# RK-W-Stacking: A Hybrid Model Combining Entropy-Weighted RFM-K-Means Clustering and Weighted Stacking Ensemble for User Consumption Prediction

Yaping Zhao
School of Cultural Industry and Business, Jilin Animation Institute, Changchun 130000, China
E-mail: zyplylyz@163.com

**Keywords:** K-means clustering, stacking, feature prediction, RFM, weighted

**Received:** April 30, 2025

*Understanding user consumption characteristics helps predict consumer behavior and design appropriate marketing strategies, which promotes spending and supports market economic circulation. However, existing models have a low recognition rate for user consumption characteristics and limited accuracy in consumer behavior prediction. To this end, a user consumption behavior prediction model was constructed in the research. The main innovation of the research method lies in the weighted reconstruction of the K-means distance calculation logic by feature variance to enhance the robustness of the non-convex data set class. The index weights are dynamically calculated through the entropy weight method to optimize the user value assessment. A Stacking ensemble framework based on error rate weighting is designed to integrate the advantages of XGBoost and random forest-based classifiers. Experimental results showed 96.3% feature recognition accuracy, 0.25% loss rate, and 0.91 AUC, demonstrating strong classification ability. Additionally, long-term consumption prediction experiments show that the proposed model reaches a maximum memory usage of 1170 MB, a maximum response time of 120 ms, while maintains an annual prediction accuracy of over 94.2%. These results indicate that the proposed model achieves high accuracy and adaptability in extracting consumption characteristics and forecasting future consumer behavior. Furthermore, the system exhibits greater stability and faster response times than comparable prediction models, providing new insights for merchants to optimize business strategies and anticipate market trends.*

*Povzetek: Predstavljen je RK-W-Stacking, hibrid za napoved porabe. Je entropijsko utežen RFM, variančno utežen K-means, ki robustno izlušči potrošniške značilnosti ter stabilno napove dolgoročno vedenje v praksi.*

## 1 Introduction

With technological advancements and the continuous improvement of living standards, traditional marketing methods have become less effective in the internet era. They fail to influence customer consumption habits and desires, creating an urgent need for businesses to explore new ways to stimulate spending [1]. Purchasing decisions are typically influenced by individual consumption habits. Predicting user consumption behavior and implementing appropriate marketing strategies can significantly improve business performance [2]. Therefore, summarizing user consumption characteristics and predicting their behavior is crucial for merchants to optimize operations and enhance market capital flow. Current methods for predicting user consumption behavior typically rely on single-model learning, which lacks specificity in extracting user characteristics and results in low overall prediction accuracy. Thus, a new approach is needed to accurately identify user features and efficiently predict consumer behavior [3]. The K-means clustering algorithm, an unsupervised learning method, is simple to implement and highly interpretable. It gourps similar characteristics to uncover hidden patterns in data, making it widely used in feature recognition [4]. Stacking overcomes the limitations of single models by leveraging multiple models to create complementary advantages. As an effective ensemble learning method, it captures different hidden relationships within the same dataset and has demonstrated strong predictive capabilities in image classification and financial stock selection [5]. Therefore, the K-means clustering and Stacking algorithms were optimized in the study, and a model for predicting user consumption characteristics was proposed. The research aims to improve the performance of user consumption behavior prediction through the following core goals: (i) Reconstruct the weight of the RFM index by using the entropy weight method to optimize the discrimination of user consumption characteristics; (ii) Improve the K-means clustering algorithm based on feature variance to enhance the clustering robustness for non-convex datasets; (iii) Design an error-weighted Stacking ensemble framework to integrate the advantages of multi-basis classifiers to improve the prediction accuracy.

## 2 Related works

K-means clustering is an algorithm that represents sample clustering through the squared Euclidean distance. Its core principle is to divide data into multiple clusters and iteratively minimize the total distance. Scholars worldwide have conducted research on this method. For example, Nie et al. raised an iterative reweighted algorithm to optimize K-means clustering. By redefining functions, this method reduced the computation of cluster centers in each iteration. Experimental results demonstrated that the optimized algorithm had a faster convergence speed, confirming its effectiveness and efficiency [6]. Brusco et al. aimed at the problem that the hierarchical clustering relied on by the Walktrap algorithm in the scale of psychological research networks might not provide the optimal solution for the sum-square optimization problem. They used the K-means clustering method (including exact and approximate methods) to replace the hierarchical clustering step in the Walktrap algorithm. The results show that the application of K-means clustering can usually obtain a better sum-square solution, improving the effect of community detection [7]. Chen et al. addressed the mean difference issue in K-means clustering identification by introducing a p-value constraint in cluster assignments. This constraint controlled the Type I error in mean difference testing between clusters. Their study validated the linear relationship between the p-value and Type I error in single-cell RNA sequencing data and demonstrated its effectiveness in accurate computation [8]. Unlike K-means, Stacking is an ensemble learning method that relies on multiple models to create complementary advantages. Many research teams have applied it in predictive modeling. Meharie et al. proposed an ensemble model combining linear regression, support vector machines, and artificial neural networks for highway construction cost prediction. Results showed that this model had a lower prediction error than single ensemble models [9]. Xue et al. introduced a multi-objective evolutionary algorithm with probabilistic Stacking, incorporating a mechanism to guide offspring crossover and mutation in multi-objective genetic algorithms. Experiments demonstrated that the algorithm achieved high classification accuracy while reducing time costs [10].

To predict surrounding rock classification during tunnel boring machine operations, Hou et al. proposed an ensemble classifier that uses a tree-based feature selection method to process operational data. Comparisons with multiple single-class classifiers verified that the ensemble model had superior performance in rock classification prediction and stronger generalization ability for imbalanced small-sample datasets [11]. User consumption behavior characteristics are essential indicators for predicting consumer behavior and evaluating market conditions. Many scholars have extracted and analyzed these characteristics. Tang et al. addressed the issue of low clustering accuracy in user characteristics by introducing a customer feature analysis method based on behavioral segmentation. The study used maximum correlation and optimal redundancy to identify the best features, followed by affinity propagation clustering for analysis. Experimental results showed high accuracy in user classification with strong practical applications [12]. Yang et al. found that existing anomaly detection methods often overlooked abnormal data when analyzing consumption amount characteristics. To address this, they introduced a one-class support vector machine based on a growth model to mitigate the impact of anomalies. Experimental results confirmed that the model effectively handled abnormal data interference [13]. Yang et al. explored the influence of peer effects on daily essential consumption by applying a spatial autoregressive model to estimate the proportions of peer-influenced and autonomous consumption. Findings indicated that households prioritized peer effects in daily expenses but valued autonomy more in social expenditures [14]. As environmental issues become increasingly severe, Saif's research team investigated whether corporate and consumer environmental responsibility strengthened green consumption behavior using structural equation modeling. Their analysis of large datasets found that consumer environmental responsibility did not directly drive green consumption behavior [15]. Khan et al. explored factors influencing green and sustainable consumption behavior by collecting survey data through Google Forms using a Likert 5-point scale. The study applied variance and multiple regression analysis, revealing that environmental awareness significantly influenced sustainable consumption behavior and that there were notable differences between male and female participants [16]. The key information of the relevant research is shown in Table 1.

Table 1: Summary of the main information of related research

| Ref. | Core Method/Contribution | Key Performance/Data |
|---|---|---|
| [6] | Iterative reweighted K-means optimization | Faster convergence (generic datasets) |
| [7] | K-means clustering replaces Walktrap's hierarchical clustering | Better sum-square solutions (psychology networks) |
| [8] | p-value constraint in K-means cluster assignments | Controlled Type I error (single-cell RNA data) |
| [9] | Stacking ensemble (linear regression + SVM + artificial neural networks) | Lower prediction error (highway construction costs) |
| [10] | Multi-objective evolutionary algorithm with probabilistic Stacking | Higher classification accuracy + reduced time cost |

| | | |
|---|---|---|
| [11] | Tree-based feature selection + Stacking ensemble classifier | Superior rock classification for imbalanced samples |
| [12] | Customer feature analysis via behavioral segmentation + affinity propagation clustering | High practical accuracy (electricity consumption) |
| [13] | Growth model-based one-class SVM for anomaly detection | Effective abnormal data handling (campus consumption) |
| [14] | Spatial autoregressive model for peer effects | Quantified autonomous/peer-influenced consumption (rural China) |
| [15] | Structural equation modeling for environmental responsibility effects | No direct green consumption drive (large datasets) |
| [16] | Variance analysis + multiple regression for sustainable behavior | Gender differences in environmental awareness (Likert-scale surveys) |

In summary, existing research has shown that analyzing user consumption behavior characteristics can help predict future consumer actions to some extent. However, challenges remain, such as low prediction accuracy in single-model approaches and weak multidimensional data processing capabilities. Stacking offers strong ensemble learning capabilities, integrating the advantages of different base models. Meanwhile, K-means can cluster complex user data to uncover hidden patterns. Therefore, this study proposes a consumption behavior characteristic prediction model based on K-means clustering and Stacking algorithms. The goal is to extract consumer behavior features accurately and provide effective consumption predictions, thereby enhancing merchants' marketing capabilities and promoting a healthy market economy cycle.

# 3 The user consumption behavior prediction model integrating K-means and stacking

## 3.1 Feature extraction design based on K-means

K-means quickly classifies user consumption behavior features after selecting the number of clusters $k$. However, its clustering performance is sensitive to the choice of $k$ and struggles to converge on datasets with weak features [17]. This study replaces the standard distance metric in K-means with a weighted Euclidean distance based on feature variance. The distance calculation logic is reconstructed, and variance-guided centralization is used to prevent local optima caused by random selection of $k$. The process of the improved K-means algorithm with feature variance weighting is shown in Figure 1.

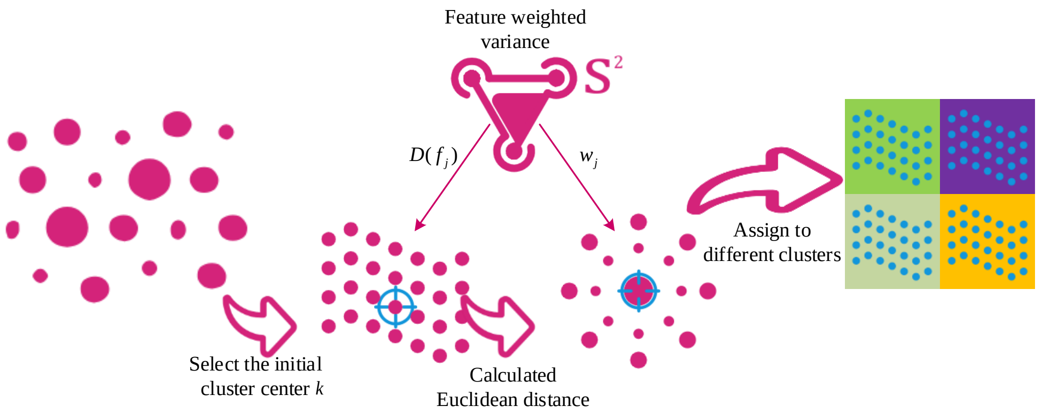



Figure 1: K-means operation flow chart based on feature variance weighting improvement

As shown in Figure 1, the improved K-means selects cluster centers based on feature variance from the dataset. The weighted Euclidean distance is calculated using the assigned weights, and samples are assigned to clusters based on their distance to the cluster centers. Distance calculation and sample assignment repeat until the cluster centers stabilize. The number of clusters k is automatically determined by the contour coefficient. Traverse the candidate k values within the preset range (from k=2 to the square root of the sample size), and perform weighted K-means clustering respectively. The contour coefficient measures the closeness of the sample to the samples of the same cluster and the separation from other clusters. Its value range is [-1, 1]. The closer it is to 1, the higher the cohesion within the cluster and the better the separation between clusters. Select the k value that maximizes the contour coefficient as the optimal number of clusters. Before applying feature variance weighting, the dataset is represented in matrix form, as shown in Equation (1).

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \quad (1)$$

In Equation (1), $m$ and $n$ represent the number of samples and the number of features, respectively. The variance of a specific feature across all samples is calculated as shown in Equation (2).

$$D(f_j) = \frac{\sum(x_j - \overline{x})^2}{m} \quad (2)$$

In Equation (2), $x_j$ represents the eigenvalue of the sample, and $f_j$ represents its feature set. The greater the variance of a sample, the higher the assigned weight. The calculation of sample weights is given in Equation (3).

$$w_j = \frac{D(f_j)}{\sum_{t=1}^{p} D(f_t)} \quad (3)$$

In Equation (3), $p$ represents the number of features in a sample, and $w_j$ denotes the assigned sample weight. The calculation of the weighted Euclidean distance based on these weights is given in Equation (4).

$$d_w(x_i, x_j) = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + \dots + w_p(x_{ip} - x_{jp})^2} \quad (4)$$

In Equation (4), $x_{ip}$ represents a single eigenvalue of sample $x_i$, and the sum of the variance weights of all features is equal to 1, that is, $\sum_{P=1}^{P} w_p = 1$. The characteristic variance in Equation (2) is the global variance of the entire dataset. Since the weight is proportional to the feature variance, features with high variance have a greater influence in distance calculation. This may lead to sensitivity to outlier variance. For example, a feature with an exceptionally high variance may dominate the distance metric, thereby affecting the clustering results. The improved K-means enhances clustering performance on non-convex datasets. To further optimize classification results, this study preprocesses feature selection using the Recency-Frequency-Monetary (RFM) analysis tool before clustering. However, the traditional RFM model lacks differentiation between users with irregular large-amount consumption and users with steady, balanced consumption [18]. To reduce errors caused by time accumulation effects, this study applies the entropy weight method to independently determine the weights of R, F, and M indicators, and the normalized values of these three weighted indicators constitute the feature vectors for clustering. The improved RFM process based on the entropy weight method is shown in Figure 2.

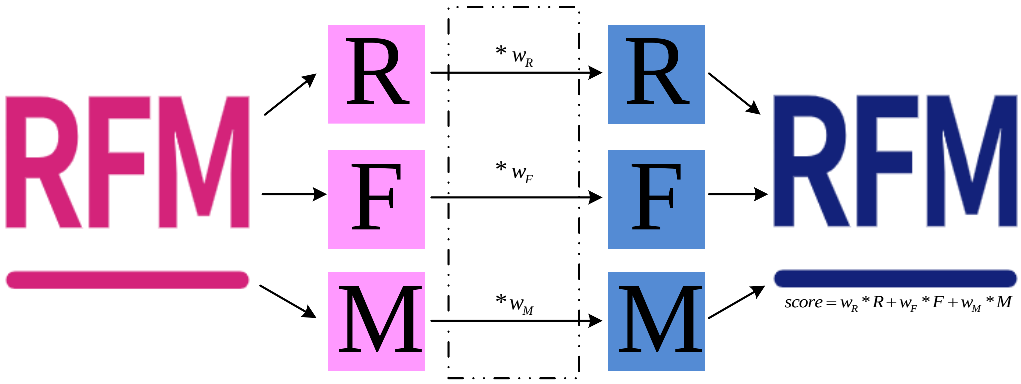



Figure 2: Improved RFM flow chart based on entropy weight method

As shown in Figure 2, the traditional weights of $R$, $F$, and $M$ remain fixed. The entropy weight method is applied to recalculate the indicator weights, allowing different indicators to reflect their varying degrees of influence on customer value. This improves the linear correlation between customer value indicators and the original data. In the traditional RFM model, the number of purchases, total spending, and observation period length are key factors. A higher $R$ leads to a higher $M$, which in turn indicates greater customer value [19]. The calculation of customer value is shown in Equation (5).

$$score = w_R \cdot R + w_F \cdot F + w_M \cdot M \quad (5)$$

In Equation (5), the weights of $R$, $F$, and $M$ are $w_R$, $w_F$, and $w_M$, respectively. Notably, the resulting weighted R, F, M features—not the linear combination in Equation (5): directly from the input feature space for K-means clustering, ensuring multidimensional pattern capture. This study recalculates indicator weights based on information entropy, as given in Equation (6).

$$\begin{cases} p_{rs} = \dfrac{x_{rs}}{\sum_{i=1}^{n} x_{rs}} \\ e_s = -k \sum_{r=1}^{n} p_{ij} \operatorname{In}(p_{rs}) \end{cases} \quad (6)$$

In Equation (6), the $s$-th indicator of the $r$-th sample is represented by $x_{rs}$. The entropy $e_s$ of $s$ determines the weight proportion on customer value, as shown in Equation (7).

$$w_s = \frac{1-e_s}{m-\sum_{s=1}^{m} e_s} \tag{7}$$

In Equation (7), $w_s$ represents the weight of the $s$-th indicator. By applying the RFM analysis tool, data complexity is reduced, making clustering more efficient. A feature extraction algorithm combining RFM and K-means is developed and named RFM-K-means. The RFM-K-means feature extraction process is shown in Figure 3.

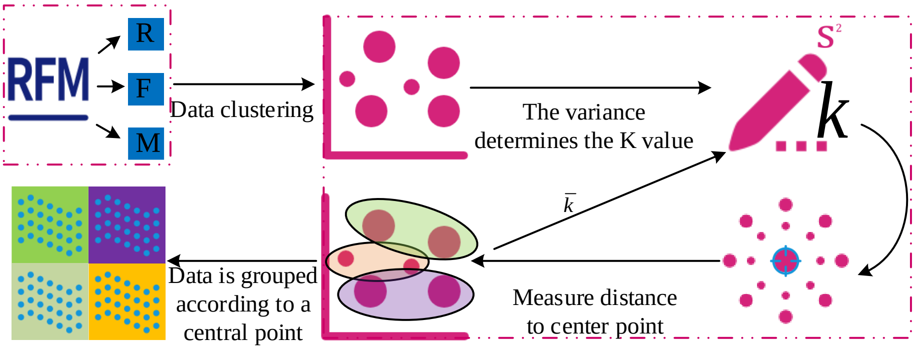



Figure 3: RFM-K-means consumption feature extraction flow chart

As shown in Figure 3, RFM-K-means uses the reweighted R, F, M values (derived from entropy-based weighting) as multidimensional input features for the improved K-means clustering, resulting in stable and accurate consumption user features. To minimize the impact of data units on output results, data normalization is performed before clustering. Both $M$ and $F$ are positively correlated with user value, whereas $R$ is negatively correlated [20]. The normalization process is shown in Equation (8).

$$Z_{rs} \frac{x_{rs} - \min(x_{1s}, x_{2s}, ..., x_{ns})}{\max(x_{1s}, x_{2s}, ..., x_{ms}) - \min(x_{1s}, x_{2s}, ..., x_{ns})} \tag{8}$$

In Equation (8), the value range is denoted by $r = 1, 2, ...n; s = 1, 2, ...m$, and the number of samples and features are represented by $m$ and $n$. The normalization process for negative indicators is given in Equation (9).

$$Z_{rs} \frac{\max(x_{1s}, x_{2s}, ..., x_{ms}) - x_{rs}}{\max(x_{1s}, x_{2s}, ..., x_{ms}) - \min(x_{1s}, x_{2s}, ..., x_{ns})} \tag{9}$$

Equation (9) normalizes consumption frequency $R$, reducing the impact of unit differences on clustering performance. The reweighted RFM values capture more potential consumption patterns, leading to more accurate feature extraction after K-means clustering. The design of the normalization strategy is based on the inherent characteristics of the RFM model. $R$ is used as a negative indicator because the longer the recent consumption time interval of the user, the higher the risk of churn and the lower the customer value. Therefore, it needs to be converted into a feature that is positively correlated with user value through reverse normalization. The consumption amount $M$ usually shows a right-skewed distribution, while the consumption frequency $F$ is relatively evenly distributed. Min-Max normalization retains the skewed features of $M$ through linear compression to avoid the dilution of the features of high-consumption users. Meanwhile, since F is evenly distributed, the same normalization method can effectively preserve its linear relationship. The research uses the entropy weight method for pre-weighting and combines Min-Max normalization to constrain the numerical range to [0, 1] while retaining the distribution characteristics of the data, ensuring the robustness of subsequent clustering against skewed data.

## 3.2 Construction of a user consumption behavior prediction model with Stacking

Although RFM-K-means offers refined clustering results for user consumption features, it cannot predict future behavior. A predictive model is needed to establish the relationship between features and outcomes. Stacking ensemble learning excels in predictive tasks, and its second-level output aligns better with practical applications. Moreover, in long-term predictions, seasonal or unexpected changes in user behavior may disrupt periodic patterns and introduce noise into the time series, resulting in a reduction in the intensity of long-term predicted patterns. Therefore, this study integrates RFM-K-means with Stacking to construct a forecasting model called RK-W-Stacking. Stacking, as a layered ensemble algorithm, consists of multiple base classifiers and a meta-classifier. Its basic operational process is shown in Figure 4.

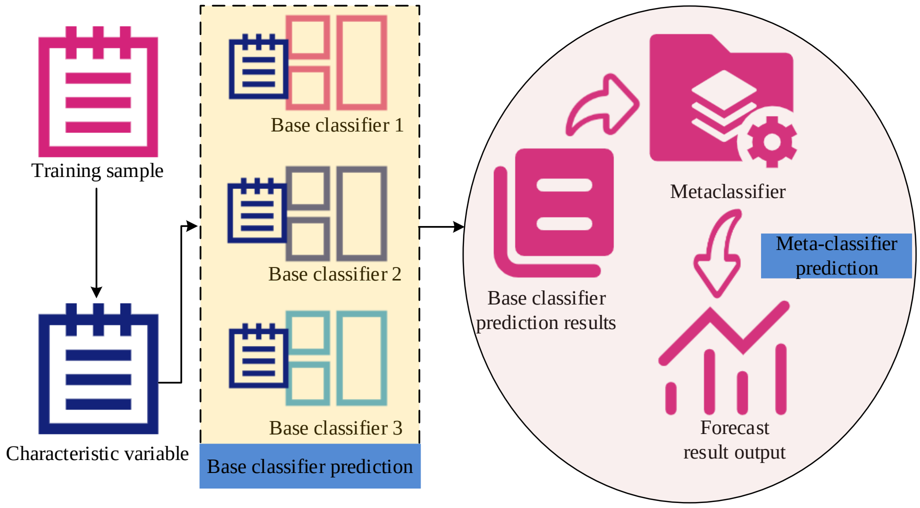



Figure 4: Schematic diagram of multiple classifiers stacked in the Stacking algorithm

In Figure 4, training samples are first processed by multiple base classifiers, each producing a prediction. These predictions serve as secondary training samples for the meta-classifier, which generates the final prediction. The performance of Stacking largely depends on the choice of base classifiers. This study uses eXtreme Gradient Boosting (XGBoost), Random Forest (RF), and Gradient Boosting Decision Tree (GBDT) as base learners, with Logistic Regression Model (LRM) as the meta-learner. The GBDT expression is shown in Equation (10).

$$F_a(x) = \sum_{i=1}^{m} T(x;\theta_a) \tag{10}$$

In Equation (10), $a$ represents the number of prediction rounds, and $T(x;\theta_a)$ denotes the number of weak learners. GBDT improves prediction accuracy by learning residuals from each classifier. XGBoost follows a similar principle but minimizes the loss function during the splitting process. The reduced loss function is given in Equation (11).

$$\max \frac{1}{2}\frac{G_L^{\ 2}}{H_L+\lambda} + \frac{1}{2}\frac{G_R^{\ 2}}{H_R+\lambda} - \frac{1}{2}\frac{(G_L+G_R)^2}{G_L+H_R+\lambda} - \lambda \tag{11}$$

In Equation (11), $L$ represents the loss function, and $\lambda$ is the regularization term. Directly using the outputs of base classifiers to train the meta-learner poses a risk of overfitting. Assigning different weights to base classifiers based on their performance helps mitigate overfitting. Therefore, before the second Stacking training, this study applies error-based weighting to the base classifiers' outputs. The weighted Stacking ensemble with XGBoost, RF, GBDT, and LRM is constructed as shown in Figure 5.

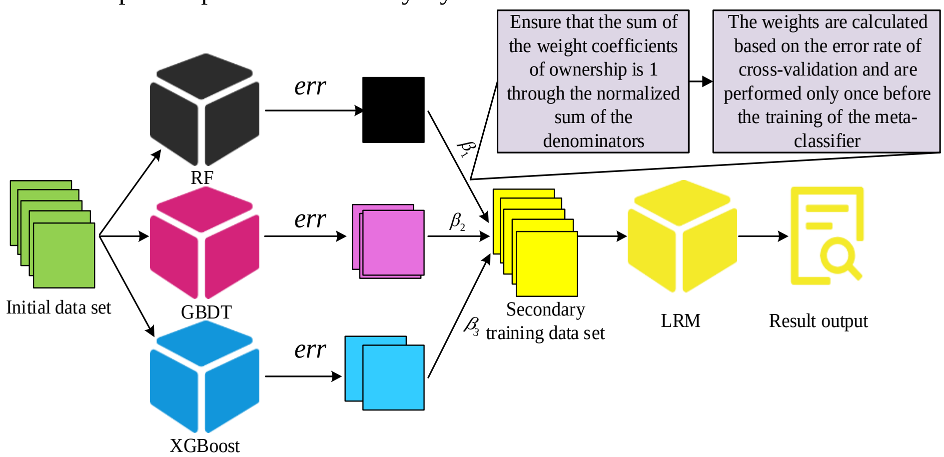



Figure 5: Construction diagram of the weighted set learner of Stacking

As shown in Figure 5, error-based weighting is applied to the base classifiers' predictions. The weight distribution is determined based on the misclassification rates obtained from initial evaluations, improving the accuracy of the LRM final prediction. The error rate of base classifiers is calculated as shown in Equation (12).

$$err = \frac{1}{c}\left(\sum \frac{n}{N}\right) \tag{12}$$

In Equation (12), $err$ represents the error rate, and $n$ and $c$ denote the number of misclassified samples and the number of cross-validation folds, respectively. The weight coefficient for training the meta-classifier is calculated as shown in Equation (13).

$$\begin{cases} \beta = \dfrac{1-err}{err} \\ g = \dfrac{\beta}{\sum \beta} \end{cases} \quad (13)$$

In Equation (13), $\beta$ and $g$ represent the learning weight derived from error rates and the weight coefficient for secondary training. The overall process of RK-W-Stacking for user consumption behavior prediction is shown in Figure 6.

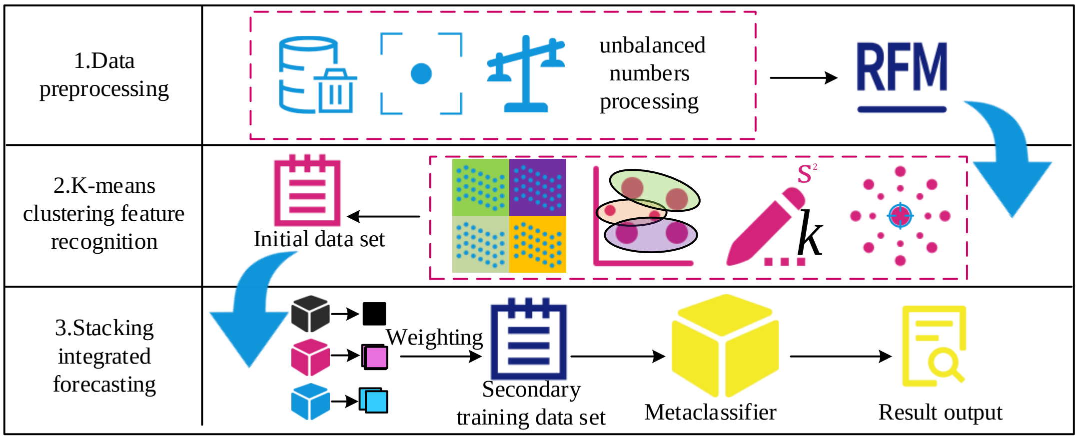



Figure 6: Flowchart of the user consumption feature prediction model of RK-W-Stacking

As shown in Figure 6, the model preprocesses data to reduce the impact of missing values, outliers, and imbalanced distributions on initial samples. RFM indicators enhance feature representation, and the improved K-means algorithm provides clear feature clustering. The clustered features serve as training data for base classifiers, whose weighted predictions act as training data for the meta-classifier, ultimately yielding an accurate prediction. Data balancing is performed using the k-nearest neighbor's method, which generates new samples based on Euclidean distance calculations, as shown in Equation (14).

$$dist(B,C) = \sqrt{\sum_{m}^{n}(b_i - c_i)^2} \quad (14)$$

In Equation (14), the minority class samples are represented by $b$, and other samples by $c$. The generation of new samples based on minority samples is given in Equation (15).

$$\bar{b} = b + rand(0,1) * (\tilde{b} - b) \quad (15)$$

In Equation (15), $\tilde{b}$ is the nearest neighbor of $b$, and $\bar{b}$ represents the newly generated sample. The proposed comprehensive prediction model improves data selection, feature classification, and training efficiency, improving feature extraction capability and prediction accuracy. The model implementation adopts version 4.1.0 of the R language. The key components rely on the following software libraries: Data preprocessing uses tidyverse 1.3.0 and caret 6.0-94 packages; The K-means clustering implements an improved algorithm based on the stats 4.1.0 package; In the base classifier, XGBoost adopts the xgboost 1.6.2 library, with a learning rate of 0.05, a maximum depth of 6, and 150 rounds of iterations set. Random Forest uses the randomForest 4.7-1.1 package, configures the number of feature selections as the square root of the total number of features, and builds 500 trees; The GBDT classifier is implemented through the gbm 2.1.8.1 library, with the optimization parameters being 300 trees and depth 5; The meta-classifier logistic regression implements L2 regularization using the glmnet 4.1-4 package, and the regularization strength is set to 0.01. The preprocessing process consists of three stages: numerical missing values are filled with the median, and typed missing values are filled with the mode. After identifying outliers based on the Z-score threshold 3.0, 1% tailing processing is performed; All features are mapped to the interval [0, 1] by Min-Max normalization. Time features are encoded with cyclic transformations using sine and cosine functions. For the problem of class imbalance, the nearest neighbor oversampling technique of k=5 is adopted, combined with the weighted Euclidean distance metric, to perform random interpolation on the minority class samples until they are balanced in size with the majority class samples.

# 4 Performance analysis of the RK-W-Stacking feature prediction model

## 4.1 Simulation of the user consumption behavior prediction model

To evaluate the performance of the RK-W-Stacking model in predicting user consumption behavior characteristics, the study compared RK-W-Stacking with the XGBoost model, the Support Vector Machine (SVM) model, and the Blending model. The KKBOX music subscription dataset used in the research is a publicly available and reflects users' paid subscription behavior. This dataset is of moderate size and contains approximately 120,000 unique user records, covering subscription activities throughout the entire year from January to December 2021. The dataset provides dozens of raw features (the raw data contains approximately 25 fields), mainly covering basic user attributes, subscription package information, and core consumption interaction behaviors. The key features selected for the study include the order frequency, consumption amount, and time of first consumption, time of last consumption and total consumption amount during the observation period. There is a certain class imbalance problem in the data set. The proportion of active subscribers (high-value users) is usually lower than that of inactive or churn users, which reflects a common phenomenon in online subscription services. To alleviate the impact of this problem on model training, the study applied the oversampling technique based on K-nearest neighbors (K=5) in the preprocessing stage. This dataset demonstrates excellent user diversity, covering user groups with different demographic characteristics and consumption capacity levels, ensuring the generalization ability of the model training results. The data is sourced from the official open data platform of KKBOX, with a time granularity of Daily, precisely recording the date of each subscription transaction. The computing environment was set up with the Windows system, high-speed solid-state storage, 64 GB+ RAM, and an NVIDIA A100 GPU. The development tool was VS Code, all algorithms were implemented in R, and C++ was used for debugging. The learning curves of different basic classifiers are shown in Figure 7.

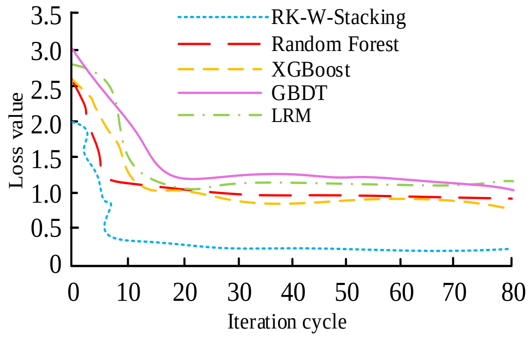



Figure 7: Learning curves of different basic classifiers

As can be seen from Figure 7, the iterative efficiencies of different basic classifiers were close, but there are subtle differences in the iterative situations. The loss value of Random Forest rapidly dropped below 1.2 in the first 7 iterations and eventually slowly decreased to 1.0. The loss value of GBDT rapidly dropped below 1.3 in the first 16 iterations and eventually slowly decreased to 1.1. The loss value of LRM rapidly dropped below 1.2 in the first 10 iterations and eventually slowly decreased to 1.1. It indicated that different basic classifiers can all learn effectively and efficiently. The pre-weighted performance of the basic model and the comparison after Stacking integration are shown in Table 2.

Table 2: Comparison of the performance of the basic model before weighting and after Stacking integration

| Model | Accuracy Before Weighting (%) | AUC Before Weighting | F1-score Before Weighting (%) | Weight in Stacking | Accuracy After Stacking (%) |
|---|---|---|---|---|---|
| XGBoost | 92.1 | 0.85 | 88.3 | 0.38 | 96.3 (↑4.2) |
| Random Forest | 90.5 | 0.82 | 86.7 | 0.35 | 96.3 (↑5.8) |
| GBDT | 91.3 | 0.84 | 87.2 | 0.27 | 96.3 (↑5.0) |
| Weighted Stacking | \ | \ | \ | \ | 96.3 |

As can be seen from Table 2, in the basic model, XGBoost has the best individual performance (accuracy rate 92.1%, AUC 0.85), followed by GBDT, and Random Forest is slightly lower. After error-weighted ensemble, the accuracy rate of the Stacking model significantly increased to 96.3%, which was 4.2 percentage points higher than that of the optimal base model (XGBoost), which verified the complementary advantages of ensemble learning. The weight distribution showed that XGBoost had the highest contribution (0.38), which is consistent with its individual performance, indicating that the error weighting mechanism effectively enhanced the role of the high-performance base classifier. The accuracy rates, loss rates, ROC curves and F1 values of different models are shown in Table 3.

Table 3: Comparison of comprehensive performance of different models

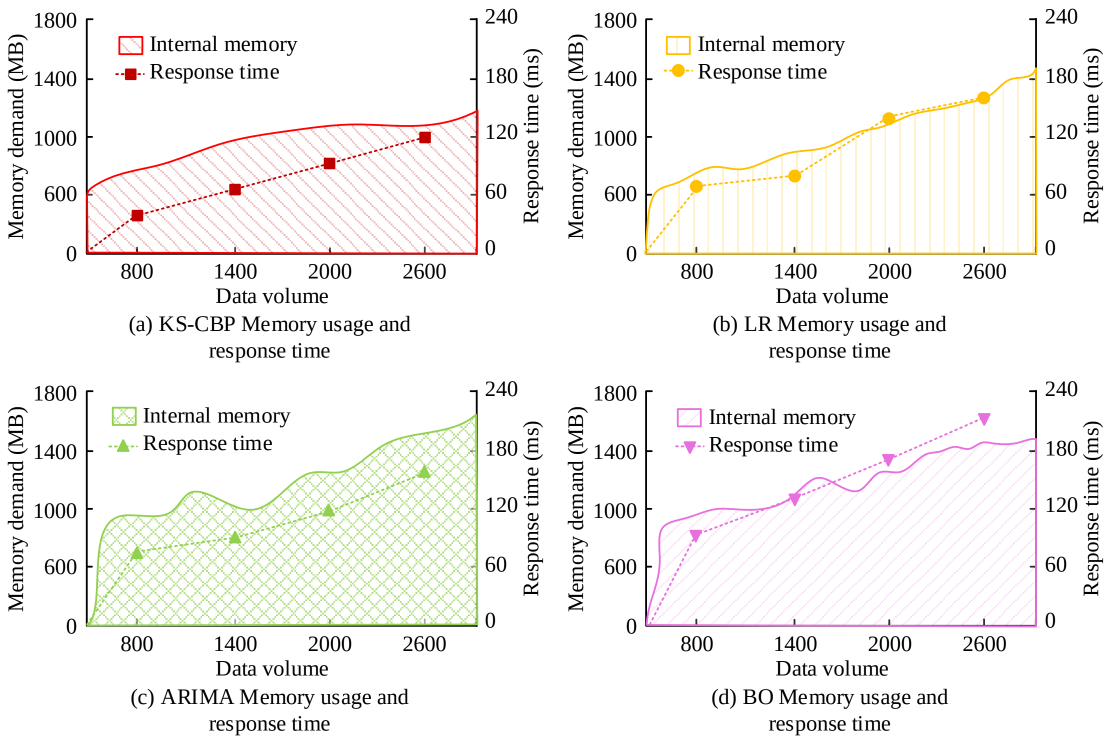
| Model | RK-W-Stacking | XGBoost | SVM | Blending |
|---|---|---|---|---|
| Accuracy (%) | 96.3 | 90.2 | 86.7 | 90.7 |
| 95% CI for Accuracy | [95.5, 97.1] | [89.1, 91.3] | [85.3, 88.1] | [89.6, 91.8] |
| Loss Rate (%) | 0.25 | 1.1 | 1.3 | 1.4 |
| Std Dev of Loss Rate | 0.03 | 0.12 | 0.15 | 0.14 |
| AUC | 0.91 | 0.79 | 0.65 | 0.73 |
| 95% CI for AUC | [0.89, 0.93] | [0.77, 0.81] | [0.62, 0.68] | [0.71, 0.75] |

| | | | | |
|---|---|---|---|---|
| F1-Score (%) | >80 | >40 | >40 | >50 |
| Std Dev of F1-Score | 1.2 | 2.3 | 2.6 | 2 |
| *p*-value (vs RK-W-Stacking) | \ | <0.001 | <0.001 | <0.001 |

As can be seen from Table 3, after stabilization, the accuracy rates of the RK-W-Stacking, XGBoost, SVM and Blending models are 96.3%, 90.2%, 86.7% and 90.7% respectively. The proposed RK-W-Stacking model was significantly superior to other models in terms of accuracy and converges within fewer iterations. RK-W-Stacking maintains a loss rate of 0.25%, while other models required more iterations to converge and had higher loss values. The results showed that RK-W-Stacking had achieved a high accuracy rate and a low loss rate in the recognition of consumption behavior characteristics, and its overall performance is stable. The area under the curve (AUC) of RK-W-Stacking was 0.91, which higher than 0.79, 0.65 and 0.73 of XGBoost, SVM and Blending models respectively, proving its powerful classification ability. The F1 score of RK-W-Stacking remained above 80%, which demonstrated its outstanding ability in complex data recognition. These results indicated that RK-W-Stacking can effectively capture positive examples, exhibit strong classification performance, and maintain a high recognition accuracy even in imbalanced datasets. Meanwhile, statistical analysis showed that the p values of XGBoost, SVM, and Blending models compared with RK-W-Stacking were all less than 0.001, indicating that the conclusion that the performance of the RK-W-Stacking model was significantly better than that of other models is statistically significant. The confidence intervals and standard deviations obtained from statistical analysis also support the accuracy of the research results.

## 4.2 Real-world application analysis of the user consumption behavior prediction model

After verifying the predictive performance of RK-W-Stacking through simulation, the study further examined its practical application value. Using the same dataset and computing environment, the study divided KKBOX's 2021 subscription consumption data into 12 months to extend the prediction period. Additionally, it introduced the Linear Regression (LR), Autoregressive Integrated Moving Average (ARIMA), and Bayesian Optimization (BO) models to compare with RK-W-Stacking in terms of program response, prediction error, and long-term forecasting accuracy. As data samples increased, memory usage and response time of the models are shown in Figure 8.



(a) KS-CBP Memory usage and response time
(b) LR Memory usage and response time
(c) ARIMA Memory usage and response time
(d) BO Memory usage and response time

Figure 8: Changes in response time and memory usage of several models

As shown in Figure 8, memory usage and response time increased to varying degrees as the data sample size grew. When processing 2,600 samples, the memory usage of RK-W-Stacking, LR, ARIMA, and BO was 1, 170 MB, 1, 430 MB, 1, 670 MB, and 1, 450 MB, respectively. RK-W-Stacking saved approximately 14.2% of memory compared to LR. At the same sample size, RK-W-Stacking's response time was 120 ms, significantly lower than the 171 ms, 172 ms, and 232 ms of LR, ARIMA, and BO, respectively. These results demonstrated that RK-W-

Stacking efficiently managed memory while maintaining fast response times, making it a more practical model. The study also analyzed the prediction accuracy of the models for user subscription behavior over 12 months, as shown in Figure 9.

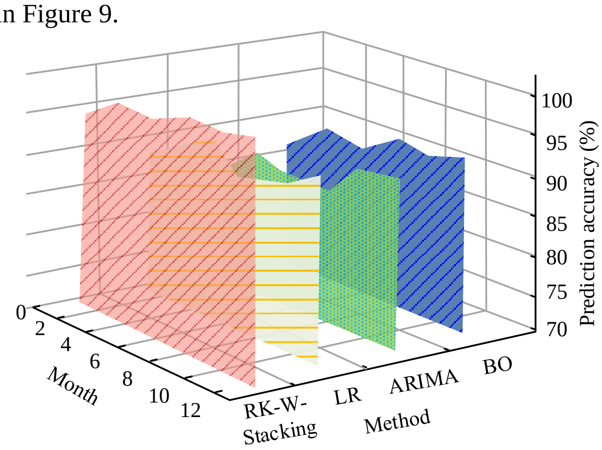



Figure 9 Analysis of the forecast accuracy of user paid subscriptions in 2024

As shown in Figure 9, RK-W-Stacking consistently maintained a prediction accuracy of over 94.2%, with an average fluctuation range of about 3%. The highest accuracy of 97.4% was observed in November. The main reason for the highest accuracy rate in November was that there are more shopping activities guided by merchants in November, which makes consumers' shopping behaviors more regular. In contrast, the overall accuracy of LR, ARIMA, and BO models remained below 90%, with greater fluctuation. ARIMA's prediction accuracy for the second quarter was only 85%, which was 11.8 percentage points lower than RK-W-Stacking's 96.8%. These results indicated that RK-W-Stacking maintained stable and accurate predictions over extended time periods, demonstrating strong adaptability and accuracy. The study further analyzed the relative error in predicting subscription behavior in 2024 and compared the monthly subscription forecasts of RK-W-Stacking with actual consumption data. The results are shown in Figure 10.

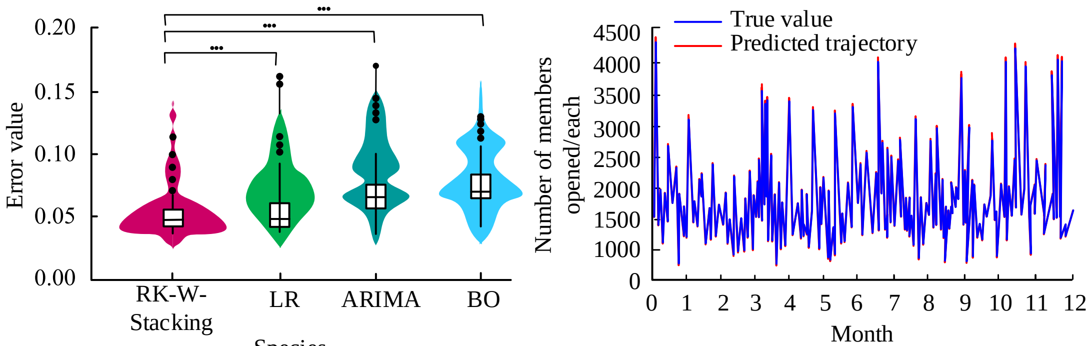



Figure 10: Analysis of prediction accuracy error and payment prediction trajectory

As shown in Figure 10(a), the relative prediction errors for RK-W-Stacking, LR, ARIMA, and BO were 0.04, 0.06, 0.07, and 0.08, respectively. RK-W-Stacking's error distribution was closer to the x-axis, with a maximum relative prediction error not exceeding 0.12. In contrast, LR and ARIMA had some outliers above 0.15, showing greater fluctuation. Although BO's errors were closer to the x-axis, they still exhibited a larger range of variation. RK-W-Stacking's prediction errors showed significant differences compared to LR, ARIMA, and BO. In Figure 10(b), the subscription prediction curve of RK-W-Stacking shows certain deviations in mid-March and early October. The prediction results of the remaining months are highly consistent with the actual results. These findings indicated that RK-W-Stacking achieved high prediction accuracy with minimal diagnostic errors, making it highly effective for long-term monitoring of consumption behavior, significantly outperforming the comparison models. To further confirm the generalization ability of the model, the study introduced the Online Retail Dataset of the University of California with different time scales and types for cross-domain testing. This dataset contains 540,000 transaction records of British e-commerce platforms from 2010 to 2011, covering features such as purchase frequency, amount, and product categories. The experiment retained the original preprocessing process (entropy weight method RFM weighting combined with improved K-means clustering), and divides the training set and the test set at 7:3. The generalization test results are shown in Table 4 as follows.

Table 4: Generalization test results

| Model | Accuracy (%) | AUC | F1-Score (%) | The prediction of a single sample is time-consuming (ms) |
|---|---|---|---|---|
| RK-W-Stacking | 93.7 | 0.89 | 87.5 | 1.4 |
| XGBoost | 88.2 | 0.81 | 79.6 | 3.7 |
| SVM | 84.1 | 0.72 | 75.3 | 2.1 |
| Blending | 87.9 | 0.78 | 80.1 | 1.7 |

As can be seen from Table 4, the accuracy rate of the research method reaches 93.7%, while that of the other methods is all lower than 90%. The AUC of the research method reached 0.89, which was significantly higher than that of other methods. Meanwhile, the prediction time for a single sample of the research method is only 1.4 ms, which is lower than that of other methods. These results proved that the research method could maintain good generalization ability in different types of data and has strong enough computational efficiency.

# 5 Discussion

The RK-W-Stacking model has achieved a significant breakthrough in the field of user consumption behavior prediction through collaborative innovation in methodology. Its core advantage was first reflected in the dynamic feature optimization ability of the entropy weight RFM framework. The research method was based on information entropy to dynamically calibrate the weights of consumption closeness, consumption frequency and consumption amount, quantify the differentiated impact of each dimension on user value, and effectively distinguish the ambiguous boundary between unconventional high-consumption users and stable low-consumption users. The reconstructed RFM features are used as the clustering input, significantly improving the feature separability. The variance-weighted K-means algorithm significantly reduces sensitivity to outliers by redefining the distance measurement logic, assigning greater weights to high-variance features, and incorporating an automatic cluster number determination mechanism based on contour coefficients. This directly contributed to a feature recognition accuracy of 96.3%. The error-weighted stacked integration architecture further amplifies the model performance. The three types of base classifiers, namely XGBoost, Random Forest and GBDT, give full play to their complementary advantages. Meanwhile, the weight allocation mechanism based on the error rate (such as the weight of the best-performing XGBoost reaching 0.38) enhances the contribution of high-performance models, suppressed the noise of weak classifiers and effectively alleviates the risk of overfitting. The logistic regression meta-classifier, by integrating the optimized prediction results, ultimately achieves an AUC of 0.91 and a loss rate of 0.25%, which was 4.2-9.6 percentage points higher than the accuracy of the single model. The robustness of the model was verified in three dimensions: computational efficiency, spatio-temporal adaptability, and cross-domain generalization ability: The peak memory usage is controlled at 1170 MB, and the response time is stable at 120 ms, which benefited from the dimensionality reduction of RFM features and the linear computational complexity of K-means. The 12-month long-term prediction maintained an accuracy rate of more than 94.2%. The use of periodic time feature coding and K-nearest neighbor balancing effectively resisted the interference of seasonal fluctuations and class imbalances. However, the research methods still have some limitations. The black box feature of stacked integration hinders decision traceability. In particular, the predictive fusion mechanism of meta-classifiers lacks transparency, restricting causal inference of marketing strategies. Key parameters such as the learning rate of XGBoost rely on empirical Settings, which may cause performance degradation in new scenarios. The current model deals with discrete consumption snapshots and does not fully model the sequence dependencies of continuous consumption behaviors. In future research, explainable artificial intelligence technologies can be integrated to reveal feature attribution, and explainable meta-classifiers (such as decision trees) can be adopted to enhance transparency. Bayesian optimization is introduced to achieve adaptive tuning of hyperparameters.

# 6 Conclusion

To address the challenges of extracting user consumption features and the low accuracy of consumption behavior prediction, this study developed an integrated consumption behavior feature prediction model. The model utilized an entropy weight method to enhance the RFM processing of initial samples, combined with an optimized K-means algorithm with improved minimum distance calculation logic and a weighted Stacking ensemble classifier. The experimental results showed that the proposed RK-W-Stacking prediction model achieved a consumption feature identification accuracy of 96.3%, a loss rate of 0.25%, and an area under the ROC curve of 0.91. The ROC curve closely aligned with the top of the y-axis, demonstrating high feature identification efficiency, low loss, and strong classification performance. The study further used the RK-W-Stacking prediction model to predict user consumption behavior over 12 months and validated its performance. The model's peak memory usage reached 1170 MB, and the response time remained under 120 ms. Throughout the year, prediction accuracy consistently exceeded 94.2%, with a relative error of 0.04. The predicted results closely matched the actual value curve. The proposed RK-W-Stacking prediction model significantly outperformed the compared prediction models in terms of accuracy and applicability. Although the proposed model demonstrated outstanding performance, the experiment did not comprehensively optimize the key parameter settings, and there is still room for improvement in the selection of base classifiers. Future work will refine the selection of base classifiers through experiments to better match secondary classification samples, reducing the impact of time effects on prediction accuracy.

# References

[1] Ali M, Ullah S, Ahmad M S, Cheok M Y, Alenezi H. Assessing the impact of green consumption behavior and green purchase intention among millennials toward sustainable environment. Environmental Science and Pollution Research, 2023, 30(9): 23335-23347. https://doi.org/10.1007/s11356-022-23811-1

[2] Vighnesh N V, Balachandra P, Chandrashekar D, Sawang S. How cultural values influence sustainable

consumption behavior? An empirical investigation in a non-Western context. Sustainable Development, 2023, 31(2): 990-1007. https://doi.org/10.1002/sd.2436

[3] Alsaad A, Elrehail H, Saif-Alyousfi A Y H. The interaction among religiosity, moral intensity and moral certainty in predicting ethical consumption: A study of Muslim consumers. International Journal of Consumer Studies, 2022, 46(2): 406-418. https://doi.org/10.1111/ijcs.12688

[4] Liu H, Chen J, Dy J, Fu Y. Transforming complex problems into K-means solutions. IEEE transactions on pattern analysis and machine intelligence, 2023, 45(7): 9149-9168. https://doi.org/10.1109/TPAMI.2023.3237667

[5] Ture B A, Akbulut A, Zaim A H, Catal C. Stacking-based ensemble learning for remaining useful life estimation. Soft Computing, 2024, 28(2): 1337-1349. https://doi.org/10.1007/s00500-023-08322-6

[6] Nie F, Li Z, Wang R, Li X. An effective and efficient algorithm for K-means clustering with new formulation. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(4): 3433-3443. https://doi.org/10.1109/TKDE.2022.3155450

[7] Brusco M, Steinley D, Watts A L. Improving the Walktrap algorithm using k-means clustering. Multivariate Behavioral Research, 2024, 59(2): 266-288. https://doi.org/10.1080/00273171.2023.2254767

[8] Chen Y T, Witten D M. Selective inference for k-means clustering. Journal of Machine Learning Research, 2023, 24(152): 1-41. https://doi.org/10.48550/arXiv.2203.15267

[9] Meharie M G, Mengesha W J, Gariy Z A, Mutuku R N. Application of stacking ensemble machine learning algorithm in predicting the cost of highway construction projects. Engineering, Construction and Architectural Management, 2022, 29(7): 2836-2853. https://doi.org/10.1108/ECAM-02-2020-0128

[10] Xue Y, Chen C, Słowik A. Neural architecture search based on a multi-objective evolutionary algorithm with probability stack. IEEE Transactions on Evolutionary Computation, 2023, 27(4): 778-786. https://doi.org/10.1109/TEVC.2023.3252612

[11] Hou S, Liu Y, Yang Q. Real-time prediction of rock mass classification based on TBM operation big data and stacking technique of ensemble learning. Journal of Rock Mechanics and Geotechnical Engineering, 2022, 14(1): 123-143. https://doi.org/10.1016/j.jrmge.2021.05.004

[12] Tang C, Tang Y, Zeng Z, Zhang L, Xiang S. Customer characteristics analysis method based on the selection of electricity consumption characteristics and behavioral portraits of different groups of people. Journal of Intelligent & Fuzzy Systems, 2023, 44(3): 4273-4283. https://doi.org/10.3233/JIFS-220615

[13] Yang X, Huang P, An L, Feng P, Wei B, He P. A growing model-based OCSVM for abnormal student activity detection from daily campus consumption. New Generation Computing, 2022, 40(4): 915-933. https://doi.org/10.1007/s00354-022-00193-z

[14] Yang J, Li J, Cao Y. Analysis of peer effects on consumption in rural China based on social networks. Applied Economics, 2023, 55(6): 617-635. https://doi.org/10.1080/00036846.2022.2092592

[15] Saif S, Zameer H, Wang Y, Ali Q. The effect of retailer CSR and consumer environmental responsibility on green consumption behaviors: mediation of environmental concern and customer trust. Marketing Intelligence & Planning, 2024, 42(1): 149-167. https://doi.org/10.1108/MIP-04-2023-0181

[16] Khan A A, Meraj M, Asif H M. A multi-dimensional exploration of university students' sustainable consumption and environmental awareness in Pakistan. Journal of Management Info, 2024, 11(1): 102-122. https://doi.org/10.31580/jmi.v11i1.3005

[17] Chen W, Wang Q. Application of improved k-means algorithm in e-commerce data processing. Informatica, 2024, 48(11): 147-166. https://doi.org/10.31449/inf.v48i11.5996

[18] Malik M S I, Nawaz A. SEHP: stacking-based ensemble learning on novel features for review helpfulness prediction. Knowledge and Information Systems, 2024, 66(1): 653-679. https://doi.org/10.1007/s10115-023-02020-3

[19] Du B, Wang M, Zhang J, Chen Y, Wang T. Urban flood prediction based on PCSWMM and stacking integrated learning model. Natural Hazards, 2025, 121(2): 1971-1995. https://doi.org/10.1007/s11069-024-06893-7

[20] Liang C. Application of maximum entropy fuzzy clustering algorithm with soft computing in migration anomaly detection. Informatica, 2024, 48(17): 171-182. https://doi.org/10.31449/inf.v48i17.6537