Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 15, No. 3, 2024

Research on User Churn Prediction of Traditional Supermarket Digital Platform

Honglei Yin, Yilei Pei *

School of Management, Beijing Union University, Beijing, China
* **Corresponding author**: Yilei Pei (Email: peiyilei@126.com)

Abstract: This paper uses the advantages of machine learning algorithms to conduct empirical research on classification prediction to explore the user churn prediction model of traditional supermarket digital platform. User reviews on the online platform of Wumart were used as the data source, and user information of Ehime Orange purchased was used as the research object. Starting from the two dimensions of user value characteristics and evaluation characteristics, the user's purchase information is collected, and five characteristic data such as the latest consumption time, consumption times, consumption amount, rating and review text are obtained by calculating the information. Later, a predictive model was built based on XGBoost, a machine learning algorithm, to predict the trend of user churn. By comparing and analyzing the contribution of characteristic variables in user churn prediction, the types of lost users are divided according to key characteristic variables such as average monthly consumption times and average consumption amount, and corresponding retention strategies are proposed. The validation found that the sentiment of the score and review text had a significant impact on user churn prediction. By analyzing the important variables affecting the prediction of user churn, this paper summarizes two types of churn users, and formulates corresponding retention strategies for users with less demand and no demand. This is of great significance for reducing the loss of digital platform users, maintaining old users, improving the profits of supermarkets, and achieving the goals of healthy and sustainable development of supermarkets.

Keywords: User churn, Traditional supermarkets, Machine learning, Digitization, Churn prediction.

1. Introduction

The rapid development of information technology and the popularization of various mobile devices are changing the traditional way of shopping, he has broken the original time and space restrictions on shopping places, so that consumers do not have to be limited to a fixed time, a fixed space to shop, people anytime and anywhere on the traditional supermarket digital platform shopping has become an increasingly common phenomenon. Although the number of users of traditional supermarket digital platforms is growing rapidly, it is also facing problems such as insufficient user stickiness and user churn. As for whether consumers choose to shop on digital platforms, scholars have studied the influencing factors of online shopping and the experience and results of online shopping, which reveals the causes and consequences of consumers' decision to use digital platforms to shop. Arora and Sahney (2019) [1] explain digital platform retail behavior from the perspective of planned behavior theory and technology acceptance models. Subsequently, researchers agreed that understanding shopping behavior on digital platforms requires consideration of environmental (e.g., time pressure), consumer (e.g., touch needs), and channel-related (e.g., price and convenience) factors (Aw et al., 2021 [2]; Mukherjee and Chatterjee, 2021 [3]). In addition, previous research results have shown that the decision to shop on digital platforms can be determined by different economic and non-economic motivations (e.g., efficient shopping, scarcity-curiosity, interest, and joy of discovery) (Aw, 2019 [4]; Royet al., 2022 [5]). The latest research has begun to explore the experiences and outcomes of shopping on digital platforms. For example, Chung et al. [6] (2022) found that shopping on digital platforms can widen the gap between expectations and product performance, ultimately reducing purchase intent. With the impact of a large number of national-level APP platforms such as JD.com, Tmall, Taobao and the growth rate of China's online shopping users has slowed down significantly, the "demographic dividend" has gradually disappeared, which shows that the cost of attracting new users is rising day by day. However, what factors will affect the loss of users in the process of consumers purchasing on traditional supermarket digital platforms? In recent years, scholars have begun to analyze and study the churn of digital platform users in industries such as online health communities [10], finance [11], and e-commerce [12]. Specifically, in the field of traditional business digital platforms, there are few relevant researches. For example, Guo Chengqi [13], Ren Hongjuan [14], and Zhu Chaona [15] used GBDT model, Stacking ensemble learning, HetGNN, and DRSA models for predictive analysis, respectively.

Although the prediction of user churn on online platforms has attracted the attention of scholars, a review of the existing literature can be found to be what factors should affect the user churn of traditional supermarket digital platforms. What is the weight of the influencing factors? These questions have not been well answered. The Beijing Regulations on the Promotion of Digital Economy [16] also clearly state that relevant departments should promote the digital transformation of traditional businesses such as supermarkets, and promote the digital promotion of traditional brands and time-honored brands, so as to promote the digital transformation of the life service industry. As an important guarantee of people's livelihood, traditional supermarkets have inherent advantages in terms of financial, material and human resources, and at the same time, they also occupy an irreplaceable position in the entire retail industry. Therefore, predicting the loss of users of traditional supermarket digital platforms can not only enrich the existing research, but also

have great significance for the sustainable development of traditional supermarkets and their digital platforms.

The RFM model is a classic model used to determine customer value in the field of marketing, and its main role is to analyze customer consumption behavior during the observation period (i.e., a period before the observation point) in order to find out the customers with important value. At present, RFM models have made some achievements in the application of predicting user churn. Wei Ling and Guo Xinyue [17] constructed an RFLP index system specifically for MOOC users' learning behavior and churn prediction based on the improved RFM algorithm model. Based on the research of Zhu Xuefang et al. [18], this study replaces the total consumption amount in the RFM model with the average consumption amount on the basis of the RFM model, aiming to eliminate the impact of the linear relationship between consumption frequency and total consumption amount on user value evaluation, so as to enhance the accuracy of the model's prediction.

2. Literature References

At present, the prediction methods of user churn at home and abroad are mainly summarized in two directions: one is to construct a user churn behavior model based on relevant theories [8, 18-21], and the other is to build a user churn prediction model based on machine learning and deep learning algorithms [7, 9, 14-15, 20]. The former mainly expresses the user's consumption feelings in a more detailed and comprehensive manner, and studies the influencing factors of user churn through statistical analysis methods such as structural equation model, focusing on subjective willingness. The latter is to evaluate the value of users by comprehensively considering their activity, loyalty, and spending power, and predict user churn in different scenarios, focusing on objective behavior. At the level of constructing user churn behavior models based on related theories, four main applied theories at home and abroad have been sorted out, including S-O-R theory [12, 19-21], PPM theory [8], RFM theory [11, 17-18], and theoretical choice theory [22]. At the same time, the study found that in addition to the most basic RFM model, scholars also selectively improved the model according to their own research content. Wei Ling [17] constructed an RFLP index system for user churn prediction of MOOCs, and proved that the improved model has higher prediction accuracy than the basic model through comparative experiments. Xing Shaoyan [18] used Zhihu Live as the data source to predict the loss of paid knowledge live broadcast users by constructing an RFML indicator system.

In the research on the construction of user churn prediction models based on machine learning and deep learning algorithms, especially in the fields of finance [11], telecommunications [7], and online APP [8], scholars at home and abroad have achieved certain results in using data mining and machine learning related technologies to predict user churn. Overall, the study broadly fell into three categories. First of all, there are two main types of machine learning, supervised learning and unsupervised learning, which are often used to solve the problem of binary classification, and many scholars use machine learning to predict user churn in the telecommunications field. Secondly, deep learning simulates the mechanisms of the human brain to interpret data, and has a wide range of applications in image speech recognition, natural language processing and other fields. Finally, the hybrid model-based method combines two or more models to make predictions, and the advantages and disadvantages of multiple models complement each other to improve the accuracy of prediction. For example, Lu Guangyue [23] and other scholars use improved KNN (K-nearest neighbors) and SVM (support vector machine algorithms) to predict the churn of telecom customers, and use weighted K-nearest neighbor classification for boundary samples and improved support vector machine classification for non-boundary samples, integrating the advantages of the two to classify. Most of the three types of research are contingency studies, and the relevant discussions lack the necessary theoretical basis to support them, and there is also a lack of theory-based and relevant countermeasures and measures research, which cannot make in-depth and accurate predictions for some fields to a certain extent. For China, the existing studies have studied user churn in different scenarios from the perspective of subjective willingness influencing factors and objective behavior prediction, and the fields involved are relatively rich, but most of the prediction studies do not take traditional supermarkets as research scenarios. In addition, the current research on the loss of users on digital platforms is very limited, especially in the prediction of user churn on traditional supermarket digital platforms, which cannot provide a reliable basis for platforms to identify user churn in the early stage and make corresponding retention measures.

In terms of practical guidance, the Regulations on the Promotion of the Digital Economy in Beijing [16], adopted by the Standing Committee of the 15th National People's Congress, also clearly pointed out that the commerce department should work with relevant departments to promote the digital upgrading of traditional businesses such as supermarkets, promote the digital promotion of traditional brands and time-honored brands, and promote the digital transformation of the life service industry. As the "reserve army" of people's livelihood, traditional supermarkets have inherent advantages in terms of financial, material and human resources, and also occupy an irreplaceable position in the entire retail industry. According to the 51st CNNIC data, the scale of online shopping users in China has reached 841 million, although the scale of online shopping users is still growing, but the growth rate has slowed down significantly. It can be seen that the "demographic dividend" has gradually disappeared, however, it is still unclear how to predict the loss of users of traditional supermarket digital platforms and retain users through practical measures. This paper attempts to enrich research in this field. On the one hand, in order to improve the accuracy of the prediction model on the user churn prediction effect, it is necessary to construct a systematic theoretical framework. The framework not only analyzes customer behavior characteristics, but also predicts future short-term behavior based on customers' past purchase behaviors. On the other hand, in order to let theory guide practice, it is necessary to support the theory that is suitable for the field of user churn, and then summarize the specific measures that can effectively implement the theory. Therefore, this study comprehensively collects, reviews and analyzes the literature of domestic and foreign scholars on user churn prediction in various scenarios, and uses XGBoost and this algorithm to construct a theoretical model of user churn on traditional supermarket digital platforms, and uses RFM model and text sentiment analysis to select user characteristics. At the same time, this paper identifies and

refines the best practices that can effectively implement the theoretical model in the past responses, in order to grasp the behavior patterns of users, improve user loyalty, and provide useful references for platforms to attract new users, maintain old users, and reduce the user churn rate of digital platforms.

3. User Feature Selection

3.1. User Value Characteristics-RFM Model

The RFM model is a classic marketing model that is used to identify the value of a customer. The three indicators of RFM are user activity indicator-Rencency (the most recent consumption), user loyalty indicator-Frequency (consumption frequency), and user spending power indicator-Monetary (consumption amount), and the first letter combination of the three indicators. Among them, Recency refers to the time of the customer's last purchase, Frequency refers to the number of purchases made by the customer during the observation period, and Monetary refers to the total consumption amount of the customer during the observation period. Through the comprehensive analysis of these three metrics, the RFM model can identify customers with significant value. It is generally believed that users with a short time interval of the most recent consumption and a large number of recent purchases have a higher recognition of the product, so the churn tendency is low. Conversely, users with a longer interval between the most recent purchases and a smaller recent consumption frequency and amount have a higher tendency to churn and have a lower value to the platform.

In this study, this paper changes the total consumption amount in the RFM model to the average consumption amount, aiming to eliminate the impact of the linear relationship between the consumption frequency and the total consumption amount on the user value evaluation, and by replacing the total consumption amount with the average consumption amount, the user's value can be more accurately evaluated, so as to improve the accuracy of prediction.

3.2. Characteristics of User Value Evaluation—Scoring and Review Text Sentiment

In the context of this study, user evaluation specifically refers to the fact that after users purchase products on the digital platform of traditional supermarkets (taking Ehime Orange on Wumart supermarket online mini program as an example), users can give ratings and textual evaluations on the price, freshness and purchase feelings of the products in the review system. These reviews can help other users better understand the pros and cons of the product, so they can make more informed purchasing decisions. In the review system, the rating is usually presented in the form of stars, generally 1-5 stars, of which 5 stars represent the highest evaluation, 1 star represents the lowest evaluation, the numerical or star-level scoring is concise and direct, and the operation is convenient, while the evaluation in the form of text can more carefully and comprehensively express the user's purchase feelings, which is the withdrawal of the user's emotional attitude. These reviews are also very valuable for businesses to understand the needs and feedback of users, so as to improve products and services, and increase user satisfaction and loyalty. The more positive the sentiment tendency of the review text, the higher the customer's consumer satisfaction, the greater the likelihood of continuing to buy on the platform, and the lower the risk. Conversely, users with a more negative emotional tendency in the review text have lower satisfaction, less likely to continue paying, and higher risk of churn. Therefore, there is some value in dissecting the personal emotions implicit in the comments and incorporating them into the user churn prediction model.

Text sentiment analysis, also known as opinion mining, is the process of collecting, processing, analyzing, inducting, and reasoning about subjective texts with emotional overtones. It is an interdisciplinary discipline involving multiple research fields such as artificial intelligence, machine learning, data mining, and natural language processing. At present, many scholars have conducted research on user reviews. For example, Feng Jianying et al. [24] took the e-commerce review data of fresh agricultural products as the object, based on different models and technologies such as dictionaries, LDA topics, and machine learning, and conducted text analysis, focusing on the analysis of the influence mechanism of review text on the sales of fresh agricultural products, the information and emotional attributes of reviews, and the impact of review contradiction on the online sales of fresh agricultural products. The research of Liu Ji [25] shows that the model combining BERT and BiLSTM (M2BERT-BiLSTM) has certain effectiveness in the sentiment analysis of unbalanced text of network public opinion. The model can effectively process unbalanced text data and improve the accuracy and reliability of sentiment analysis. At the same time, the study also gives suggestions on the direction of online public opinion guidance for the "new coronavirus pneumonia" incident, which provides a valuable reference for enterprises and government departments.

Text sentiment analysis is primarily based on dictionaries or machine learning methods. Compared with complex machine learning-based methods, the dictionary-based method can quickly and accurately quantify sentiment, so this study uses a dictionary method to achieve sentiment analysis.

This study is based on the ontology database of emotional vocabulary of Dalian University of Technology as the basic emotional dictionary, and each word corresponds to the polarity of an emotion under each type of emotion, where 0 represents neutral, 1 represents positive emotion, and 2 represents negative emotion. Each emotion word has five levels of emotional intensity: 1, 3, 5, 7, and 9, which is much more refined than other dictionaries. In order to facilitate the subsequent calculation, the emotional intensity of 1, 3, 5, 7 and 9 was assigned 1, 3, 5, 7 and 9 points, respectively, and the polarity value representing negative emotion was assigned to -1. The quantification method of the sentiment value of the emotion word is shown in equation (1): where word-sentiment represents the sentiment value of the emotion word, polarity represents the sentiment polarity, and degree represents the sentiment intensity. In order to get a more accurate sentiment value, it is necessary to extract the emotional words that appear more frequently in comments. Considering the influence of adverbs of degree and negation on the emotional intensity and polarity of affective words, it is necessary to formulate reasonable rules to revise them. Based on the above rules, the sentiment value for each comment text is calculated as follows: text-sentiment=p * weightadv * word-sentimenti. Based on the 189 degree-level words in the Sentiment Analysis Vocabulary Collection (Beat Version), the five degree-level words ("extreme/most", "very", "comparatively", "slightly" and "owed") were weighted, and

the polarity of these words was quantified as an index. In the environment of mood words, a detection window of size 5 is set, and based on this, according to the original emotional polarity and intensity of the mood words, the weights corresponding to the degree adverbs and negative words in the detection window are multiplied, and the detailed weight settings and some word examples are shown in Table 1.

Table 1. Examples of adverbs of degree and negative words and their corresponding weights

category		examples of words	weight	number
extentad verb	extremely/ most	the most, the extreme, the abnormal,	2.00	11
	very	Very, extraordinary, really, especially, too specially, so much	1.50	349
	compare	more,	1.25	18
	little	slightly, quite well, somewhat, a little,	0.50	65
	owe	not much	0.25	1
Negative words		no, can't, won't, don't	-1.00	281

4. Model Building

4.1. Algorithm Selection

At this stage, logistic regression, decision tree, random forest, XGboost and other machine learning classifiers are the most frequently used machine learning classifiers, among which the multi-classifier system can usually obtain superior generalization performance than the single classifier, so it is favored by more and more scholars. In this paper, by referring to the research of Zhu Xuefang [18] et al., it is concluded that the multi-classifier system represented by XGboost has obvious advantages in prediction, and on this basis, the application of the method on digital platforms is carried out.

4.2. Model Evaluation Metric Selection

In this study, the user churn prediction model of the traditional supermarket digital platform is comprehensively evaluated, and the following four situations may occur when the churn prediction of the digital platform users is carried out, as shown in Table 2.

Table 2. Confusion matrix

Actual type	Forecast type	
	Predicted as churn	Predicted to be non-churned
Actually churn	TP	FN
Actually non-churned	FP	TN

In Table 2, TP indicates that the predicted is churned, and in fact churned; FN indicates that the prediction is a non-churned user, which is actually a churned user; FP predictions are churned users, which are actually non-churned users; TN said that the prediction was a non-churned user, and in fact a non-churned user.

For traditional classification algorithms, specificity and sensitivity are generally used as evaluation indicators, but for unbalanced datasets, it is inaccurate to use specificity, sensitivity, and accuracy to evaluate the performance of the classifier. In this paper, accuracy, precision, recall, and F1 are used as the evaluation indicators for the performance of the classifier for the classification of unbalanced data.

5. Empirical Part

5.1. Data Collection

The research data in this paper comes from the digital platforms of traditional supermarkets, and the shopping reviews of Ehime Orange on Wumart's online mini program are used as the data source.

The first part of the collected content is the user value characteristic data, including the date of the user's last purchase on the online platform, the average number of purchases per month on the online platform of traditional supermarkets, and the average amount of each payment. The second part is the feedback of consumers after online shopping, including the rating and review text of the product, service, etc. Among them, the user value characteristic data cannot be directly crawled, so it is necessary to crawl the purchase records of a certain type of product in the user's dynamic list, collect the date and payment amount corresponding to each user's purchase behavior, and then carry out simple statistics and calculations to obtain the date of the user's last participation in the platform purchase, the total number of purchases and the total amount paid, and further calculate the variables such as the average monthly consumption number and the average consumption amount. The variables are described in Table 3.

Table 3. specifies the variables

Description of the variable	Variable name	Description of the variable
User price Value feature	The interval between the most recent consumption time	The number of days from the date of the most recent purchase on the Wumart online platform to the data observation point
	The average number of monthly spending	The average number of purchases made on Wumart's online platform per month
	The average amount spent	The average amount spent on the Wumart supermarket online platform
User reviews valence characteristics	Score	Five-star rating, only full-star ratings are supported
	Comment on text sentiment	The sentiment value of the comment text is calculated based on the sentiment dictionary method

The obtained sample data was collated to remove those that were not meaningful (e.g., comments with only a few punctuation marks or numbers) to result in a sample of 593 users. These samples can be used for subsequent data analysis and modeling, helping us to better understand users' purchase behaviors and evaluations, so as to formulate more effective marketing strategies and improve user satisfaction and loyalty.

Based on the preliminary investigation of the purchase behavior characteristics of users on traditional supermarket digital platforms, this study sets the user churn time threshold

to 90 days, that is, if the user's last purchase time is greater than 90 days, it is a churned user, and less than 90 days is a non-churned user. Churned users are 1 and non-flow users are 0.

Table 4. Partial comments for crawling

User ID	Score	Comment time	Comment content
Miao***7	5	2022/9/23	Fantastic, fresh, no bad fruit, sweet!
****n	1	2022/11/14	There are 2 rotten fruits, and the time is too long for the wife to swallow
h***7	1	2022/10/27	It was delicious, better than I expected, this is the second time I bought it, it was good overall, there were a few with less moisture, which was acceptable
****C	1	2022/11/11	Find customer service as soon as the goods are received, and it hasn't been solved for two days, so will you just wait?
Hong*** meaning	3	2022/10/30	The moisture is large, so the sweetness is average

The essence of the churn prediction problem is a binary classification problem, in which if the number of samples in the two categories is far apart, there will be a serious problem when training the model. In real life, the proportion of churned users among the purchasing users on the digital platform of traditional shopping malls is much smaller than that of non-churned users, which causes the distribution of these two types of samples to be extremely uneven, which affects the classification results of the classifier, so it is necessary to balance the sample data. However, the traditional SMOTE algorithm synthesizes a new sample by selecting a small number of samples without considering the surrounding samples, which easily causes the new small number of samples to overlap with most of the samples in the surrounding samples, resulting in a large amount of noise.

The data cleaning technology ENN is widely used in the processing of overlapping sample data, so the SMOTE algorithm and ENN can be combined to form a pipeline, that is, oversampling is carried out first and then data cleaning is carried out to improve the shortcomings of the data processed by the SMOTE algorithm. In this paper, the SMOTE+ENN algorithm is used to synthesize minority samples to improve the optimization performance of the classifier [26].

The SMOTE+ENN algorithm is mainly based on SMOTE oversampling, and the overlapping data is cleaned through the ENN algorithm to achieve the purpose of balancing the sample data. The specific steps are as follows: firstly, the SMOTE algorithm is used to generate new minority samples to obtain an expanded dataset. Then, each sample in the new dataset was predicted using the KNN (generally K=5) method, and if the prediction result was significantly different from the actual result, the sample was excluded.

5.2. Comment Text Sentiment Value Calculation

First, the text data was denoised, including converting some Chinese Traditional Chinese to Chinese Simplified Chinese. Remove facial emojis made up of various lines and special symbols; Convert some English words into Chinese, such as great, nice, good; Convert some simple pinyin sounds such as bang (stick) into their corresponding characters.

In this study, the Jieba Chinese word segmentation module in Python was used to process the obtained comment text with sentence breaking, word segmentation, stop word screening and part-of-speech marking. Finally, based on the complete sentiment dictionary and the calculation steps in section 3.2, the quantitative score of comment sentiment is obtained. Dalian Institute of Technology emotional vocabulary question bank (part) see Table 5.

Table 5. Emotional vocabulary ontology database of Dalian University of Technology (partial)

Words	Types of parts of speech	The number of word mean-ings	Sentiment classification	strength	polarity
fresh	noun	1	PB	5	1
rot	adj	1	NN	3	-1
grudging	verb	2	NN	5	0
stinking	adj	1	ND	5	-1
fresh	adj	1	PB	5	1
disappointed	adj	1	NJ	5	-1
fantastic	adj	1	NJ	7	-1

5.3. Experimental Results and Conclusions

5.3.1. Model Training and Testing Results

In this study, four user value level characteristics (recent consumption interval, average monthly consumption times, average consumption amount) and two user evaluation level characteristics (rating, review text sentiment) were all the characteristic variables. After the data under all variables were standardized, the dataset was divided into layers, and 70% of the lost samples and 70% of the non-lost samples were randomly selected as the dataset, which were processed by SMOTE to become the balanced samples, and then the prediction model was constructed based on the XGBoost machine algorithm.

Firstly, in order to verify the effectiveness and robustness of the user churn prediction method based on the sentiment characteristics of scoring and comment text based on user value characteristics, this paper intends to use two different datasets for comparative experiments. However, since the classification standard of user churn uses the user's most recent time consumption interval, it is necessary to classify the model and eliminate the importance of the characteristic variable. Dataset 1 is the data under the user value characteristics of the traditional supermarket digital platform (except for the user's recent consumption time interval), and dataset 2 is the data of all characteristics (except for the user's recent consumption time interval). Based on the analysis of the performance of dataset 1 and dataset 2 on the same machine learning model, it can be concluded that the accuracy, recall, precision and F1 value under dataset 2 are all above 0.9, which has good accuracy and recall, indicating that the classifier has good performance. At the same time, dataset 2 is higher than dataset 1 in terms of accuracy, recall, precision, and F1 to varying degrees, that is, the combination of scoring and review text sentiment features to establish a prediction model based on scoring and review text sentiment features can better solve the problem of user churn prediction on traditional shopping mall digital shopping platforms, and verify the significant effect of scoring and review text sentiment features on user churn prediction.

Table 6. Test results

		Accuracy	Summons rate	Precision	F1
Data set 1	Training set	0.97	0.97	0.942	0.956
	Cross-validation sets	0.967	0.967	0.946	0.955
	Test set	0.864	0.864	0.746	0.8
Data set 2	Training set	0.951	0.951	0.953	0.939
	Cross-validation sets	0.931	0.931	0.915	0.916

5.3.2. Importance of Characteristic Variables

On the basis of the above training, this paper finds that the prediction effect of the XGBoost model on user churn has obvious advantages, which verifies the results of scholar Zhu Xuefang [18]. In the training process of the XGBoost model, this paper normalizes the feature variables to determine and identify the contribution of each feature variable in user churn prediction, and then analyzes the key feature variables by outputting the importance ranking of the feature variables, as shown in Figure 1.

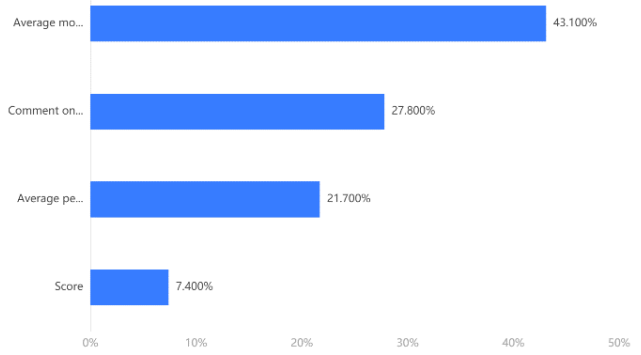

Figure 1. Importance of Feature Variables

As can be seen from Figure 1, the extracted feature variable that contributes the most to the prediction of user churn is the average number of monthly consumptions, which is consistent with daily experience and common sense, that is, the more times the average average user pays to participate in knowledge live broadcast per month, the greater the user stickiness and the lower the risk of churn. We can see that the sum of the importance of the top two and the sum of the top three is 70.9% and 92.6%, which indicates that the feature variables extracted in this study are reasonable. Moreover, from the perspective of the importance distribution of each feature variable at different time points, there is neither an overly important feature variable nor a feature variable with 0 importance, which indicates that the feature variables extracted in this study are reasonable. Therefore, from this point of view, these characteristic variables can be used to predict user churn well.

5.3.3. The Type of Churned User

On this basis, the k-means clustering technology is used to find the similarity between lost users in digital platforms such as traditional shopping malls, so as to construct differentiated lost user groups, which provides convenience for different lost user groups to formulate personalized retention strategies.

K-means clustering divides the sample set into several groups according to the distance between the samples, so that the distance between the points in the group is as small as possible, and the distance between groups is as large as possible. On this basis, the k-means clustering function provided by SPSS26 is used to analyze the algorithm, and finally two different user groups are obtained. The cluster center is the core of each taxon and is representative. By comparing and analyzing the values of the center point of each group on each characteristic factor, the different characteristic characteristics of each group can be analyzed.

It can be seen from Table 7 that the characteristics of average monthly consumption, average consumption amount, sentiment value of comment text, and rating users are significant in the clustering of user churn, indicating that these four characteristic indicators can be used as clustering indicators, and all users are divided into two types through clustering.

Table 7. Cluster Difference Analysis

	Cluster Category (Mean±SD)		F	P
	Category 2 (n=420)	Category 1 (n=17)		
The average number of monthly spending	1,063±0,268	3,124±1,005	652.417	0.000***
The average amount spent	32.748±10.553	186.435±63.892	1499.224	0.000***
Score	4,233±1,561	5.0±0.0	4.091	0.044**
Comment on the sentiment value of the text	14.14±20.807	26.824±29.267	5.86	0.016**
Note: ***, **, and * represent the significance levels of 1%, 5%, and 10%, respectively				

Table 8. Center points for each churned user group

		Clustering	
		Cluster category 1	Cluster category 2
User Characteristics	The average number of monthly spending	3.12	1.06
	The average amount spent	186.44	32.75
	Comment on the sentiment value of the text	26.82	14.14
	Score	5.00	4.23
Percentage (%)		3.89	96.11

6. Conclusion

Based on the above research, the conclusions of the study are as follows:

1) The ratings and comments of purchasing users have a significant effect on the prediction of user churn.

In terms of the selection of predictors, the sentiment characteristics of user ratings and review texts collected from traditional supermarket online platforms have a significant impact on the prediction of user churn behavior, and these characteristics are integrated into the prediction model of user churn behavior, and the model will have a better prediction effect.

The user value characteristics obtained on the basis of the

RFM model are only a comprehensive reflection of customer behavior and a set of objective data, which cannot truly reflect the real feelings of customers shopping on traditional supermarket digital platforms, so there are certain limitations in predicting user departure. The sentiment of shopping rating and review text is an expression of customers' subjective emotions and feelings, which can reflect customers' satisfaction with shopping on digital platforms and their willingness to continue to pay to a certain extent, so these two characteristic data have a certain impact on user churn prediction. At the same time, the user value characteristics extracted from the user reviews of Ehime orange also have a certain reference value for the user loss of other products.

2) The contribution of characteristic variables in user churn prediction was compared and analyzed, and it was concluded that the average monthly consumption number accounted for the largest proportion in user churn prediction, followed by comment text sentiment.

The importance of each feature variable is ranked by XGBoost, and the importance of each feature variable in the prediction can be obtained according to the contribution of each feature variable to the user churn prediction. On this basis, after ranking the importance of different characteristic variables, it is concluded that the importance of these characteristic variables is the average monthly consumption times, comment text sentiment, rating, and average consumption amount. Different types of churned users may have different reasons and characteristics, and at the same time, they can be divided into different types according to the importance of feature variables, and corresponding retention strategies can be given for different types to improve the retention effect.

7. Management Implications

In view of the problem of user churn of the digital platform of traditional supermarkets (hereinafter referred to as the "digital platform"), this paper divides the lost users of Wumart online platform into experiential users and economic users according to the key characteristic factors such as the average monthly consumption, the average consumption amount, and the sentiment of the comment text, and proposes corresponding retention strategies for different lost users.

1) Experiential users: Digital platforms should focus on optimizing the product supply chain and after-sales service system.

For experiential users, the characteristic of this group is that the user's review text sentiment will fluctuate greatly with the purchase experience, that is, when the user's satisfaction is not high, his review text is more negative, which will lead to the loss of customers. Therefore, for this part of the customer, the platform should attach great importance to the customer's first contact with the platform, optimize the interface design of the platform, and highlight the characteristics of the product name, so as to leave a good first impression on the first-time customer. In addition, the platform must not only have enthusiastic and professional customer service, but also optimize its own supply chain, in addition to packaging to have a high product quality control, in the process of transportation, for different products to ensure that the time, quality, quantity, and safety to the customer. During the period, there will inevitably be some losses, and the platform also needs to have a relatively complete after-sales service system to protect the rights and interests of consumers and give customers a good shopping experience. In order to give customers a better shopping experience

Based on this, the platform can expand its own product range and pay attention to customer care for the old user group, which is an important asset of the enterprise, and their loyalty and word-of-mouth influence are very important. Therefore, enterprises need to take measures to maintain the relationship with old users, and can use message reminders and other methods to establish contact with these users, so as to maintain the relationship between the platform and old customers. At the same time, when establishing contact with old users, it is necessary to analyze the business scenario and actual situation to avoid user loss due to excessive marketing.

2) Economic users: digital platforms launch promotional activities from time to time to increase customers' willingness to spend.

For economic users, the significant feature of this group is that the average consumption amount is small, indicating that such users have a demand for the product, but they are more cautious in terms of money investment. For such users, the platform can launch some promotional activities from time to time, send coupons, discounts, gifts and other ways to let them fully enjoy the preferential services, little by little increase the customer's emotional identity and value recognition of the product, and then increase their willingness to consume, increase the possibility of subsequent purchases.

References

[1] Arora, S. and Sahney, S. (2019), “Examining consumers' webrooming behavior: an integrated approach”, Marketing Intelligence and Planning, Vol. 37 No. 3, pp. 339-354.

[2] Aw, E.C.X., Basha, N.K., Ng, S.I. and Ho, J.A. (2021), “Searching online and buying offline: understanding the role of channel-, consumer-, and product-related factors in determining webrooming intention”, Journal of Retailing and Consumer Services, Vol. 58, 102328.

[3] Mukherjee, S. and Chatterjee, S. (2021), “Webrooming and showrooming: a multi-stage consumer decision process”, Marketing Intelligence and Planning, Vol. 39 No. 5, pp. 649-669.

[4] Aw, E.C.X. (2019), “Understanding the webrooming phenomenon: shopping motivation, channel-related benefits and costs”, International Journal of Retail and Distribution Management, Vol. 47 No. 10, pp. 1074-1092.

[5] Roy, Subhadip et al. I "showroom" but "webroom" too: investigating cross-shopping behaviour in a developing nation [J]. INTERNATIONAL JOURNAL OF RETAIL & DISTRIBUTION MANAGEMENT, 2022, 50(12):1475-1493.

[6] Chung, Sorim et al. It is different than what I saw online: Negative effects of webrooming on purchase intentions [J]. PSYCHOLOGY & MARKETING, 2022, 39(1):131-149.

[7] A Kehong, HU Xiaodong. Telecom user churn prediction method based on GAN data reconstruction [J]. Telecommunications Science, 2023, 39(3):135-142.

[8] YUAN Shunbo, ZHANG Hai, DUAN Hui. Research on the Influencing Factors of User Churn Behavior of Mobile Government APP from the Perspective of PPM [J]. Journal of Information, 2021, 40(2):182-188.

[9] HU Yanfang, XIONG Wen, GAO Wei. Online game user churn prediction method based on Spark platform [J]. Computer Engineering and Science, 2022, 44(10):1730-1737.

[10] Wang Ruojia, Yan Chengxi, Guo Fengying, et al. Research on user churn prediction in online health community based on user portrait [J]. Data Analysis and Knowledge Discovery, 2022, 6(2):80-92.

[11] ZHAO Hong, DING Ru. Comparative study on feature extraction methods for user churn prediction in Internet financial enterprises [J]. Frontiers of Engineering Management Science and Technology, 2018, 37(6):61-66.

[12] CUI Linan, QI Lili. Research on consumer churn behavior of audio reading e-commerce platform from the perspective of SER [J]. Journal of Tianjin Vocational College of Commerce, 2022, 10(3):67-75.

[13] GUO Chengqian. Research on churn prediction of online shopping users based on data mining [D]. Jilin University, 2016.

[14] Ren Hongjuan. Research on churn prediction of online shopping customers based on Stacking ensemble learning [D]. Guangxi University of Science and Technology, 2020.

[15] ZHU Chaona. E-commerce user churn prediction based on HetGNN and DRSA model [D]. Shanghai University of Finance and Economics, 2021.

[16] Ren Qian. Helping the construction of the "Beijing Benchmark" of the global digital economy——The first instance

[17] Wei Ling, Guo XinyueMOOC User Churn Prediction Based on Improved RFM and GMDH Algorithm [J]. China Distance Education, 2020, 0(9):39-43+61+76-77.

[18] Xing Shaoyan, Zhu Xuefang. An empirical study on the prediction of user churn of paid knowledge live streaming [J]. Journal of Information Resources Management, 2022, 12(4):121-130+140.

[19] YUAN Shunbo. An empirical study on the loss intention of social reading users [J]. Zhejiang Academic Journal, 2022(2):99-110.

[20] LIU Guiqin, XU Xinhua. Discussion on the influencing factors of library user churn based on machine learning [J]. New Century Library, 2020, 0(1):9-13.

[21] LIU Lei, YUAN Shunbo, ZHANG Hai. An empirical study on the churn intention of online video paying users [J]. Journal of Jiaxing University, 2020, 32(3):96-106.

[22] CHEN Yu, HUANG Liangfeng. Research on user churn behavior of e-book reading client from the perspective of rational choice theory [J]. Library Forum, 2019, 39(9):118-126.

[23] Lu Guangyue, Zhang Hongjian, Yan Zhenguang, Wu Yang. Journal of Xi'an University of Posts and Telecommunications, 2019, 24(02):21-25. DOI:10.13682/j.issn.2095-6533.2019.02.005.

[24] Feng Jianying, Wu Dandan, Wang Bo, et al. A review of the impact of Chinese online review text analysis on the e-commerce of fresh agricultural products [J]. Transactions of the CSAM, 2021(S1 vo 52): 504-512.

[25] LIU Ji, GU Fengyun. Non-equilibrium text sentiment analysis of network public opinion based on BERT and BiLSTM hybrid method [J]. Journal of Information, 2022, 41(04):104-110.

[26] Sun Dan, Shi Weili, Rao Lanxiang, Meng Shasha, Guo Xiaoming, Li Yilun. Credit card fraud detection method based on improved hybrid sampling and XGBoost algorithm [J]. Computer and Modernization, 2022(09):111-118.