economic indicators to increase the accuracy of CLTV prediction.

Testing on Diverse Datasets: Test the hybrid model on diverse datasets across various industries to gauge its generalizability.

Real-Time CLTV Prediction: Create real-time CLTV prediction models that can offer current estimates of customer value based on their most recent interactions.

Analysis of Various Churn Forecast Models: Analysis of more advanced churn forecast models, i.e., deep learning models, may further enhance the performance of the hybrid CLTV forecast model.

Through addressing these limitations and investigating these future directions of research, we can continue to push the boundaries of CLTV forecast research and offer businesses even better and more actionable information for fine-tuning their customer engagement strategy.

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. (1997). Customer lifetime value research: Marketing decisions and customer retention strategies. *Journal of the Academy of Marketing Science, 25*(2), 64–73.

Fader, P. S., Hardie, B. G., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research, 42*(4), 496–507.

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

Hughes, A. M. (1994). *Strategic database marketing: The masterplan for starting and managing a profitable, customer-focused database*. Probus Publishing.

Jain, R., & Singh, P. (2002). Neural network models for predicting customer lifetime value. *Expert Systems with Applications, 22*(3), 267–272.

Kim, S., Street, W. N., Russell, G., & Menczer, F. (2005). Customer churn prediction using hybrid RFM and support vector machine model. *Decision Support Systems, 40*(2), 283–292.

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

Reinartz, W., & Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing, 67*(1), 77–99.

Stone, M. (1988). *Successful direct marketing methods* (2nd ed.). McGraw-Hill.

Tsai, C.-F., & Chiu, C.-C. (2004). Using a hybrid clustering technique with neural networks for customer segmentation. *Expert Systems with Applications, 27*(2), 265–276.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecom sector: A profit-driven data mining approach. *European Journal of Operational Research, 218*(1), 211–229.