Scholar Publishing Group
International Journal of Business Management and Economics and Trade
https://doi.org/10.38007/IJBMET.2026.070202
ISSN 2790-0959 Vol. 7, Issue 2:11-18

The Integration and Application of Ai Technology in Marketing Data Analysis and Business Decision-Making

Yiyu Yang

Columbia University School of Professional Studies, 203 Lewisohn Hall, 2970 Broadway, MC 4119, New York, NY 10027

Keywords: Customer value segmentation, churn prediction, machine learning, artificial intelligence, marketing data analysis

Abstract: In the context of digitalization and stock competition in consumer finance, the loss of high-value credit card customers (20% contributing 80% of profits) directly affects corporate profits. The traditional RFM model has a single dimension and collinearity, making it difficult to capture behavioral changes; K-means clustering has difficulties in determining K values and dynamic adaptability; Static machine learning models are not easily able to reflect complex factors such as economic fluctuations. This study integrates transaction frequency, amount, interaction behavior, transaction changes, and credit characteristics to construct an improved RFM+CBC model that breaks through traditional dimensional limitations. Optimize K-means clustering by combining contour coefficient and Calinski Harabasz index, determine indicator weights using entropy weight method, and achieve precise customer segmentation. Based on logistic regression, random forest, XGBoost, and Stacking fusion models for churn prediction, the results show that customer value segmentation significantly improves prediction performance. XGBoost has an accuracy rate of 97.73%, a recall rate of 93.10%, and a Stacking accuracy rate of 97.98%among high-value customer groups, verifying the core value of segmentation in predicting high-value customers. In the future, it is possible to expand multi-source data (such as social media behavior), develop dynamic segmentation and real-time prediction systems, and apply deep learning technology to further strengthen the integration and application of artificial intelligence in marketing data analysis and business decision-making, enhancing customer loyalty and long-term benefits.

1 Introduction

Against the backdrop of accelerating digital and platform transformation in the field of consumer finance, credit cards, as the core carrier connecting consumption and finance, have entered the era of stock competition from the stage of scale expansion. The loss of high-value customers (following the Pareto principle[1], 20% of customers contribute 80% of profits) directly affects enterprise revenue and market share. Although existing research emphasizes the importance of customer segmentation and data-driven decision-making, there are still limitations: traditional RFM models[2] struggle to fully capture changes in customer behavior due to issues such as collinearity of dependent variables and a single dimension; Clustering algorithms such as K-means face technical challenges in determining K values, selecting initial centers, and dynamic adaptability; Although machine learning models have evolved from traditional statistical methods to ensemble learning, static models are difficult to reflect the impact of complex factors such as economic fluctuations and market competition on customer behavior, and data processing and feature engineering quality directly affect model performance. The urgent need for enterprises to improve the accuracy of customer management has driven this study. By integrating multidimensional data such as consumer behavior and credit characteristics, an innovative customer value evaluation framework is developed, and advanced machine learning technology is combined to optimize the churn prediction model, achieving the transformation from "product center" to "customer center". The specific goals include building an RFM+CBC model that integrates transaction frequency, amount, interactive behavior, transaction changes, and credit characteristics to break through the limitations of traditional RFM dimensions. The K-means clustering is optimized using silhouette coefficients[3] and Calinski Harabasz index, and the entropy weight method is used to determine indicator weights to achieve accurate segmentation of customer groups. Logistic regression, random forest, XGBoost, and Stacking fusion models are used for churn prediction to verify the key role of customer value segmentation in improving prediction performance. The contribution of this study is to pioneer a multidimensional customer value assessment model that integrates behavior and credit characteristics to provide a more comprehensive customer segmentation tool. Through model fusion and algorithm optimization, it significantly improves the prediction accuracy and recall rate of high-value customer groups (such as the XGBoost model achieving 97.73% accuracy and 93.10% recall rate among high-value customers), verifies the core value of customer segmentation in improving prediction efficiency, provides scientific basis for financial institutions to formulate differentiated customer retention strategies, promotes strategic transformation from "scale growth" to "value mining", and ultimately achieves customer loyalty enhancement and long-term revenue growth, fully reflecting the integration and application value of artificial intelligence technology in marketing data analysis and business decision-making.

2 Theoretical Background

2.1 Business application of customer segmentation and value assessment

Customer segmentation[4]is the process of dividing customers into different categories based on specific goals, enabling businesses to provide customized services and resource allocation tailored to the unique needs and values of each group. This concept was proposed by Wendell Smith in 1956, and with technological advancements, it has evolved from traditional statistical analysis to data mining and big data technology. The subdivision dimension has also expanded from a single behavior or value to a comprehensive evaluation of behavior value. Customer value research began in the 1980s, emphasizing the differences in the contributions of different customers to the enterprise (such as the 80/20 rule where 20% of high-value customers contribute 80% of profits). Through segmentation, enterprises can optimize resource allocation, enhance high-value customer service, and achieve maximum benefits. Customer churn prediction aims to accurately identify potential churn customers, analyze the churn trends of customers with different values, guide enterprises to allocate limited resources, implement precise marketing, and ultimately promote profit growth. This research that combines customer value and segmentation provides a scientific basis for business decision-making, helping enterprises transition from "scale growth" to "value mining" strategy.

2.2 The core value and application logic of RFM model

In the management of bank credit card customers, the RFM model evaluates customer value through three dimensions: the most recent consumption time (Recency), consumption frequency (Frequency), and consumption amount (Monetary). It is a key tool for enterprises to identify lost customers and reduce customer relationship management costs. The theoretical advantage of this model lies in its simplicity and practicality - by integrating ratings or ratings from three dimensions, companies can divide customers into different groups, each group exhibiting different purchasing potential and loyalty due to differences in RFM values. This segmentation enables businesses to design more targeted marketing strategies, enhance customer satisfaction and loyalty, and ultimately achieve revenue growth. As a classic tool in customer relationship management (CRM)[5] and direct marketing, the RFM model is still widely used in customer rating, segmentation, and maintenance strategy development.

3 Research method

3.1 Overview of Analysis Methods in Customer Segmentation Practice

In customer segmentation practice, correlation analysis[6] identifies the strength of the correlation between features and target variables (such as customer churn) through statistical methods: chi square test is used to test the independence between categorical variables, and the statistical formula is

$$Y^2 = \sum_{j=1}^{l} \frac{(O_j - F_j)^2}{F_j} \tag{1}$$

When the P value is less than 0.01, the correlation is significant; The Pearson correlation coefficient measures the degree of linear correlation, and the formula is

$$r = \frac{COV(Y,Z)}{\sigma_r \sigma_z} = \frac{\sum(Y_J - \bar{Y})(Z_J - \bar{Z})}{\sqrt{\sum(Y_J - \bar{Y})(Z_J - \bar{Z})^2}} \tag{2}$$

with a value range of [-1,1]; Spearman correlation coefficient is used to evaluate monotonic relationships, and the formula is

$$r_T = 1 - \frac{6\sum d_j^2}{n(n^2-1)} \tag{3}$$

Suitable for non-linear scenarios. Cluster analysis is based on the K-means algorithm [7], and the process includes determining the K value, selecting the initial center, assigning data points, updating the center, and iterative optimization. The goal is to minimize the sum of squares within the cluster (WCSS). The commonly used elbow method (WCSS descent inflection point), contour coefficient (the closer the value is to 1, the better the effect), and Calinski Harabaz index (variance ratio criterion) are selected for the K value. The analysis of indicator weights adopts the entropy weight method and standardizes the data through positive indicators

$$Y_{norm} = \frac{Y - Y_{min}}{Y_{max} - Y_{min}} \tag{4}$$

Y,Negative indicator,Calculation of specific gravity

$$P_{jk} = \frac{x_{jk}}{\sum_{j=1}^{m} x_{jk}} \tag{5}$$

Information entropy calculation

$$F_K = -\ln(m)\sum_{j=1}^{m} P_{jk}\ln(P_{jk}) \tag{6}$$

and weight calculation

$$X_k = \frac{1-F_K}{\sum_{k=1}^{n}(1-F_K)} \tag{7}$$

realize objective weight allocation. These methods collectively support the precise segmentation and value assessment of customer groups.

3.2 Customer churn concept and response strategy framework

Customer churn[8]is a core business issue in highly competitive industries such as telecommunications, finance, and insurance, directly affecting enterprise revenue and driving up customer acquisition costs. The customer lifecycle is divided into an assessment period (establishing a brand, identifying potential high-value customers), a formation period (obtaining initial profits, reducing promotional costs), a stabilization period (optimizing experiences through CRM systems, maintaining long-term relationships), and a degradation period (predicting and retaining customers who are about to leave). Customer churn is defined as the behavior of stopping the use of products or services, turning to competitors, and specific manifestations include service cancellation, inactivity, no longer interacting, and account closure. Different industries need to clarify specific definitions based on business goals. Loss warning requires monitoring user behavior (such as decreased activity), setting warning indicators (such as consecutive months of no login), applying predictive models, and conducting customer surveys; The recovery strategy includes personalized discounts, customized services, active communication, product improvement, quick response to problems, and CRM system tracking and interaction. By comprehensively applying these strategies, enterprises can effectively prevent churn, improve satisfaction, and recover users, ultimately achieving long-term maintenance and value mining of customer relationships.

3.3 Machine learning classification and its application in predicting customer churn

Machine learning can be divided into three categories based on training methods and data characteristics: supervised learning, unsupervised learning, and semi supervised learning. Supervised learning uses labeled data to train models to predict unknown sample labels (such as customer churn prediction), which is suitable for regression and classification problems; Unsupervised learning utilizes data structures such as clustering and dimensionality reduction to process unlabeled data; Semi supervised learning combines limited labeled data with a large amount of unlabeled data to improve learning performance. Customer churn prediction, as a binary classification problem in supervised learning, follows the process of "clarifying the problem data collection preprocessing model selection and training evaluation strategy formulation". Common evaluation indicators include accuracy, precision, recall, F1 score, and AUC value. In practice, logistic regression, as a classic binary classification model, maps linear outputs to probabilities through the Sigmoid function[9]and is suitable for linearly separable data; Random Forest constructs multiple decision trees through Bootstrap sampling and feature random selection, and uses a voting mechanism [10]to improve prediction stability (as shown in Figure 1)

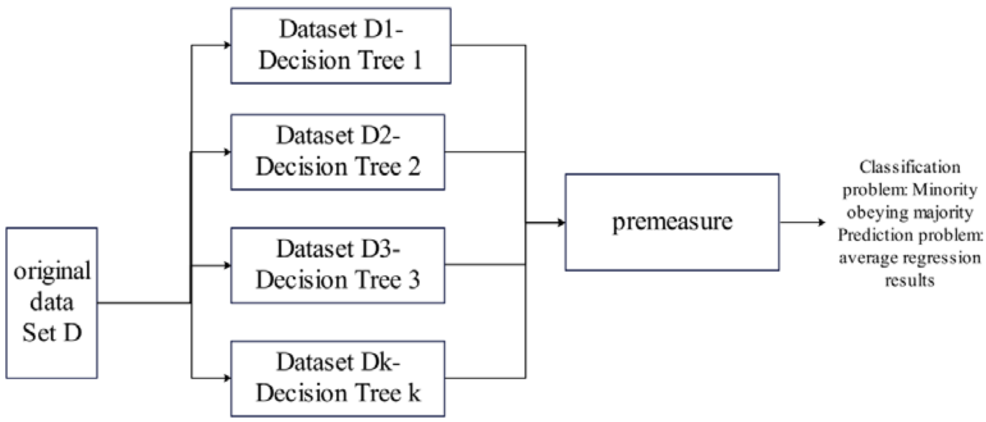

Figure 1 Ensemble Learning Framework Based on Decision Trees

XGBoost, implemented as a gradient boosting tree, optimizes model complexity through loss functions and regularization terms, supporting missing value handling and distributed computing; Stacking model fusion trains a meta model (such as logistic regression) based on the prediction results of the base model layer (such as logistic regression, random forest, XGBoost), and captures more information by utilizing the differences between different models. These algorithms have their own advantages in predicting customer churn: logistic regression provides good interpretability, random forest has strong ability to handle high-dimensional data, XGBoost reduces overfitting with efficient computation and regularization, and Stacking further improves prediction performance through model fusion. By comparing the characteristics of different models and task requirements, the optimal model or fusion strategy can be selected to provide scientific basis for formulating customer retention strategies (such as personalized discounts and service improvements), ultimately achieving a reduction in customer churn rate and long-term relationship maintenance.

4 Results and discussion

4.1 Process Bank Credit Card Customer Data Preparation and Preprocessing Process

This chapter focuses on the processing and analysis of a bank credit card customer dataset, which is sourced from a bank credit card customer information set released on the Kaggle platform in 2020. The dataset includes 10127 samples, 20 customer characteristics, and 1 target variable (customer churn indicator), of which 6 are categorical features (such as gender and education level) and 14 are numerical features (such as age and transaction amount). Revealing data distribution characteristics through descriptive statistics: In terms of numerical features, customers are concentrated in the age range of 26-73 years old (average 46 years old), with an average of 2.35 family members, 35.93 months of card holding, an average credit limit of 8631.95 CNY” , an average total transaction amount of 4404.09 CNY” an average transaction frequency of 64.86 times, and an average usage ratio of 0.27 (standard deviation of 0.28, indicating significant differences in usage). The distribution of category characteristics shows that the proportion of lost customers is about 16.1% (1627 people), which belongs to imbalanced data; There are slightly more females in the gender dimension, but the male turnover rate is higher; The churn rate of unmarried customers is higher than that of married and single groups; Low income groups have a higher turnover rate, while high-income groups have better loyalty; Blue card customers have the highest number and high churn rate, while gold/platinum card customers have a lower churn rate. Data preprocessing involves multi-step operations: deleting unique identifiers (customer numbers) with no predictive value; Since there are no missing values in the data, there is no need to fill in the data; Categorical variables are encoded using label encoding(such as gender, churn markers) and one-hot encoding(such as education level, marital status, etc.); Create new features based on the RFM model (such as the number of active months and average transaction amount in the past 12 months); Finally, the feature scale is standardized and unified through Min max to avoid algorithm bias towards high-value features. These preprocessing steps lay a high-quality data foundation for subsequent customer segmentation, churn prediction model construction, and value evaluation.

4.2 Model experiment

We used a confusion matrix and five metrics to evaluate the model: accuracy, precision, recall, F1, and AUC. Accuracy may seem simple, but it is easy to deceive people when the data is imbalanced; Accuracy focuses on reducing misjudgments, recall focuses more on identifying all potential lost customers, F1 is a compromise between the two, and AUC measures overall differentiation ability. For customer value segmentation, we use the RFM+CBC model (including recent active months, transaction frequency, amount, interaction behavior, transaction changes, and credit characteristics) and K-means clustering. When selecting the K value, we tried the elbow method (SSE keeps decreasing, difficult to judge), contour coefficient, and Calinski Harabasz index, and finally determined K=3. As a result, 4747 high-value customers, 3395 medium value customers, and 1985 low value customers were identified. The differentiation retention strategy is designed around these three groups.

4.3 Effect analysis

We compared the performance of logistic regression, random forest, XGBoost, and Stacking fusion models in predicting credit card customer churn. The data comes from 1520 customers, divided into three groups based on value: low (296 people), medium (507 people), and high (717 people), while also making predictions for the entire customer base. For low value customers, XGBoost is the best: with an accuracy rate of 95.26% and a recall rate of 91.43%, AUC 93.89%。 Stacking is slightly inferior. XGBoost still leads among mid value customers with an accuracy rate of 97.20% and a recall rate of 87.50%, AUC 92.79%。 The accuracy rate of high-value customers is generally high, with Stacking accuracy slightly higher at 97.98%, but XGBoost's recall (93.10%), F1 (92.31%), and AUC (95.81%) are more balanced. Considering the overall performance and efficiency, it is recommended to use XGBoost for subsequent verification. We also validated the effectiveness of customer value segmentation: the predictive performance of grouping by value was better than that of an equal number of random groups. For example, XGBoost's AUC on low value customers is 3.08% higher than that of random grouping, and all models in the medium and high-value groups are significantly better than those in random grouping. The test results of the validation set of 1519 customers (281/541/697 low/medium/high-value customers) are shown in Table 1

Segmenting strategies can indeed improve prediction accuracy, especially in the high-value customer segment. The model suggests stratifying customers of different values: although low value customers are predicted accurately, do not invest too much resources. Use customized services and discounts to increase satisfaction; The median customer recall rate is relatively low, and we need to rely on special discounts and better services to win over customers; High value customers are the main force, and personalized services, exclusive privileges, and financial consulting all need to keep up,fostering deeper engagement and long-term loyalty. In addition, banks need to monitor changes in customer behavior, adjust communication and service strategies at any time, implement loyalty programs, innovate products, and warning systems, and then strengthen customer relationships through feedback mechanisms. Only in this way can resources be allocated where they will have the greatest impact, improve customer satisfaction, and naturally stand firm in the market.

Table 1 Segmented Churn Prediction Results

Customer Segment	Model Name	Accuracy	Precision	Recall	F1 Score	AUC Value
Low-Value	Logistic Regression	82.84%	58.25%	79.05%	67.07%	81.48%
Low-Value	Random Forest	94.42%	85.84%	89.52%	87.65%	92.67%
Low-Value	XGBoost	95.26%	87.67%	91.43%	89.51%	93.89%
Low-Value	Stacking	94.84%	88.15%	88.57%	88.36%	92.60%
Medium-Value	Logistic Regression	84.39%	32.88%	85.71%	47.52%	84.99%
Medium-Value	Random Forest	96.47%	76.67%	82.14%	79.31%	89.95%
Medium-Value	XGBoost	97.20%	80.33%	87.50%	83.76%	92.79%
Medium-Value	Stacking	96.91%	79.66%	83.93%	81.74%	91.00%
High-Value	Logistic Regression	87.41%	54.88%	77.59%	64.29%	83.34%
High-Value	Random Forest	97.73%	91.53%	93.10%	92.31%	95.81%
High-Value	XGBoost	97.73%	91.53%	93.10%	92.31%	95.81%
High-Value	Stacking	97.98%	94.64%	91.38%	92.98%	95.25%

5 Conclusion

We use an improved RFM+CBC model, combined with transaction behavior and credit card usage habits, to divide customers into three value groups: high, medium, and low. Then use algorithms such as logistic regression, random forest, XGBoost, and Stacking to predict which customers are likely to churn - with the best predictive performance for high-value customers. This system enables banks to develop differentiated retention strategies for different groups: providing personalized services to high-value customers, finding ways to increase stickiness for medium value customers, and optimizing resource investment for low value customers. The goal is to reduce churn rates, enhance competitiveness and efficiency. The next step is to incorporate more data sources, such as social media behavior and mobile payment habits, explore new scenarios for online and mobile banking, develop dynamic segmentation and real-time prediction systems, and then try deep learning. This can better integrate AI into marketing analysis and business decision-making, making customer maintenance and growth strategies smarter and more efficient.

References

[1] Giannini M. Focus your resources with Pareto principle[J].Automotive news, 2025 (TN.7188): 100.

[2] Zheng Y, Gan W, Chen Z, et al. SeqRFM: Fast RFM analysis in sequence data[J]. Information Sciences, 2026, 723(c):122652.

[3] Lai H, Huang T, Lu B L, et al. Silhouette coefficient-based weighting k-means algorithm[J]. Neural Computing and Applications, 2025, 37(5):3061-3075.

[4] Paulraj K, Murraliraj R, Kumar S S M, et al. Sales price prediction with customer segmentation[J]. AIP Conference Proceedings, 2025, 3279(1):020008.
[5] Rubinstein M A. Reimagining the airport-customer relationship through customer relationship management[J] .Journal of Airport Management, 2025, 20(1)..
[6] Park H, Bai S, Ye Z, et al. Graph Canonical Correlation Analysis[J]. 2025.
Gan G. A k-Means Algorithm with Automatic Outlier Detection[J]. ELECTRONICS, 2025, 14(9):1723.
[7] Kumari B, Vinaya Prasad T. Customer Churn Prediction using Data Science[J]. International Scientific Journal of Engineering & Management, 2025, 4(7).
[8] Bosso V D A, Nardini R M, De Sousa M A D A, et al. An Area-Efficient and Low-Error FPGA-Based Sigmoid Function Approximation[J]. Applied Sciences (2076-3417), 2025, 15(21).
[9] Ulitin N V, Shadrina G R, Anisimova V I, et al. Interpretation of the Structure-Glass Transition Temperature Relationship for Organic Homopolymers with the Use of Increment, Random Forest, and Density Functional Theory Methods[J].Journal of Structural Chemistry, 2025, 66(5).