Predictive Analysis of Customer Churn in Community-Supported Agriculture Based on RFM Modeling

Xiaoying Yan*, Mei Yin, Renren Li

Software Engineering Institute of Guangzhou, Guangzhou 510300, Guangdong Province, China

*Corresponding author:93847153@qq.com

Abstract. High customer turnover has been a significant challenge in China's community-supported agriculture (CSA) industry. Establishing a customer churn prediction and intervention management mechanism based on consumption data analysis is of great significance for the sustainable and healthy operation of many Chinese CSA family farms. In this paper, we utilize RFM models (Recency, Frequency, and Monetary) and algorithms to rank and classify the consumption ability of CSA customers on a regular basis. This is done by analyzing their recent purchase time, number of times of consumption, and consumption data. In order to determine the reasons behind CSA customer loss and intervene early, it is important to continuously enhance the knowledge and level of intelligent management in the CSA industry. This will effectively support the healthy and stable development of the community-supported agriculture industry.

Keywords: RFM model; community-supported agriculture; customer churn; early warning models

1 Introduction

In 2008, Dr. Shi Yan introduced the Community Supported Agriculture (CSA) model to China. In 2008, Dr. Shi Yan introduced the Community Supported Agriculture (CSA) model to China. The following year, she founded the first CSA farm in Beijing, known as the Little Donkey Farm. This approach, which emphasizes the direct connection between agricultural producers and consumers in order to produce healthy agricultural products, has garnered widespread attention. Currently, CSA farms are located in all provinces, autonomous regions, and municipalities in China, with the exception of Tibet[1]. However, CSA producers are facing problems such as low willingness to continue cooperation and a high turnover rate to varying degrees. Currently, CSA farms in China are mostly family-run, with a short time in business, small scale, and a lack of management experience. This makes it difficult to retain customers for an extended period. This situation makes the production and marketing activities of CSA farms more uncertain[2]. Due to capacity constraints, smallholder CSA family farms face difficulties in accessing the supermarket chain channel. The main source of new customers primarily relies on word-of-mouth from existing customers and the promotion through

B. K. Kandel et al. (eds.), *Proceedings of the 2023 3rd International Conference on Business Administration and Data Science (BADS 2023)*, Atlantis Highlights in Computer Sciences 19,
https://doi.org/10.2991/978-94-6463-326-9_39

the farm's self-media channels, which in turn generates small-scale private traffic. The cost of acquiring new customers is much higher than retaining existing ones. Establishing a mechanism to predict and intervene in customer loss management, based on historical consumption data, is of great significance for the sustainable and healthy development of many CSA family farms in China. This mechanism can help reduce operating costs.

2 RFM model

The RFM model, proposed by Arthur Hughes of the Database Marketing Institute in the United States, is a widely used tool for measuring customer value based on customers' historical consumption behavior. The RFM model is based on the customer's historical consumption behavior. In this model, R (recency) represents the interval between the customer's most recent purchase and the time of analysis. It reflects the customer's recent consumption activity. F (frequency) represents the number of times the customer has made purchases within a specific period of time. M (monetary) refers to the total amount of the customer's consumption within that same period[3]. Purchases tend to reflect customers' purchasing power and satisfaction with the product and its services[4]. It is important to note that quantity or weight data may also be used when M-values are applied. The reason for this is that there may be significant price variations based on the type and quality of the produce. In other words, the selection may differ, but the daily nutritional requirements are still fulfilled to the same extent. Therefore, in order to avoid statistical errors caused by pricing factors, the M-value can be used when analyzing purchase weight or purchase quantity data. The RFM model is highly practical in enterprise customer relationship management due to its clear thinking and simplicity of use[5]. Additionally, it has significant predictive ability in determining customers' consumption behavior[6]. The feasibility of RFM metrics has been proven in numerous experiments involving customer segmentation and precision marketing[7].

3 RFM model analysis process and principles

In this paper, the recent purchase behavior, consumption times, and purchase volume of CSA customers are used as the basis for segmentation and screening. The shorter the R-value from the analysis time, the higher the level of recent customer engagement with CSA products. The F-value represents the number of times CSA products have been purchased within a specific measurement period. It is important to exclude transactions that have been refunded or returned from the statistics. Agricultural products differ from general consumer products. Customers' preferences for CSA farm products, as well as price and other factors, may lead to bias in evaluating the M-value. Therefore, in this paper, the M-value is defined as the volume of CSA products purchased. Therefore, the M-value in this paper is defined as the purchase volume of CSA products. The specific value extraction method of the model is shown in Table 1. Additionally, users have the option to assign weights to RFM based on the actual operations of the farm.

In addition, according to the R-value calculation rule, the longer the time between the most recent purchase date and the measurement deadline, the higher the likelihood of customer churn. Therefore, the R-value calculation should be expressed as R* (-1) in order to preserve its objectivity. In summary, this paper presents the CSA customer RFM scoring model as follows: score = α R*(-1) + β F + γ M.

Table 1. Meaning of RFM model features

model	RFM
Recent purchases (R)	Date of last purchase - Measurement cut-off date
Number of purchases (F)	Number of purchases - number of returns during the measurement period
Purchases (M)	Sum of the number of all transactions during the measurement period

In the specific quantization process, the system needs to address the magnitude gap problem before performing operations because the three values of R, F, and M have different magnitudes and units of magnitude. Here, we can utilize min-max standardization processing to linearly transform and scale the original data of R, F, and M. This transformation ensures that the original values are scaled proportionally and converge to the interval from 0 to 1[8]. The conversion function can be represented by the following equation:

$$x^* = \frac{(x - xmin)}{(xmax - xmin)} \tag{1}$$

In Equation (1), xmin and xmax represent the minimum and maximum values in the R, F, and M arrays, respectively, and x* represents the converted value. This method may result in changes to xmin and xmax when new data is added, requiring recalculation. It is more appropriate for applications where the values of R, F, and M are more concentrated. Therefore, the Z-score standardization method can also be used to convert data by standardizing the mean and standard deviation of the original data, as depicted in Equation (2). The data processed using this method follows a Gaussian distribution, with a mean of 0 and a standard deviation of 1. The Z-score standardized conversion function is shown in Equation (3):

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{2}$$

$$x^* = \frac{(x - \mu)}{\sigma} \tag{3}$$

In Equation (3), μ represents the mean value of each group of sample data from the R, F, and M arrays, while σ represents the standard deviation of each group of samples. Taking R and M data as an example, the gradient descent process before and after the standardization process is shown in Figures 1 and 2, respectively.

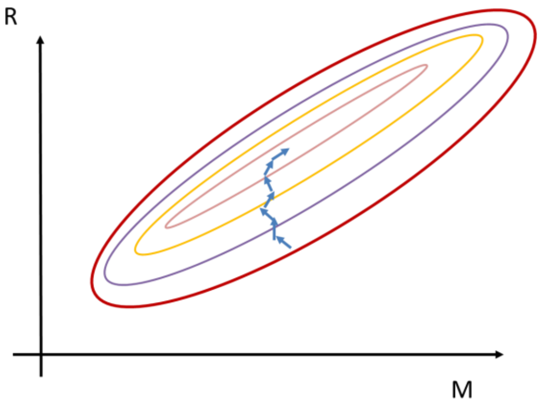

Fig. 1. Data gradient descent process before standardized processing

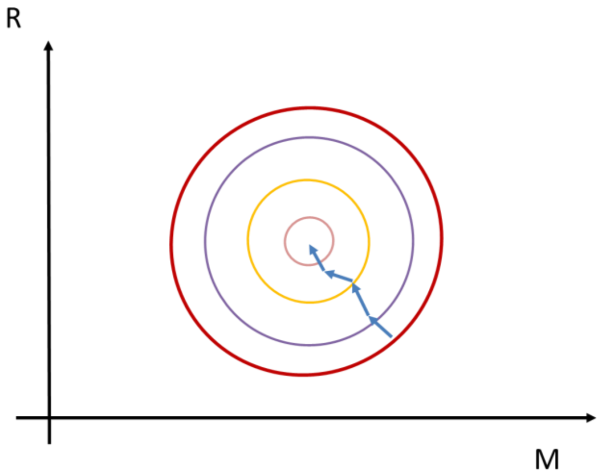

Fig. 2. Gradient descent process of data after standardized processing

In the case of the same machine learning rate, the number of iterations required for non-normalized data is significantly higher than the number required after normalization. Therefore, the R, F, and M values are all mapped within the interval from 0 to 1. This is done to ensure comparability between the data and to expedite data analysis.

After calculating the final score of CSA customers' RFM values, the system then arranges the final customer RFM data from smallest to largest (which can be divided into quartiles or deciles as needed). We will use the quartile as an example. CSA customers in the lower quartile (i.e., 0-25% range) are classified as churn warning customer groups[9]. The system sends out commands to remind CSA farm operators and assist in identifying reasons for low customer RFM values that need to be addressed promptly. The RFM score levels and intervention decisions are presented in Table 2.

Table 2. CSA Customer Churn Early Warning Rating Scale

RFM Score Level	0~25%	26%~50%	51%~75%	76%~100%
Customer Churn Intervention	Strong intervention	Weak intervention	Non-intervention	Non-intervention

The key difference between interventions on customer churn lies in the number of communications and the design of preferences. It is worth noting that fruits and vegetables are frequently purchased items. To provide early warning reminders, it is important to set independent thresholds for recent purchase dates and number of purchases. The specific settings of the R, F, and M thresholds are determined by the product variety and pricing of the CSA farm store. For example, if a customer has not made any purchases for more than two months, it will be considered a strong indication of churn. In such cases, the system will prompt the CSA farm to promptly attend to the customer, identify the reasons for their potential churn, and take corrective measures to prevent it. In addition to the above, the RFM scores of CSA customers may always be lower than the average level of other customers due to their family preferences and demographics. Therefore, when deciding whether to set up a churn warning, it is also imperative to consider the customer's own transaction history for the same period of time, either monthly or quarterly, for CSA products. If the transaction data is significantly lower than the average value for the same period, then a warning will be issued.

4 The overall architecture of the CSA customer churn early warning system based on the RFM model

CSA customer churn early warning system includes modules such as data collection, data cleaning, data standardization, RFM calculation, early warning decision-making, customer care, and data connectivity. Customers' historical consumption data can be obtained through the farm's online store integration or manual entry. After performing data cleansing and standardization, the RFM score calculation and classification of customer information, as well as targeted customer care to enhance customer RFM value for customer retention, the new consumption data will be reintroduced into the data collection module to initiate a new cycle, as depicted in Figure 3.

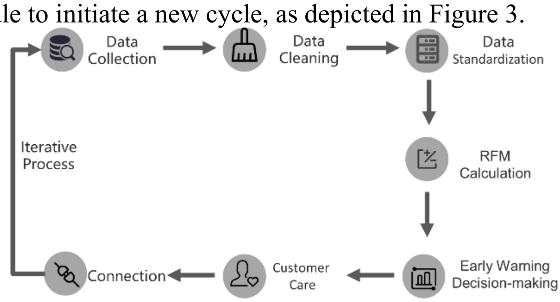

Fig. 3. Schematic diagram of CSA customer churn early warning analysis based on RFM modeling

The functions of the modules of the CSA Customer Churn Early Warning System are as follows:

1. Data Collection Module: The historical consumption records of CSA customers will be collected through the API interface of each e-commerce platform or through manual entry. The collected information will include customer ID, purchase time, number of transactions, types of agricultural products purchased, amount and quantity purchased, as well as returns and exchanges.
2. Data Cleaning Module: De-weighting, null value checking, and outlier correction are performed on the collected original consumption records. Here, it is also necessary to pay attention to merging consumption records from different e-commerce platforms within the same family, as well as different IDs within the same family, and so on.
3. Data Standardization Module: This module standardizes data such as purchase time, number of transactions, types of agricultural products purchased, and the amount and quantity of purchases. Eliminating the disparity in magnitude ensures the accuracy and objectivity of data comparisons across different magnitudes and units.
4. RFM Calculation Module: The RFM score is calculated using the formula score = α R * (-1) + β F + γ M. Since the value of R can be negative, the total RFM score may also be negative. If needed, the RFM score can be adjusted by adding a constant greater than 1 to ensure that the final score is positive. For example: score = αR*(-1) + βF + γM + 10.
5. Early Warning Decision Module: The module arranges CSA customer RFM scores in descending order and sets thresholds. It provides early warning for customer information that exceeds these thresholds, indicating potential churn.
6. Customer Care Module: The customer care approach of CSA products needs to be more targeted. Through analyzing historical transaction data and online communication, among other methods, we can determine whether customers have been lost due to factors such as changes in demand, competition, or product or service faults (excluding factors such as customer death or bankruptcy). Based on this analysis, we can then implement targeted strategies to correct and optimize products, services, or business processes, as well as engage in timely communication, in order to effectively win back customers.
7. Data Connection Module: This module connects the consumption data from customer care to the data collection module for a new round of data processing cycles.

5 Conclusion and discussion

Unlike standardized industrial products that can withstand storage, agricultural products have unique characteristics and are not resistant to storage. Therefore, a sustainable and stable customer base is crucial for the survival and growth of CSA farms. In China, the CSA industry is primarily operated by small farmers, and challenges such as a short history of development and inadequate management persist, hindering the industry's progress. A significant number of CSA family farms assess customer attrition based on whether or not they have disengaged from the community. This form of post-loss concern often results in a situation where it is difficult to regain customers. The absence of data-driven prognostic thinking by farm operators results in customer attrition, higher operating expenses, and increased management risks.

Early detection of customer churn and targeted customer care through digital channels, along with continuous improvement of knowledge-based and intelligent operation and management in the CSA industry, will effectively contribute to the healthy and stable development of the community-supported agriculture industry.

Acknowledgment

Guangdong Province Philosophy and Social Science General Project: Research of the practice model of U+CSA public welfare assistance to farmers in the China Great Bay Area (Project No. GD21CYJ17).

References

1. Shi Y., Cheng C.W., Zhang X.X., et al. 2019 China CSA Industry Report [R]. Beijing: SMART Vacation Industry Intelligence Platform, 2019:18-52.
2. Chen W.P. Producers' strategies to build consumers' food trust in community-supported agriculture scenarios--a case study of Gao family farmers in Anlong village, Sichuan[J]. China Rural Economy,2013(02):48-60.
3. Shi A.X., Zhang J. Research on e-commerce user value classification based on improved RFM model[J]. Computer Technology and Development,2022,32(12):123-128.
4. Anitha P,Patil M M. RFM model for customer purchase behavior using K-Means algorithm[J]. Journal of King Saud University - Computer and Information Sciences,2019(prepublish).
5. Tie J.C. Research on automatic credit card customer feature mining method based on RFM model[J]. Shanghai Management Science,2023,45(04):55-60.
6. Zhang R.X., Zhang J.X., Li S.Y. Customer screening and collaborative filtering recommendation algorithm based on improved RFM model[J]. Network Security Technology and Application,2023(07):42-45.
7. Bughin J. Digital user segmentation and privacy concerns[J]. Journal of Direct, Data and Digital Marketing Practice,2011,13(2).
8. Zhuo L., Sun X. A digital cluster user classification method based on an improved RFM model[J]. Computer Application Research, 2020, 37(9):5. DOI:10.19734/j.issn.1001-3695.2019.05.0112.
9. Tanweer A,Aseem S,Rafeeq A, et al. A Hybrid Approach of Prediction Using Rating and Review Data[J]. International Journal of Information Retrieval Research (IJIRR),2022,12(2).