# Research Review on Customer Life Cycle Value Prediction Based on RFM Analysis and Ensemble Learning

Yuhe Bai
School of Science, Zhejiang Sci-Tech University, Hangzhou, 310018, China
b13354561130@163.com

## ABSTRACT

Accurate prediction of customer lifetime value (CLV) is a fundamental point relied upon for optimizing resources and formulating marketing plans. The traditional RFM method can illustrate the general characteristics of customer behavior, but it is difficult to handle nonlinear problems; a single ensemble learning method, although highly accurate, lacks interpretability. Therefore, this paper combines RFM and ensemble learning to examine their integrated form in terms of CLV, discussing the integration methods of input features, customer classes, and Stacking separately, and concentrating on analyzing the applicable scenarios and actual effects of random forest, LightGBM, and Stacking models. This also points out that existing research still has common deficiencies in RFM time window adaptation, customer value evolution, and model transparency, and proposes the possibility of improving from dynamic features, multi-model cooperation, and interpretability perspectives, providing a reference for subsequent research.

**Keywords:** Customer Lifetime Value; RFM Analysis; Stacking; Random Forest LightGBM

## 1. INTRODUCTION

At present, the market competition is extremely fierce, and the cost of acquiring customers is continuously rising. Enterprises need to manage their existing customers as long-term resources. Various analyses related to customer lifetime value (CLV) are the basic methods for judging the long-term contribution of customers and are of great significance for designing targeted marketing strategies and making reasonable arrangements of resources.

However, there are still many practical difficulties in accurately calculating CLV at present. The traditional RFM method divides customers based on consumption intervals, consumption frequencies, and consumption totals, and is simple to calculate and well-explained, but it only performs static weighted processing of each variable and is difficult to reflect the non-linear and time-series characteristics of customer behavior, making the prediction accuracy unreliable. After introducing machine learning, integrated methods such as random forests, LightGBM, and Stacking have improved the prediction accuracy due to their strong nonlinear fitting capabilities, but their interpretability is relatively weak, and enterprises cannot find the exact behaviors that drive customer value and face obstacles in practical application.

Based on RFM features and combined with ensemble learning is the current key topic for CLV prediction. There are many domestic and foreign literatures that have analyzed the possibility of combining the two. The proposed solutions can be roughly classified into three categories: feature input, customer segmentation, and Stacking hierarchical integration. However, current research still has shortcomings: the RFM time window does not have the ability to dynamically adjust, customer segmentation does not provide explanations for the changes in value over time, and the base learners or meta-features used in Stacking lack optimization ideas for quantitative analysis. Therefore, this paper systematically reviews various fusion methods of RFM and ensemble learning, compares the advantages and disadvantages of various models, points out the common problems of existing work, and makes prospects for feasible future research.

## 2. REVIEW METHOD

To systematically synthesize the research on the integration of RFM analysis and ensemble learning in prediction tasks, this paper adheres to the retrieval and screening principles of systematic literature reviews.

1) Database and Retrieval Time: This study uses the core databases of CNKI, Web of Science, IEEE Xplore, and Google Scholar. The retrieval time range is from 1987 to 2025. Academic translation.

2) Search keywords: “Customer Lifetime Value”, “CLV”, “RFM”, “Customer Segmentation”, “Random Forest”, “LightGBM”, “Stacking Ensemble Learning”, “Customer Churn Prediction”.
3) Inclusion criteria for the literature: (1) The research topic involves the combination of RFM indicators and ensemble learning methods; (2) The research objective is CLV prediction, customer value classification, customer churn prediction or purchase intention prediction; (3) Published in peer-reviewed journals, conference proceedings or high-quality dissertations; Excluding pure theoretical derivations, those without empirical experiments, and conference abstracts.
4) Screening process and quality assessment: The initial search yielded approximately 280 relevant articles. After removing duplicates and conducting preliminary screening of titles and abstracts, 140 irrelevant or duplicated studies were excluded; further reading was conducted for a second screening, and finally, 25 core articles were included. The main references cited in this paper are all from the above-mentioned included literature database. The quality assessment focuses on the scale of empirical data in the research, the rationality of the model comparison design, and the reproducibility of the experimental conclusions.

## 3. EVOLUTIONARY PATH OF CLV PREDICTION METHODS

The methods for predicting customer lifetime value have undergone an evolution process from probability statistical models to RFM analysis and then to ensemble learning. Each method iteration aims to address the shortcomings of the previous stage, but also has its limitations. This section will systematically review the theoretical basis and inherent limitations of the three types of methods, providing a reference framework for the discussion of the integration model in the following paragraphs.

### 3.1 Probability and statistics modeling method

The predictive research of CLV is based on probability statistics. The basic idea is to convert the historical purchase facts of customers into the probabilities of repeat purchase or churn, and then calculate the long-term value based on the average transaction value. Schmittlein et al. [1] proposed the Pareto/NBD model, which regards customer consumption as a Poisson process and customer churn as a geometric event. This model provides a complete probabilistic treatment for the prediction of non-contractual customer CLV. Fader et al. [2] based on this proposed the BG/NBD model, which does not strictly limit the form of customer churn, making parameter estimation simpler. Gupta et al. [3] summarized the relevant models and pointed out that the most recent consumption should be regarded as a more important judgment basis. Among domestic scholars, Chen Mingliang [4,5] conducted empirical analysis and prediction attempts on customer life cycle profits earlier. The logical reasoning of the probability model is clear, and the parameters have practical significance. However, the distribution form cannot be freely selected. If customer behavior violates the assumptions, it will lead to prediction errors. In addition, the used data only includes the two states of repeat purchase and churn, and information such as consumption amount has not been included. This point is exactly the problem that the subsequent RFM method needs to solve.

### 3.2 RFM Analysis Method

The RFM model extracts three types of customer behavior indicators to determine customer value. The recent consumption interval R measures customer activity, the consumption frequency F measures customer stickiness, and the total consumption amount M measures customer consumption contribution. By grading and scoring these three indicators and dividing them into customer value echelons, the life cycle value of customers can be quickly estimated.

The foreign scholar Hughes [6] was the first to use RFM as an analysis standard and thereby proposed a three-dimensional customer stratification method. Subsequent works have supplemented or corrected this approach based on this thinking. Gupta et al. [7] applied the RFM indicators to new scenarios and proved that the proposed method is applicable to multiple types of customers. In domestic research, some scholars have reweighed the significance of indicators in combination with the actual operation mode of e-commerce and retail, making the customer value judgment closer to the domestic consumption reality. Thus, various local improved RFM models [8, 9, 10] have been developed.

The RFM model requires relatively little investment and takes into account multiple dimensions of information such as consumption frequency and consumption amount. It compensates for the shortcomings of probability models. However, this model calculates indicators based on a fixed period, making it difficult to analyze the temporal evolution of customer behavior. Moreover, the weights of the indicators are determined manually, which is highly subjective and cannot fully utilize the nonlinear connections contained in each indicator. The prediction results have obvious limitations. These limitations provide space for the introduction of ensemble learning.

### 3.3 Ensemble learning method

The basic idea of ensemble learning is to combine multiple base models for joint prediction to reduce the error of a single model [11]. Random forest uses a large number of decision trees and is suitable for handling high-dimensional information; LightGBM has high training efficiency and low memory usage, and can also obtain accurate solutions in large samples; The Stacking method is designed in a two-layer structure, where each base model gives a prediction, and the upper layer summarizes and analyzes the results obtained from various algorithms, thereby better leveraging the advantages of different algorithms.

Taherkhani et al. [12] conducted a comparative study, proving that tree-based ensemble methods are more suitable for handling complex behavioral phenomena than linear methods. Dey et al. [13] concluded that tree ensemble models have the most accurate predictions. Kasula et al.'s [14] PRISMA systematic review listed Stacking and hybrid ensemble methods as recent trends. Although ensemble learning has strong nonlinear fitting capabilities and can significantly improve the accuracy of CLV prediction, the internal calculation logic of the model is opaque, and it is impossible to clearly explain the impact of each indicator on customer value. There is a black box defect, and it is difficult to directly support business decision-making of enterprises.

## 4. THE INTEGRATION MODEL OF RFM AND ENSEMBLE LEARNING

The key to integrating RFM with ensemble learning lies in incorporating the RFM metrics into the ensemble learning framework, enabling the model to balance both fitting accuracy and business interpretability. To achieve this goal, existing research has mainly developed three integration paths: directly inputting RFM as features, re-modeling based on RFM customer segmentation, and using the Stacking framework for multi-level integration. This section will respectively elaborate on the core ideas, implementation methods, and applicable conditions of each integration mode.

### 4.1 Feature input mode

The feature input mode is the most direct, where the RFM indicators are directly used as input features and passed into the integrated learning model. The model itself analyzes the nonlinear relationships among R, F, and M. The research under this path mainly focuses on directly predicting CLV, while also covering two derivative tasks: warning of customer churn and discrimination of purchase intentions. These two types of derivative studies can respectively provide indirect reference basis for CLV modeling from the perspectives of customer churn risk and potential consumption benefits.

Ding Junmei et al. [15] in the prediction of telecom customer churn input the derived RFM features into the random forest. The experiment shows that the model predicts stably in medium-sized data, verifying the feasibility of this mode. Zou Peng et al. [16] input the RFM features into LightGBM to complete the CLV classification prediction for financial customers. It was found that LightGBM has significantly higher accuracy in processing large-scale data than traditional methods. Taherkhani et al. [12] input RFM into multiple models for comparison, and found that under the same features, the accuracy of different algorithms varies significantly, which helps in model selection.

This mode is applicable to scenarios where customer consumption behaviors vary little and the data scale is medium-sized. It has fewer preprocessing steps and is convenient to operate. However, the time window used by RFM is fixed, and only three types of original variables are input, which cannot constitute rich features. Existing extensions are mostly targeted at specific scenarios and lack a general method.

### 4.2 Customer segmentation model

The basic idea of customer segmentation is to first group customers based on the RFM indicators, and then build models separately for each group to reduce the interference between the behaviors of customers of different values. In practice, usually the K-means clustering algorithm is used to cluster the three-dimensional RFM indicators first to complete customer segmentation; then the obtained clustering information is used to split the data, and integrated learning models are trained separately for each subset. This path encompasses various tasks such as CLV value classification, customer churn prediction, and paid conversion prediction. All types of research are based on RFM clustering to reduce the heterogeneity of customer consumption behaviors. The customer group divisions obtained through clustering can provide support for subsequent calculations of customer lifetime value.

Zhou Huan et al. [9] conducted stratification of retail customers based on RFM and confirmed the effectiveness of RFM clustering in retail scenarios with significant differences in consumption frequency. Cheng Dong et al. [17] compared the performance of statistical methods and machine learning in customer segmentation tasks and pointed out that machine learning is superior to statistical methods in handling nonlinear problems. The proposed approach is conducive to independent modeling after clustering. Li Meiyu et al. [11] integrated K-means based on RFM clustering to build a random forest and experiments showed that customer behavior patterns within the same group were similar, and the prediction accuracy of clustering was higher than the overall prediction. Some other scholars have employed RFM combined with K-means to achieve customer stratification [18]. However, they only rely on fixed-cycle data to complete static grouping, lacking analysis of the temporal evolution of customer value, and can only be used as a preprocessing step for CLV modeling.

This model is applicable to scenarios where customer groups are significantly different and differentiated marketing is required. However, clustering is based on static RFM and cannot fully reflect the characteristic that customer value changes over time [19], and the number of clusters is set based on empirical judgment, which is prone to overfitting when dealing with small samples or uneven sample distribution.

### 4.3 Stacking Multi-level Integration Model

The core idea of multi-layer integration is to set multiple differentiated base learners at the first layer to make predictions separately, and then the meta-learner combines the output results of each base model to give the final prediction. This framework can be directly applied to the regression prediction of customer lifetime value, and is also often set up in classification scenarios such as customer churn and consumption behavior discrimination. The indicators such as customer retention and repurchase probability output by the classification model can also serve as intermediate features and be input into the model in layers to assist in the derivation of CLV results.

Asadi Ejgerdi and Kazerooni [20] constructed a stacked integration including Random Forest and LightGBM. The experiments showed that it outperformed the single comparison method in multiple indicators, confirming that different types of base models can form complementary information and effectively improve the overall prediction effect of the model. Cai Jixian's framework [21] has an accuracy improvement of about 8% compared to a single model, also confirming the complementary characteristic, but did not explain its source. Wang Yulin et al. [22] applied cost-sensitive Stacking to customer churn prediction, exploring the feasibility issues involved in imbalanced data.

This model has the best prediction accuracy in scenarios with sufficient data and complete features. However, it faces three challenges: first, the base learners lack standardized screening criteria; second, the construction of meta-features lacks systematic comparison; third, the computational complexity is higher than that of a single model, and the multi-layer nesting exacerbates the black box problem, with the explainability being the weakest among the three fusion modes.

## 5. COMPARISON AND ANALYSIS OF THREE TYPES OF INTEGRATION MODES

The previous text respectively elaborated on three integration paths of RFM and ensemble learning, and each of them has its own emphasis in the modeling logic. However, in practical scenarios, researchers and enterprises often find it difficult to determine the optimal integration mode in terms of performance. This section will compare the three types of integration modes from three dimensions: prediction accuracy, data adaptability, and model interpretability, in order to provide a reference basis for mode selection in different business scenarios.

### 5.1 Comparison of Prediction Accuracy

Based on the current empirical results, the Stacking method has the highest accuracy but the gain obtained is limited. LightGBM performs significantly better than Random Forest when dealing with large-scale data, while the latter is better at handling small and medium-scale data. The experiment mentioned by Ding Junmei et al. [15] also confirmed that Random Forest has more stable predictions in small and medium-scale datasets, and in the scenario of predicting customer churn in the aviation industry, the AUC of this fusion method reached 0.92 [23]. However, Random Forest lacks iterative error correction capabilities and the fitting effect has an upper limit. Zou Peng et al. [16] input RFM features into LightGBM to complete the prediction of financial customer CLV classification, and found that the accuracy of LightGBM when processing large-scale data is far superior to that of Random Forest. Stacking theoretically has the highest accuracy. Asadi Ejgerdi and Kazerooni [20] constructed a stacked ensemble and found that $R^2$ could be increased by 3-5%. Cai Jixian [21] used the Stacking framework to predict the purchasing behavior of auto parts customers, and also concluded that the prediction accuracy of the integrated model is higher, but Haddadi and Hamidi [24] found in the B2B SaaS scenario that improper combination of base models may cause noise and reduce accuracy. In summary, the accuracy gap among the three types of models is not overwhelming. Whether the gain of Stacking has commercial significance needs to be judged in combination with costs.

### 5.2 Data adaptability comparison

The three types of models have significantly different requirements for data conditions. Random Forest has the lowest requirement for data quality, while LightGBM is highly effective in handling large samples but prone to overfitting in small samples. Stacking has the highest requirements for data volume and features. Random Forest makes the fewest assumptions about the input distribution and can better cope with scenarios with multiple features or many missing values. LightGBM can directly handle categorical features without additional encoding, but high-dimensional small samples are prone to overfitting. Dey et al.'s study [13] on e-commerce CLV indicates that the adaptability of Stacking depends on the diversity of the base models combination. When the data volume is moderate, there is little difference between Stacking and a single model, and simple models are more practical.

### 5.3 Interpretability comparison

Interpretability is the most prominent difference among these three types of models. Random Forest can output the ranking of feature importance and the corresponding contribution, but it cannot analyze the mutual influence between features [3]. LightGBM can trace the single sample reasoning process [19], but it is difficult to apply it to actual business when the path is too long in high-dimensional scenarios. Stacking has the weakest interpretability, and the meta-learner only takes the numerical results of the base models as input, making it difficult to trace the corresponding relationship between the initial variables and the predictions [9]. Akter et al. [25] proposed an explanation scheme based on SHAP, but SHAP is an approximate aggregation rather than a complete logical decomposition, and there are limitations in terms of stability.

## 6. CURRENT MAIN ISSUES AND PROSPECTS

Combining RFM features with ensemble learning for customer lifetime value prediction can effectively improve the prediction accuracy compared to traditional single models. However, existing studies still have significant shortcomings in the construction of temporal features, dynamic clustering, fusion modeling norms, and model interpretability.

(1) The adaptive issue of the RFM time window: All existing studies calculate the RFM indicators using fixed-length windows, and the window length is set based on experience. A too short window will underestimate the value of low-frequency customers, while a too long window will weaken the changes in customers' recent behaviors, both of which are prone to cause misjudgments. There is no clear quantitative criterion for the selection of window step size and attenuation coefficient. In the future, based on the distribution of customer consumption frequency, an automatically adjustable window can be designed, and a more reasonable RFM statistical period can be obtained.

(2) The problem of depicting the dynamic migration of customer value: The current RFM segmentation model mostly adopts a single static clustering approach. The labels assigned to customer segments do not change in response to their consumption patterns, and over time, errors accumulate. Subsequently, a rolling time-series method could be considered. By updating the RFM features and customer segments at regular intervals, the grouping can be made consistent with the customer's consumption at each stage.

(3) The systematic optimization problem of the Stacking framework: The current implementation schemes of Stacking vary significantly. The combination of base learners relies on subjective experience, and the meta-features only take the model's prediction results, rarely integrating the actual business significance of the original RFM indicators or clustering categories; most studies ignore the data's temporal characteristics and conduct cross-validation randomly. In the future, a unified Stacking process should be established, specifying the processing methods for base models, meta-features, and time series cross-validation.

(4) The issue of improving model interpretability: Integrated models generally have the problem of being "black boxes", and the multi-layer stacking method has no direct connection with the features and predictions of the original RFM. Current research generally aims for high prediction accuracy, and the application of interpretation tools such as SHAP and LIME in Stacking customer value prediction is relatively rare. To make model evaluation take into account interpretability, a hierarchical attribution system should be established, where the meta-model, sub-model, and the initial RFM features are examined separately, and their contributions are explained at different levels to form a complete source of features.

## 7. SUMMARY

This paper reviews the customer life cycle value prediction method that integrates RFM with ensemble learning.

(1) At the feature construction level of RFM, the time windows used for each indicator of RFM were analyzed. The fixed windows were compared with customer behaviors, and it was pointed out that the fixed windows were difficult to adapt to the transaction frequencies of different customers.

(2) At the customer group modeling level, the static clustering thinking based on RFM was summarized, and the basic principle of this method was explained, which is to improve the prediction effect by reducing behavioral variability.

(3) At the ensemble integration modeling level, the single ensemble model and the Stacking multi-layer integration method were introduced respectively. The advantages and disadvantages of the three integration modes were compared from three aspects: prediction accuracy, data adaptability, and interpretability.

Customer life cycle value prediction is the key to enterprise precise management. The combination of RFM and ensemble learning can effectively improve the value prediction effect. Existing research has achieved many theoretical results on static datasets, but when facing real dynamic business scenarios, it still faces many challenges.

## REFERENCES

[1] D. C. Schmittlein, D. G. Morrison, and R. Colombo, “Counting your customers: Who are they and what will they do next?,” Manage. Sci., vol. 33, no. 1, pp. 1–24, January 1987.
[2] P. S. Fader, B. G. Hardie, and K. L. Lee, “RFM and CLV: Using iso-value curves for customer base analysis,” J. Mark. Res., vol. 42, no.4, pp. 415–430, November 2005.
[3] S. Gupta, D. Hanssens, B. Hardie, et al. “Modeling customer lifetime value,” J. Service Res., vol. 9, no. 2, pp. 139–155, Nov. 2006.
[4] M. Chen, “Empirical research on the change trend of customer life cycle profit,” Stat. Res, no. 6, pp. 40–44, June 2002.
[5] M. Chen, “Study on forecast method of customer full life cycle profit,” Sci. Res. Manage., vol. 24, no. 4, pp. 102–109, July 2003.
[6] A. M. Hughes, Strategic Database Marketing, 3rd ed. New York: McGraw-Hill, 2005.
[7] S. Gupta, D. R. Lehmann, and J. A. Stuart, “Valuing customers,” J. Mark. Res., vol. 41, no. 1, pp. 7–18, February 2004.
[8] M. Zhao and J. Qi, “Research on customer lifetime value based on purchase behavior RFM and review behavior RFMP model,” Stat. Inf. Forum., vol. 29, no. 9, pp. 91–98, September 2014.
[9] H. Zhou, “Implementation process of retail customer segmentation based on RFM analysis model,” J. Jinling Inst. Technol., vol. 24, no. 1, pp. 33–36, March 2008.
[10] A. Shi and J. Zhang, “Research on e-commerce user value classification based on improved RFM model,” Comput. Technol. Develop., vol. 32, no. 12, pp. 123–128, December 2022.
[11] M. Li, Y. Liu, Y. Wang, and Q. Zhu, “Predicting user pay conversion intention based on stacking ensemble learning: Case study of free value-added games,” Data Anal. Knowl. Discov., vol. 8, no. 2, pp. 143–154, February 2024.
[12] L. Taherkhani, “Analysis and optimization of customer lifetime value prediction using machine learning and deep learning models by RFM techniques,” Int. J. Web Res., vol. 8, no. 2, pp. 79–92, 2025.
[13] S. Dey, D. Radha, and V. S. K. Devi, “Predicting Customer Lifetime Value in E-Commerce: A Data-Driven Approach to Enhance Customer Retention Strategies,” in 2025 3rd Int. Conf. Intell. Data Commun. Technol. Internet Things (IDCIoT), Bengaluru, India, 2025.
[14] V. K. Kasula, C. Tumma, and B. Konda, “A Comprehensive Review of Artificial Intelligence Models for Lifetime Value Optimization,” in 2025 2nd Int. Conf. Res. Methodol. Knowl. Manage., Artif. Intell. Telecommun. Eng. (RMKMATE), Chennai, India, 2025.
[15] J. Ding, G. Liu, H. Li, et al., “Application of improved random forest algorithm in telecom customer churn prediction,” Pattern Recognit. Artif. Intell., vol. 28, no. 11, pp. 1041–1049, November 2015.
[16] P. Zou and Z. Chen, “Construction and application of low-value personal customer label model of commercial banks based on LightGBM,” FinTech Time, vol. 31, no. 2, pp. 27–33, February 2023.

[17] D. Cheng, Y. Sun, and W. Xue, “Robust measurement of non-contractual customer lifetime value: Comprehensive calculation of classical methods and machine learning algorithms,” Manage. Rev., vol. 31, no. 4, pp. 83–98, April 2019.
[18] S. Han, “User feature analysis based on RFM model and K-Means algorithm,” 2020.
[19] P. P. Patil, “A data-driven approach to customer segmentation and customer lifetime value prediction in retail and e-commerce,” M.S. thesis, Nat. College Ireland, Dublin, Ireland, 2024.
[20] N. Asadi Ejgerdi and M. Kazerooni, “A stacked ensemble learning method for customer lifetime value prediction,” Kybernetes., vol. 53, no. 7, pp. 2342–2360, July 2024.
[21] J. Cai, “Research on customer purchase behavior prediction based on improved stacking and implementation of customer relationship management system,” M.S. thesis, Zhejiang Normal Univ., Jinhua,
[22] Y. Wang, “Research on radio and television customer churn prediction model based on ensemble learning,” Video Eng., vol. 48, no. 7, pp. 59–66, July 2024.
[23] L. Yang, Z. Bai, and Y. Kou, “Analysis of airline customer churn by random forest algorithm based on RFM model,” Comput. Mod., no. 1, pp. 100–104, January 2021.
[24] A. M. Haddadi and H. Hamidi, “A hybrid model for improving customer lifetime value prediction using stacking ensemble learning algorithm,” Comput. Hum. Behav. Rep., vol. 18, p. 100616, 2025.
[25] J. Akter, A. Roy, S. Rahman, et al., “Artificial intelligence-driven customer lifetime value forecasting: Integrating RFM analysis with machine learning for strategic customer retention,” J. Comput. Sci. Technol. Stud., vol. 7, no. 1, pp. 249–262, 2025.