Engineering and Technology Journal e-ISSN: 2456-3358

Volume 11 Issue 08 August-2026, Page No.- 11256- 11272

DOI: 10.47191/etj/v11i08.17, I.F. – 8.595

Exploring the Relationship Between Visit Frequency and Customer Retention in Study Cafes Using AI-Based Predictive Modeling

Joonghyun Park[1], Seungchan Lee[3], Hoon Ko[2,*]

[1]Software-Oriented University, Sunmoon University, 70, Sunmoon-ro 221 beon-gil, Tangjeong-myeon, Asan, 31460, Chungcheonnam-do, Republic of Korea.

[2]Department of Computer Engineering, Sunmoon University, 70, Sunmoon-ro 221 beon-gil, Tangjeong-myeon, Asan, 31460, Chungcheonnam-do, Republic of Korea.

[3]GOSCA Ltd., Seoul, Republic of Korea.

ABSTRACT: This study empirically investigates actual customer repurchase behavior by analyzing behavioral log and payment data collected from study cafe users in an attendance-based learning service environment. Unlike prior studies that primarily relied on survey-based measures, this study integrates attendance records, payment data, and explainable artificial intelligence (XAI) techniques to provide a data-driven understanding of customer retention behavior. It compares the predictive performance of traditional regression models with machine learning approaches. Specifically, logistic regression, random forest, XGBoost, and neural networks were employed, with SHAP analysis applied as an XAI technique. The results indicate that visit frequency has a significant positive effect, supporting H1, while stay-duration variables show only limited and inconsistent effects, providing partial support for H2 and H3. Total payment in May negatively affects subsequent repurchase, suggesting possible saturation or substitution effects, thereby supporting H4. Age demonstrates a negative effect (supporting H5), and regional differences are captured by the XGBoost model (supporting H6). In terms of predictive performance, machine learning models outperformed logistic regression, with XGBoost achieving the strongest overall results among the evaluated models (ROC-AUC = 0.676; PR-AUC = 0.642). Overall, this study contributes to the literature by presenting empirical evidence based on behavioral data and by highlighting the practical interpretability of integrating XAI techniques. From a managerial perspective, the findings provide actionable insights for designing customer retention strategies based on visit frequency, spending behavior, and regional characteristics, thereby supporting AI-driven decision-making in attendance-based service environments.

KEYWORDS: Customer Retention; Machine Learning; Customer Analytics; Behavioral Data Mining; Predictive Modeling; Explainable AI

INTRODUCTION

In today's competitive service industry, repeat customers—those who engage in repurchase behavior—are considered critical assets directly tied to a company's long-term profitability. Among attendance-based commercial service environments, study cafes represent a learning-oriented service sector in which customer visit frequency and usage patterns may substantially influence customer retention and repurchase behavior. In attendance-based service sectors such as study cafes and fitness centers, customer visit frequency and usage patterns have been shown, through industry practice and prior research, to significantly influence repurchase outcomes (retention). However, much of the existing literature has relied on survey-based self-reported data, while analyses employing objective measures such as attendance records or payment records remain relatively scarce. Although recent studies have increasingly applied machine learning techniques to customer retention prediction, empirical research integrating attendance logs, transactional payment data, and explainable artificial intelligence (XAI) within attendance-based service environments remains limited. Consequently, the behavioral mechanisms linking service usage patterns to actual repurchase outcomes have not been sufficiently examined from both predictive and interpretive perspectives. Recent advances in AI and big data analytics now enable more precise measurement and analysis of service usage, thereby creating new opportunities for personalized marketing strategies. Accordingly, this study integrates attendance records and payment data to empirically examine the relationship between visit frequency and actual repurchase behavior, while also exploring the applicability of AI-based predictive models. To further enhance predictive performance and model interpretability, machine learning models and XAI techniques are jointly employed and evaluated using real behavioral log data collected from study café users. To this end, the following hypotheses (H1–H6) are proposed.

HYPOTHESES

(H1) A higher number of visit days increases the likelihood of repurchase.

(H2) Total stay duration has a positive but limited effect on repurchase.

(H3) Average stay duration positively affects repurchase, but the effect is conditional or inconsistent.

(H4) Higher total payment decreases the likelihood of subsequent repurchase (saturation/substitution effect).

(H5) Repurchase rates differ significantly across age groups (with older customers generally showing lower repurchase rates).

(H6) Repurchase rates differ significantly across regions.

This study seeks to overcome the limitations of survey-based research by demonstrating the applicability of quantitative analyses that incorporate service usage records and payment data, thereby validating the effectiveness of data-driven marketing strategies. Specifically, the objectives of this research are threefold: (1) to address the limitations of prior survey-based studies, (2) to empirically examine the relationship between visit frequency and actual repurchase behavior using service usage records, and (3) to evaluate the applicability of AI-based predictive models. By integrating behavioral log data, transactional records, machine learning models, and XAI techniques, this study extends the customer retention literature and provides practical insights for data-driven retention management in attendance-based service environments. The remainder of this paper is structured as follows: Section 2 reviews the theoretical background and related literature; Section 3 outlines the proposed methodology; Section 4 presents the empirical results; and Section 5 concludes with key implications.

Background and Related Works

Background

Repurchase intention refers to a customer's tendency to repeatedly purchase the same product or service, and it is closely associated with customer satisfaction, brand loyalty, and service quality [1] [2]. In attendance-based commercial service environments such as study cafes, customer retention and repurchase behavior are considered important indicators directly associated with long-term operational sustainability and profitability. In service industries, repurchase intention is directly tied to long-term profitability, making its accurate measurement and prediction a critical research priority. Prior studies have identified visit frequency and usage patterns as key antecedents of repurchase intention. For example, Paper [3] examined fitness center users and found a significant positive correlation between visit frequency and re- enrollment, with visit patterns during the first quarter after joining serving as crucial predictors of long-term retention. Similarly, Paper [4] analyzed class participation data and demonstrated that students with higher participation frequency achieved significantly higher re-enrollment rates and GPAs, thereby suggesting the positive relationship between attendance and retention outcomes. Alongside attendance-oriented studies, customer retention analysis using artificial intelligence (AI), behavioral log analysis, and big data analytics has recently attracted considerable attention in commercial service industries, including retail, subscription services, and e-commerce platforms. Representative approaches include the RFM (Recency, Frequency, Monetary) framework and machine learning–based churn prediction. Paper [5], for instance, utilized e- commerce payment data to predict customer churn with the XGBoost model and emphasized the potential for personalized marketing strategies. The same study further addressed the data imbalance problem in churn prediction by applying oversampling techniques such as SMOTE, ADASYN, and GNUS in combination with Random Forest and XGBoost, reporting that the XG-Boost–SMOTE combination achieved the highest accuracy. Recent studies have further demonstrated that behavioral log data and transactional records can provide more objective and practically useful insights than conventional survey-based approaches in customer retention prediction. Prior attendance-based studies have largely focused on educational outcomes, participation, and retention within academic settings, whereas customer retention behavior in commercial learning environments has received comparatively less attention. Moreover, attendance records and transactional payment data have rarely been examined together within a unified analytical framework. Despite these advances, little empirical evidence is available on how attendance records and payment data can be integrated to explain actual repurchase behavior in attendance-based commercial service environments such as study cafes.

Related Works

A wide range of studies have explored the relationship between learning space visits, attendance patterns, and academic performance. For instance, Paper [6] analyzed correlations among visit frequency, grades, and re- enrollment, suggesting that more frequent visits were associated with better outcomes. Paper [7] examined library seat reservations, class attendance, and re-enrollment, demonstrating that learning experiences contributed to higher retention. Paper [8] investigated the role of library usage purpose, suggesting that usage intentions influenced both attendance patterns and academic achievement. Addressing post-pandemic shifts, Paper [9] com- pared offline, online, and blended course formats, confirming that changes in instructional delivery substantially affected attendance and performance. Building on these traditional approaches, more recent research has employed machine learning models on large-scale administrative attendance records, demonstrating high predictive accuracy. Paper [10] applied ML techniques to attendance records to predict academic outcomes, illustrating the potential of data-driven approaches. Paper [11] examined the relationship between academic attendance frequency, achievement, and dropout rates, integrating ML-based features to highlight the feasibility of attendance prediction. From a system design perspective, Paper [12] proposed an LMS-based early warning system that leverages ML to identify potential dropouts in advance. Paper [13] compared multiple ML algorithms for predicting attendance and academic achievement across different student groups, emphasizing the importance of selecting appropriate models. Paper [14] applied ML classification models to course-specific attendance records and clarified their associations with academic outcomes. Finally, Paper [15] provided a comprehensive analysis of library attendance, grades, GPA, and retention, demonstrating that attendance patterns directly influence academic performance. Beyond educational environments, recent studies have increasingly applied machine learning and data-driven analytics to customer retention, churn prediction, and behavioral analysis within commercial service industries. Paper [20] reviewed AI-based customer churn prediction approaches and emphasized the growing importance of machine learning techniques in retention management. Paper [21] analyzed transactional customer behavior in retail environments and demonstrated that behavioral purchase patterns strongly influence customer loyalty and retention outcomes. Paper [22] comparatively evaluated multiple churn prediction models and reported that predictive performance varies significantly depending on algorithm selection and behavioral feature design. These findings indicate that spending behavior should be interpreted together with the characteristics of the service environment rather than as a standalone indicator of future repurchase. In attendance-based service environments such as study cafes, higher total payments may reflect more intensive service usage or temporary service saturation, which can reduce the need for immediate subsequent purchases despite greater cumulative spending. In addition, Paper [23] highlighted the importance of explainable AI (XAI) techniques in improving the interpretability and practical applicability of machine learning–based prediction systems. TTaken together, prior studies suggest that attendance-based behavioral patterns, transactional records, and AI-driven predictive analytics can provide meaningful insights into customer retention and repurchase behavior. More specifically, attendance-related studies consistently reported positive relationships between participation frequency and retention-related outcomes, whereas commercial retention studies emphasized the predictive value of transactional and behavioral data. In addition, in learning-oriented service environments such as study cafes, service usage patterns may vary across age groups because learning needs and usage purposes differ among customers. Therefore, customer age represents a meaningful factor when examining differences in repurchase behavior within attendance-based commercial service environments. However, most attendance-based studies focused on educational performance rather than customer repurchase behavior, while commercial retention studies rarely incorporated attendance-related behavioral indicators. Consequently, empirical studies integrating attendance records, payment data, and explainable AI techniques within learning-oriented commercial service environments such as study cafés remain limited. Table 1 synthesizes and compares these studies in terms of their research issues, methodologies, key findings, and remaining research gaps.

Table 1. Analysis of Previous Research

Category	Ref.	Summary of Existing Issues	Approach	Results	Unsolved issue
Learning / Attendance Studies	[6]	Link between study space visits and academic/emotional outcomes not proven	Surveys on well-being, engagement, performance	More visits linked to better outcomes	Need objective data and causal research
	[7]	Limited long-term evidence in library instruction and re-enrollment	8-year panel (log data + logistic regression)	Participation increases re-enrollment likelihood	Future work on session intensity, quality
	[8]	Unclear link between library use and retention/performance	Survey + regression (purpose, frequency, grades, persistence)	Frequent purposeful users achieve more	Use actual logs (e-resources, entry records)
	[9]	Unclear impact of pre/post-pandemic format on attendance	Comparative cohort (in-person vs digital)	Attendance patterns shifted	Need long-term analysis with control variables
	[10]	Limited accuracy/interpretability in retention prediction	ML on 50,000 students (proxy features)	Data integration improves performance	Need finer attendance data
	[11]	Limited feature importance analysis	ML with participation frequency, dropout, outcomes	Participation predicts attrition	Need intervention studies

	[12]	Lack of real-time retention risk frameworks	Big data + ML using LMS activity	Improved early intervention timing	Include offline data with ethics review
	[13]	Lack of comparative ML studies	Experiments with academic + participation indicators	Performance varies by dataset	Need domain-specific variables
	[14]	Limited validation of dropout models	ML with enrollment/activity frequency	Strong detection performance	Need study-space logs + cross-institution tests
	[15]	Limited evidence on workshop attendance effects	Analysis with attendance, GPA, retention	Positive link to retention	Need randomized trials
Commercial Retention & AI Studies	[20]	Difficulty identifying customer churn patterns in service industries	Customer churn prediction using machine learning	ML models improved churn prediction accuracy	Need domain-specific behavioral features
	[21]	Limited integration of transactional customer behavior analysis	Behavioral analytics using transaction and purchase history	Behavioral patterns strongly related to customer loyalty	Need real-time behavioral adaptation analysis
	[22]	Limited comparative evaluation of retention prediction models	Comparative analysis of churn prediction algorithms	Predictive performance differed by algorithm and dataset	Predictive performance differed by algorithm and dataset
	[23]	Lack of interpretable AI frameworks for customer analytics	Explainable AI (XAI) and SHAP-based model interpretation	Improved interpretability of predictive models	Need validation in real-world service environments
Proposed Study	This Study	Limited integration of attendance records, payment data, and AI-based retention prediction in study café environments	Behavioral logs + payment data + ML/XAI	Comparative analysis of retention-related behavioral patterns	Need external and longitudinal validation

Data & Methodology

Data Collection

This study examines the relationship between behavioral usage patterns and actual repurchase behavior in study cafes as a representative attendance-based service sector. Data were collected from 25,666 users across 272 branches in 17 provinces of Korea, and after preprocessing, the final sample consisted of 22,478 users. This research design is similar in approach to Paper [10], which predicted student retention using early-semester usage records. Table 2 provides an overview of the data collection process. Because all transaction records were collected within a short and contiguous four-month period, the potential influence of inflation and long-term economic fluctuations on payment-related variables was considered minimal.

Table 2. Overview of Data Collection

Item	Description
Industry/Target	Study cafe users
Branches/Regions	272 branches across 17 provinces in Korea
Sample Size	22,478 individual users after preprocessing
Collection Period	May 2025 (independent variables) – August 2025 (dependent variables)
Data Composition	Entry/exit logs (visit days, stay duration), payment records (amount, repurchase), demographics (age, region)
Research Strength	Utilizes actual behavioral log data, overcoming limitations of survey-based studies

The dataset was obtained through a research collaboration with a commercial study café service provider and was used exclusively for academic research purposes. Prior to analysis, all personally identifiable information was removed, and the dataset was fully anonymized so that individual users could not be identified. The final analytical sample was derived by excluding observations with missing values, unrealistic age records, and extreme outliers according to the preprocessing criteria described in Table 3. Furthermore, because the dataset includes users from 272 branches across all 17 provinces of Korea and covers a broad range of customer ages and usage patterns, it provides a reasonably representative sample of study cafe users within the Korean attendance-based service market.

Data Preprocessing

To ensure data quality, we applied the preprocessing procedures summarized in Table 3, resulting in a refined dataset that more accurately reflects the real service context [16].

Table 3. Data Preprocessing Procedures

Step	Description	Purpose
Outlier Handling	Removed top and bottom 1% of total and average stay duration	Mitigation of skewness and extreme values
Age Criterion	Excluded users under 10 and over 60	Elimination of unrealistic cases
Missing Values	Removed observations with missing key variables	Ensuring data stability
User-level Matching	Matched May usage variables with June-August purchase outcomes at the individual user level	Securing unit-level consistency

Variable Definition

The variables presented in Table 4 are consistent with those identified in prior studies as key determinants of retention. They also corroborate the findings of previous research [17], which indicate that retention can be predicted using only a minimal set of behavioral and demographic variables. This study was conducted as a secondary analysis using anonymized service usage records.

Table 4. Variable Definition

Item	Variable Name	Definition	Unit	Note
Dependent Variable (Y)	retained_june	Repurchase in June	0/1	Based on payment
	retained_july	Repurchase in July	0/1	Based on payment
	retained_august	Repurchase in August	0/1	Based on payment
	retained_90	At least one repurchases between June-August	0/1	Cumulative indicator
Independent Variable (X)	visit_days	Number of Visit days in May	Days	Attendance Log [3][4][6]
	total_duration_min	Total stay duration in May	Minutes	Cumulative stay time
	avg_duration_min	Average stay duration in May	Minutes	Per-visit average
	total_payment_may	Total payment in May	KRW	Payment log
Control Variable	age	Age	Continuous	Demographic [7]
	region_city_group	Region (province group)	Categorical	Provinces in Korea [8]

Although total_duration_min and avg_duration_min are both duration-related variables, they were conceptually distinguished in this study. Specifically, total_duration_min reflects cumulative engagement over the observation period, whereas avg_duration_min represents per-visit engagement intensity. To reduce potential interpretation bias caused by overlapping behavioral characteristics, the variables were interpreted primarily in terms of their relative directional consistency and practical effect size rather than solely based on statistical significance.

Analysis Method

The research design, illustrated in Table 5, was structured to empirically test the hypotheses (H1–H6) presented in Section 1, and the analysis was conducted sequentially as defined in the table. The methodology of comparing logistic regression with AI-based models has also been employed in studies predicting hospital readmissions from medical data and analyzing student dropout risks, thereby validating both model performance and interpretability [19]. Random Forest (RF), XGBoost (XGB), and Multi-Layer Perceptron (MLP) were selected because they represent complementary machine learning paradigms, including ensemble bagging, gradient boosting, and neural network–based nonlinear learning, respectively. This combination enables a balanced comparison between traditional statistical methods and diverse AI-based approaches using the same behavioral feature set. For model implementation, Logistic Regression was performed using the LBFGS solver with a maximum of 1,000 iterations. Random Forest was configured with 150 decision trees and a maximum tree depth of 10, while XGBoost employed 200 estimators with a maximum tree depth of 6, a learning rate of 0.1, a subsample ratio of 0.8, and a column sampling ratio of 0.8. The Multilayer Perceptron (MLP) consisted of two hidden layers (64 and 32 neurons) with the ReLU activation function and the Adam optimizer, and early stopping was applied to reduce overfitting. Feature engineering was intentionally minimized because all input variables were directly derived from actual behavioral logs and transactional records. Model performance was evaluated using a stratified 70:30 training–testing split with a fixed random seed (42), and identical input variables were used across all models to ensure a fair comparison. The hyperparameter settings of the machine learning models are summarized in Table 6.

Table 5. Summary of Analysis Methods

Step	Method	Purpose	Related Hypotheses	Reference
1	Descriptive Statistics	Identify sample characteristics	Explore H5, H6	—
2	Correlation Analysis	Explore relationships between variables and repurchase	Preliminary check of H1–H4	—
3	Logistic Regression	Prediction using traditional regression	Test H1–H6	[18]
4	Machine Learning (RF, XGB, MLP)	Compare ML-based performance	—	[19]
5	XAI (Feature Importance, SHAP)	Strengthen interpretation of behavioral predictors and model outputs	—	—
6	Robustness Check	Quartile, interaction, and subgroup analyses	Supplement H1–H6	—

Table 6. Hyperparameter Settings of Machine Learning Models

Model	Parameter	Setting
Logistic Regression	Solver	LBFGS
	Max iterations	1000
Random Forest	Number of trees	150
	Maximum depth	10
XGBoost	Number of estimators	200
	Maximum depth	6
	Learning rate	0.1
	Subsample ratio	0.8
	Column sampling ratio	0.8
MLP	Hidden layers	(64, 32)
	Activation	ReLU
	Optimizer	Adam
	Early stopping	Yes
Validation	Train/Test split	70:30 (stratified)
	Random seed	42

Analysis Results

Descriptive Statistics

Table 7 summarizes the descriptive statistics of the key variables. In May, users visited an average of 4.24 days, with a total stay duration of 1,311 minutes and an average stay of 234 minutes. The mean payment amount during the same period was KRW 59,250.

Table 7. Descriptive Statistics of Key Variables

Variable	Mean	Std. Dev.	Q1	Median	Q3	Min	Max
Visit Days (days)	4.24	5.40	1	2	5	1	31
Total Duration (minutes)	1,311	2,570	150	361	1,167	1	25,849
Average Duration (minutes)	234	178	110	190	316	1	1,431

Total Payment in May (KRW)	59,250	108,229	6,500	17,000	64,500	0	1,116,000
Age (years)	22.36	10.07	15	20	26	10	60

Table 8 reports repurchase rate distributions: 24.2% in June, 27.6% in July, and 23.9% in August, with a cumulative 90-day repurchase rate of 46.3%. These figures suggest that roughly half of the users made at least one repurchase within three months.

Table 8. Distribution of Repurchase Rates

Category	Repurchase = 0 (n, %)	Repurchase = 1 (n, %)
June	17,045 (75.8%)	5,433 (24.2%)
July	16,271 (72.4%)	6,207 (27.6%)
August	17,112 (76.1%)	5,366 (23.9%)
90 days	12,079 (53.7%)	10,399 (46.3%)

Correlation Analysis

Table 9. Pearson Correlation between Independent Variables and Repurchase

Variable	June Repurchase	July Repurchase	August Repurchase	90-day Repurchase
Visit_days	0.013	0.117	0.096	0.066
Avg_duration_min	-0.032	0.032	0.018	-0.016
Total_duration_min	-0.003	0.091	0.073	0.041
Total_payment_may	-0.009	-0.051	-0.043	-0.045

The correlation analysis results presented in Table 9 can be summarized as follows.

First, visit_days consistently showed a positive correlation across all periods, with relatively stronger effects in July (0.117) and August (0.096), supporting H1 and suggesting that visit frequency is an important behavioral factor associated with repurchase behavior.

Second, avg_duration_min displayed small coefficients with inconsistent signs across months, indicating limited predictive power and providing only partial support for H3.

Third, total_duration_min generally exhibited a positive correlation but weaker explanatory power than visit days, suggesting overlapping effects and partially supporting H2.

Fourth, total_payment_may showed weak negative correlations in July (-0.051), August (-0.043), and the 90 days (-0.045). This implies that customers with higher initial spending in May were less likely to repurchase later, reflecting a possible saturation or substitution effect and supporting H4.

Logistic Regression Results

Behavioral variables from May were used as independent variables, and repurchase status in June, July, August, and over the 90-day period served as dependent variables. Logistic regression analysis was then conducted, and the model is specified as follows.

$$\Pr(Ym = 1) = \frac{1}{1+\mathrm{exp}e^{-(\alpha+\beta_1\cdot \mathrm{visis\_days}+\beta_2\cdot \mathrm{avg\_duration\_min}+\beta_3\cdot \mathrm{total\_payment\_may}+\beta_4\cdot \mathrm{age})}} \quad (1)$$

Here, Ym denotes repurchase status for each month or for the cumulative 90-day period, as defined below.

$$m \in \{june, july, August, 90days) \quad (2)$$

Table 10. Logistic Regression Results (Odds Ratio & p-value)

Variable	OR (June)	OR (July)	OR (August)	OR (90 days)	P (June)	P (July)	P (August)	P (90 days)
visit_days	1.0689	1.1180	1.1232	1.1109	0.0000	0.0000	0.0000	0.0000
avg_duration_min	1.0000	1.0006	1.0006	1.0002	0.6654	0.0000	0.0000	0.0160
total_duration_min	0.9999	0.9998	0.9998	0.9998	0.0001	0.0000	0.0000	0.0000
total_payment_may	1.0000	1.0000	1.0000	0.9908	0.0000	0.3424	0.0326	0.2123
age	0.9859	0.9958	0.9949	0.9908	0.0000	0.0054	0.0013	0.0000

Table 10 presents the logistic regression results (odds ratios and p-values) for the four models (June, July, August), and the analysis results are as follows.

First, visit_days consistently had a significant positive effect across all models. In the July model, for instance, each additional visit day increased the odds of repurchase by about

1.118 times ($p < 0.001$), and a five-day increase raised the odds by roughly 1.75 times, supporting H1.

Second, avg_duration_min was statistically significant in some models (July, August, 90 days), but its effect size was very close to 1, indicating limited substantive influence and providing only partial support for H3.

Third, total_duration_min was statistically significant across all models. Still, odds ratios remained close to 1, suggesting weak explanatory power possibly reflecting overlapping behavioral characteristics with visit frequency, and thus only partial support for H2.

Fourth, total_payment_may was generally non-significant or weakly negative, indicating that customers with higher May spending were less likely to repurchase later, consistent with a saturation or substitution effect and supporting H4.

age consistently showed negative effects across all models, implying that older customers were less likely to repurchase. This reflects that the main users of study cafes are adolescents and young adults, supporting H5.

Fig. 1. Forest plot of odds ratios with 95% confidence intervals by variable and model, illustrating that the positive effect of visit days is more pronounced than that of other variables.

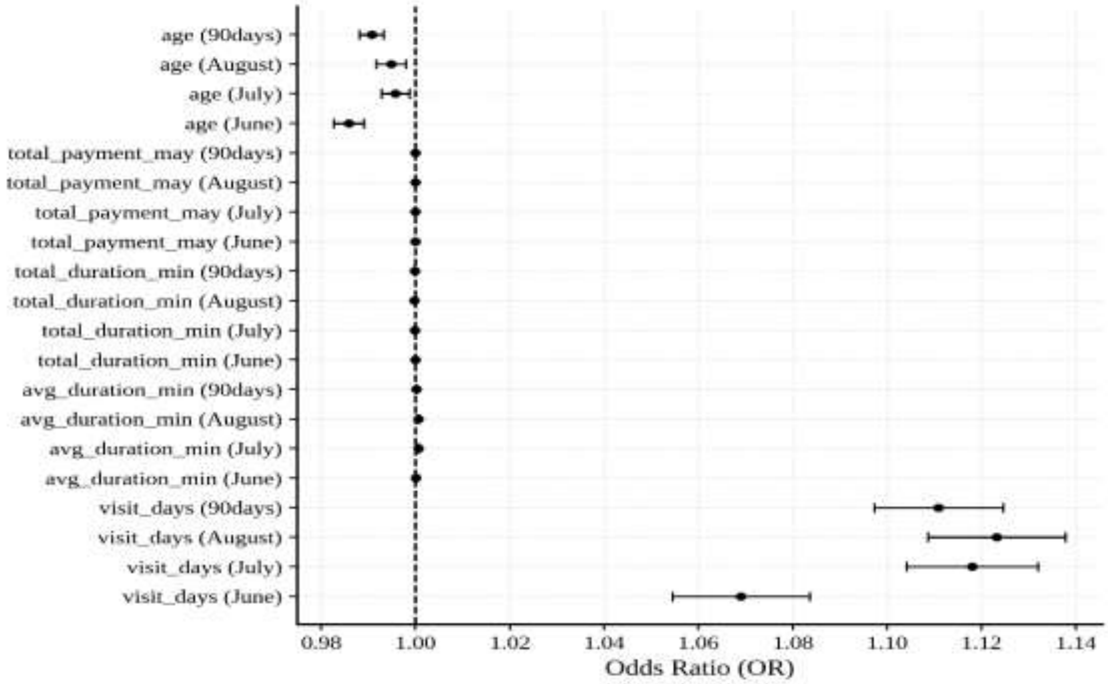

Fig. 1. Odds Ratios with 95% CI (Forest Plot). The figure summarizes the estimated effects of the behavioral variables on repurchase across multiple logistic regression models.

Fig. 2 predicted repurchase probability in the July model, showing a monotonic increase as visit days rise from 0 to 30.

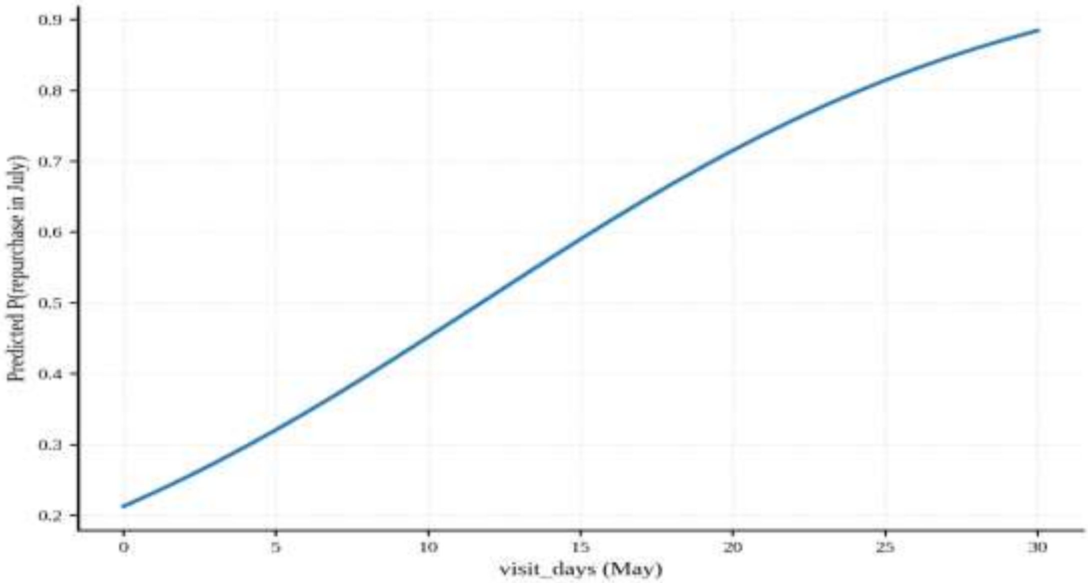

Fig. 2. Predicted Repurchase Probability by visit days (July Model). Predicted repurchase probability increases progressively as the number of visit days increases.

Quartile Analysis

Quartile analyses were conducted by dividing May visit_days and avg_duration_min into four groups (Q1–Q4) to examine the effects of customer behavioral characteristics on repurchase behavior. Results are summarized in Table [11] and [12] and Fig. [3] and [4].

Table 11. Repurchase Rates by Quartiles of visit_days

Quartile	N	June (%)	July (%)	August (%)	90-day (%)
Q1 (≤1)	9,791	21.10	19.76	16.80	38.10
Q2 (2–5)	7,566	26.37	30.50	27.13	51.41
Q3 (6–10)	2,628	30.10	38.28	33.07	57.04
Q4 (≥11)	2,493	23.31	38.43	32.05	51.34

Based on the analysis of Table 11, the average duration of stay did not exhibit a consistent pattern. A slight increase was observed in the Q2–Q3 interval, but it declined again in Q4, indicating that duration of stay is a limited variable in explaining repurchase behavior and only partially supports H3.

Table 12. Repurchase Rates by Quartiles of avg_duration_min

Quartile	N	June (%)	July (%)	August (%)	90-day (%)
Q1 (≤120)	6,944	24.77	24.16	21.53	45.16
Q2 (121–240)	7,284	25.25	28.06	24.51	47.54
Q3 (241–360)	3,965	24.29	30.79	26.51	48.40
Q4 (≥361)	4,285	21.26	29.50	24.15	43.90

As shown in Fig. 3, the bar chart compares repurchase rates in June, July, and August across the quartiles (Q1–Q4) of visit frequency. The results indicate a clear trend in which higher visit frequency is associated with higher repurchase rates.

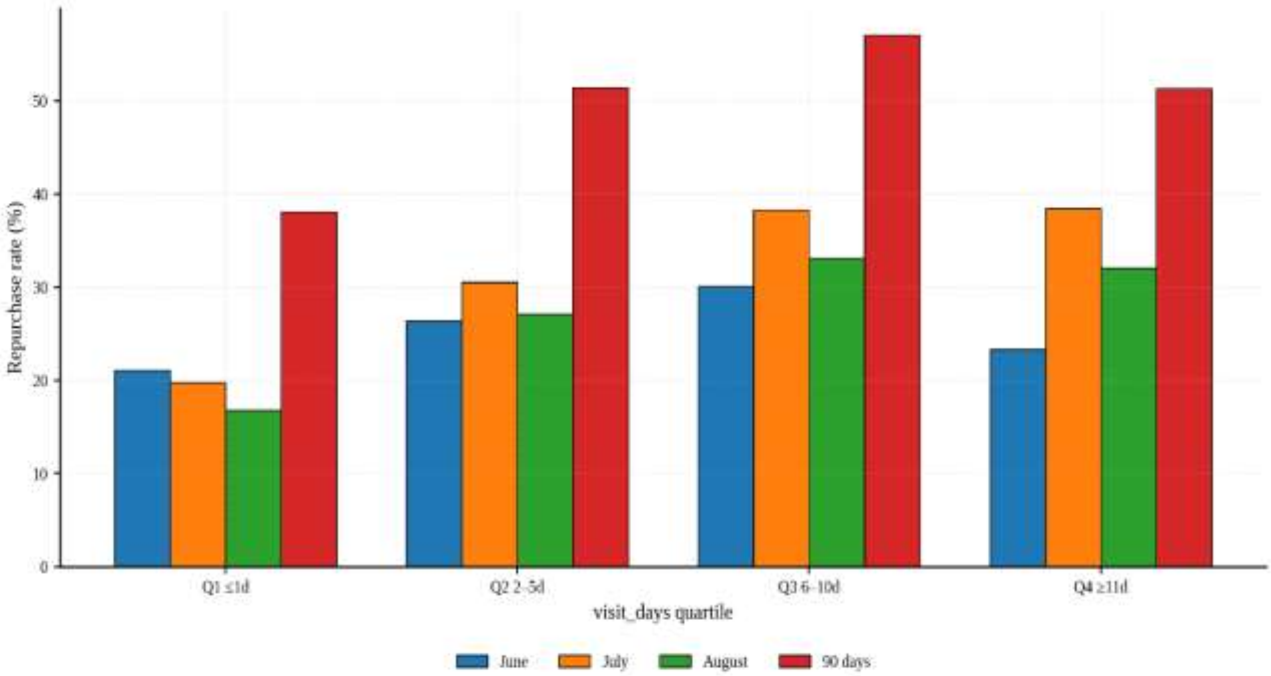

Fig. 3. Repurchase Rates by Quartiles of visit_days. Customers with higher visit frequency generally exhibited higher 90-day repurchase rates.

Fig. 4 illustrates repurchase rates by quartiles of average stay duration (avg_duration_min), showing slight increases in the middle quartiles (Q2 and Q3) but an overall inconsistent pattern across the four groups. Taken together, the results presented in Tables 11 and 12 and Figs. 3 and 4 can be summarized as follows.

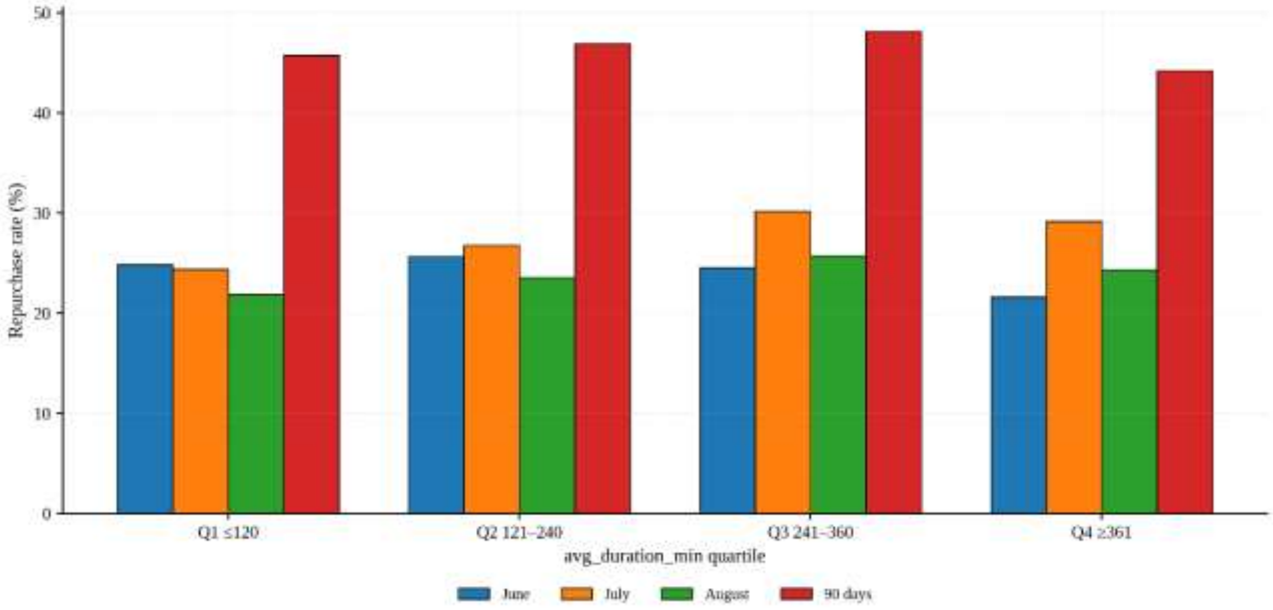

Fig. 4. Repurchase Rates by Quartiles of avg_duration_min. Average stay duration showed only a limited and inconsistent relationship with repurchase behavior.

First, repurchase rates by visit days quartiles showed a clear upward trend. The 90-day repurchase rate was only 38.1% for the Q1 group (≤ 1 day), but it rose substantially to 57.0% for the Q3 group (6–10 days). The Q4 group (≥ 11 days) showed a slight decline to 51.3%, yet overall, customers with medium to high visit frequency exhibited higher repurchase rates. This finding supports H1.

Second, repurchase rates by avg_duration_min did not exhibit a clear monotonic pattern. There was a slight increase in Q2 (121–240 minutes) and Q3 (241–360 minutes), but a decline was observed in Q4 (≥ 361 minutes). This indicates that the absolute length of stay has limited value in directly predicting repurchase behavior and provides only partial support for H3.

Third, these results suggest that customer retention strategies encouraging “frequent but shorter visits” (e.g., stamp programs or weekly mission events) are more effective in enhancing long-term repurchase rates. In contrast, strategies focused solely on extending stay duration are unlikely to directly improve repurchase rates. These additional analyses further demonstrate that the observed relationships remain consistent across different behavioral groups, thereby supporting the robustness of the proposed findings.

Baseline Predictive Performance

The logistic regression model, defined with 90-day repurchase status as the dependent variable and May’s behavioral variables as explanatory variables, is specified as follows.

$$\Pr(Y_{90} = 1) = \frac{1}{1+\exp e^{-(\alpha+\beta_1 \cdot \text{visis\_days}+\beta_2 \cdot \text{avg\_duration\_min}+\beta_3 \cdot \text{total\_payment\_may}+\beta_4 \cdot \text{age})}} \quad (3)$$

A fixed train-validation split (70%/30%) was adopted to preserve temporal consistency between historical behavioral observations and subsequent repurchase outcomes. Since the present study focuses on time-dependent service usage behavior, maintaining chronological separation was considered more appropriate than random resampling-based validation approaches. This split also provided a sufficient number of observations for model training while preserving an independent dataset for unbiased performance evaluation. Although K-fold cross-validation was considered, a fixed train-validation split was adopted because preserving the temporal relationship between historical behavioral observations and subsequent repurchase outcomes was considered more appropriate for the objectives of this study. The baseline logistic regression model, evaluated with a cut-off of 0.5 and using a 70% training and 30% validation split, yielded the performance results shown in Table 13.

Table 13. Logistic Regression Baseline Performance (cut-off = 0.5)

Metric	Accuracy	Precision	Recall	F1-score	ROC-AUC
Value	0.531	0.394	0.024	0.046	0.593

Only 76 out of 3,120 actual repurchasers were correctly classified, illustrating the poor recall performance. In other words, the model generated a large number of false negatives, missing most repurchasers and confirming that the baseline logistic regression showed limited practical predictive utility.

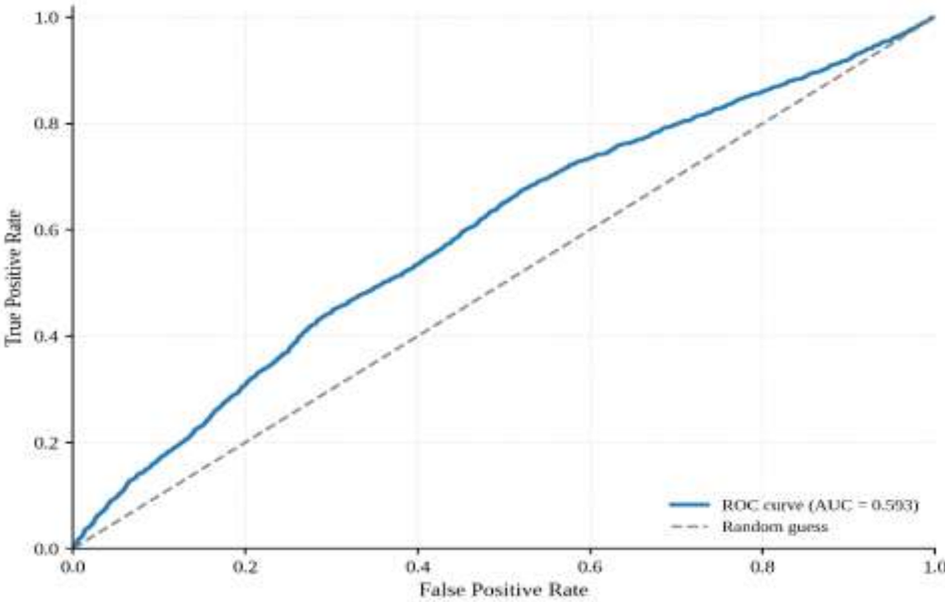

Fig. 5. ROC Curve (Baseline Logistic Regression). The ROC curve illustrates the limited predictive performance of the baseline logistic regression model.

Fig. 5 presents the ROC curve of the baseline logistic regression model, with an AUC of 0.593, confirming the limitations of the baseline model. To enhance its performance, a cut-off optimization experiment was conducted using the precision–recall curve to identify the threshold that maximizes the F1-score. The analysis indicated an optimal cut-off of approximately 0.288, at which the performance metrics were Precision = 0.463, Recall = 1.000, and F1-score = 0.633.

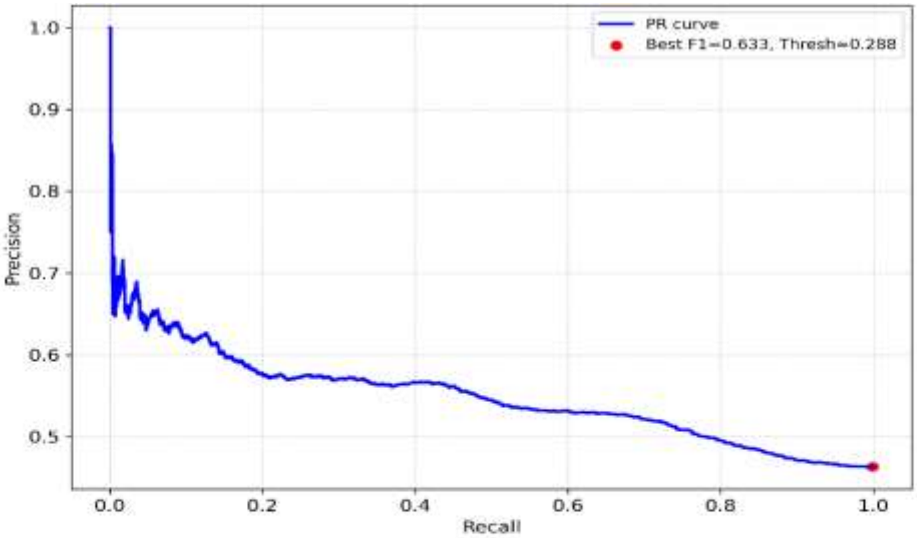

Fig. 6. Precision–Recall Curve (Baseline Logistic Regression). The precision–recall curve illustrates the trade-off between precision and recall, highlighting the limited predictive capability of the baseline logistic regression model.

Fig. 6 presents the precision–recall curve of the logistic regression model. At the optimal cut-off (0.288), false negatives dropped to zero, resulting in a recall of 1.0. However, this process also caused false positives to increase sharply, reducing precision to the baseline level (46.3%). The overall findings can be summarized as follows.

First, the baseline logistic regression model exhibited low predictive performance for repurchase due to its simple linear structure and limited set of variables.

Second, although adjusting the cut-off substantially improved recall, it caused a loss of precision, thereby reducing the model’s practical applicability.

Third, this demonstrates that customer behavioral data exhibit nonlinear and interactive characteristics that are difficult to capture with simple regression.

To address these limitations, Section 4.6 applies AI-based models—Random Forest, XGBoost, and Neural Networks—to compare their predictive performance with logistic regression while evaluating the trade-off between precision and recall across models.

Performance Comparison of AI/Machine Learning Models

To address the limited predictive performance of the baseline logistic regression model, this study employed representative machine learning and deep learning models. Specifically, Random Forest (RF), XGBoost (XGB), and a Multilayer Perceptron (MLP) neural network were trained and validated using the same data split (70% for training and 30% for validation) and the same set of independent variables. Table 14 reports the comparative performance of logistic regression, Random Forest, XGBoost, and the neural network model. Fig. 7 compares the ROC curves of Logistic, RF, XGB, and NN models. Logistic regression had the lowest performance (AUC = 0.596), whereas RF (0.671), NN (0.663), and XGB (0.676) all performed better. XGBoost showed the best classification performance with an ROC-AUC of 0.676.

Table 14. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
Logistic (base)	0.575	0.575	0.315	0.407	0.596	0.553
Random Forest	0.633	0.626	0.511	0.563	0.671	0.627
XGBoost	0.640	0.633	0.529	0.577	0.676	0.642
Neural Network	0.625	0.630	0.458	0.530	0.663	0.626

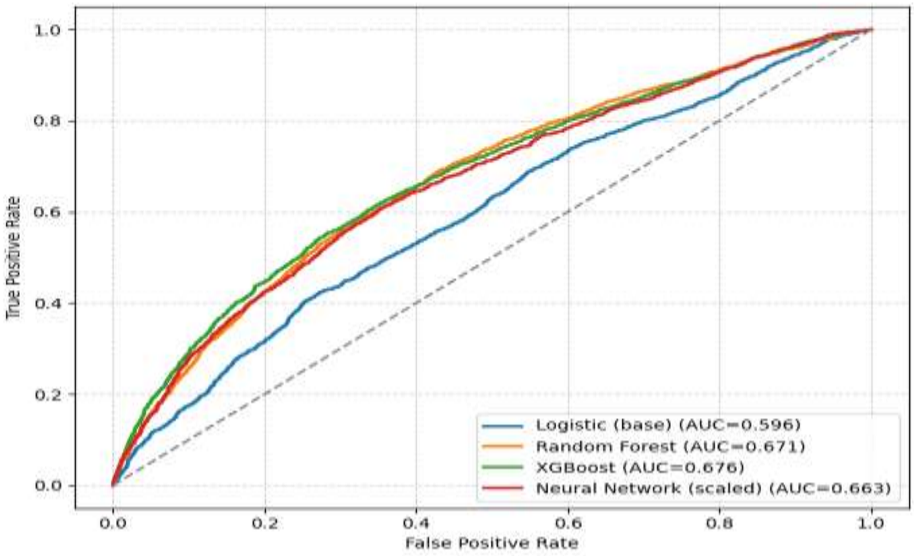

Fig. 7. ROC Curve Comparison (Logistic, RF, XGB, NN). XGBoost achieved the highest ROC-AUC, demonstrating the strongest overall predictive performance among the evaluated models.

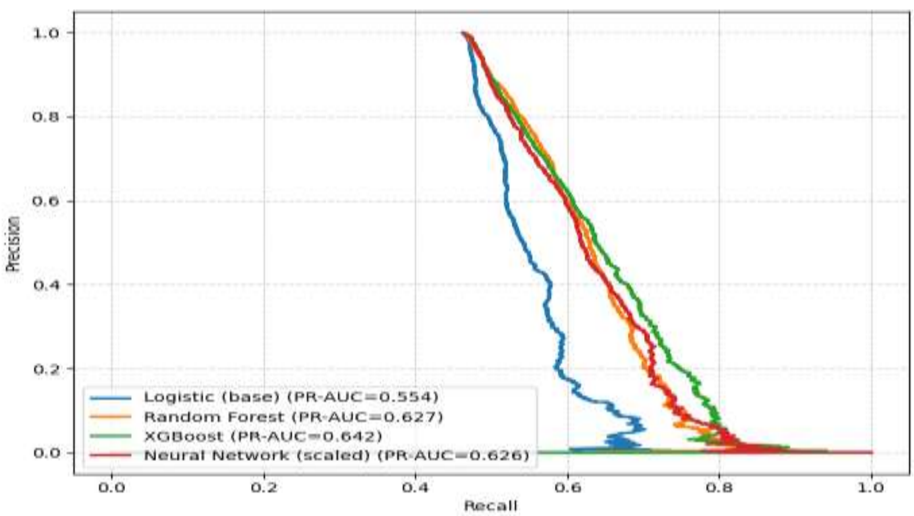

Fig. 8.Precision–Recall Curve Comparison (Logistic, RF, XGB, NN). XGBoost achieved the largest area under the precision–recall curve, indicating superior performance in identifying repurchasing customers.

Fig. 8 presents the precision–recall curves for the same models. Logistic regression again performed worst (PR-AUC = 0.553), while RF (0.627), NN (0.626), and XGB (0.642) all outperformed it. XGBoost had the best performance, with the largest area under the curve (PR-AUC = 0.642). Key findings are as follows.

First, logistic regression remained at baseline performaAnce with an accuracy of 0.575, ROC-AUC of 0.596, and PR-AUC of 0.553.

Second, Random Forest showed improved performance, achieving a recall of 0.511, an F1-score of 0.563, and an ROC-AUC of 0.671.

Third, XGBoost demonstrated the best overall performance, with an accuracy of 0.640, recall of 0.529, ROC-AUC of 0.676, and PR-AUC of 0.642.

Fourth, the scaled neural network achieved an accuracy of 0.625, ROC-AUC of 0.633, and PR-AUC of 0.626, showing a performance level similar to that of Random Forest. Unlike the original neural network, which showed unstable baseline performance, the application of feature standardization yielded more stable predictive results.

Overall, machine learning models demonstrated comparatively stronger predictive performance than logistic regression, with XGBoost showing the strongest overall performance among the evaluated models. Moreover, the neural network also achieved meaningful performance when appropriate preprocessing was applied. The superior performance of XGBoost can be attributed to its ability to capture complex nonlinear relationships and interactions among behavioral variables while effectively handling heterogeneous feature distributions. In contrast, Logistic Regression assumes linear relationships between predictors and outcomes, which may limit its ability to model complex customer behavioral patterns. Although Random Forest also demonstrated competitive performance, XGBoost achieved better generalization by sequentially correcting prediction errors through gradient boosting. These findings suggest that gradient boosting models are particularly well suited for predicting customer retention using behavioral log and transactional data collected in attendance-based service environments.

Feature Importance and Interpretation (SHAP Analysis)

To better interpret the prediction results, feature importance was compared between Random Forest and XGBoost, and SHAP analysis was further conducted for the XGBoost model.

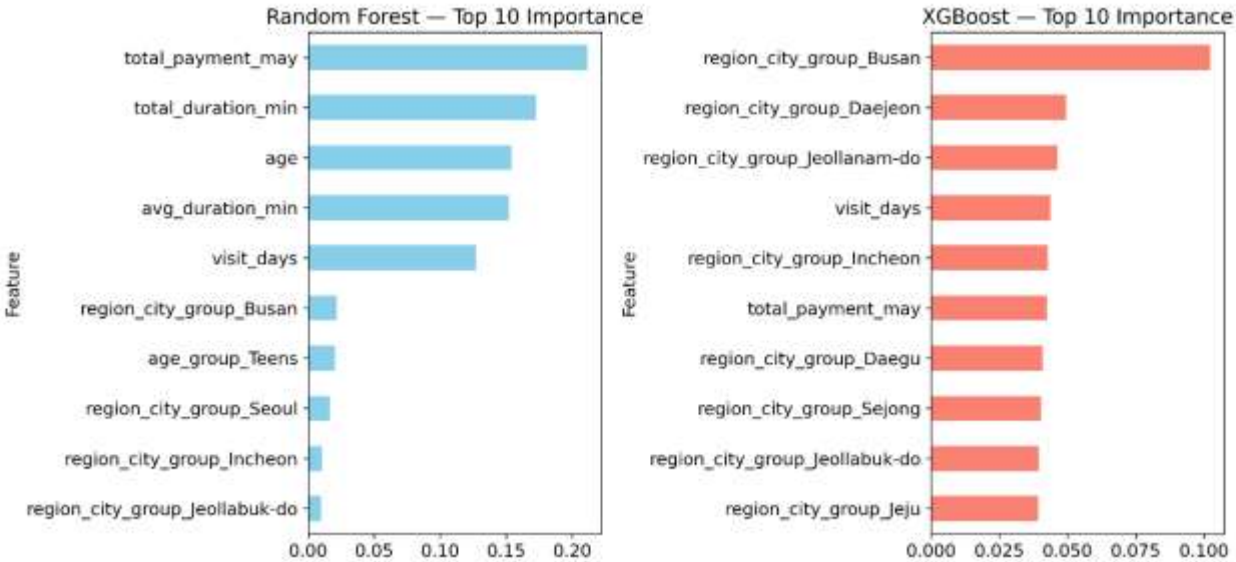

Fig. 9. Feature Importance (RF vs. XGB). The figure compares the relative importance of behavioral variables estimated by Random Forest and XGBoost.

Fig. 9 presents the feature importance results (RF vs. XGB). Random Forest ranked continuous variables such as total_payment_may, total_duration_min, and age at the top, whereas XGBoost emphasized regional variables such as region_city_group. This reflects the structural differences between the two algorithms. These results support H6, while the relatively low importance of stay-duration variables provides only partial support for H2 and H3.

Fig. 10 presents the SHAP summary plot for XGBoost, showing that the directions and relative contributions of variables can be interpreted in the SHAP-based analysis. The results are summarized as follows.

First, as visit_days increased, the SHAP values shifted in a positive direction, being associated with higher repurchase likelihood. This finding is consistent with H1.

Second, higher values of total_payment_may show negative contributions, indicating that customers with higher spending were less likely to repurchase. This finding is consistent with H4 and reflects a saturation or substitution effect.

Third, age contributed negatively, showing that older customers were less likely to repurchase. This finding is consistent with H5.

Fourth, avg_duration_min and total_duration_min contributed to some extent, but their directions and magnitudes were unstable, indicating only limited effects. This provides partial support for H2 and H3.

Fifth, regional variables such as region_city_group were highlighted prominently in specific models, suggesting potential regional differences in repurchase behavior. This finding is consistent with H6.

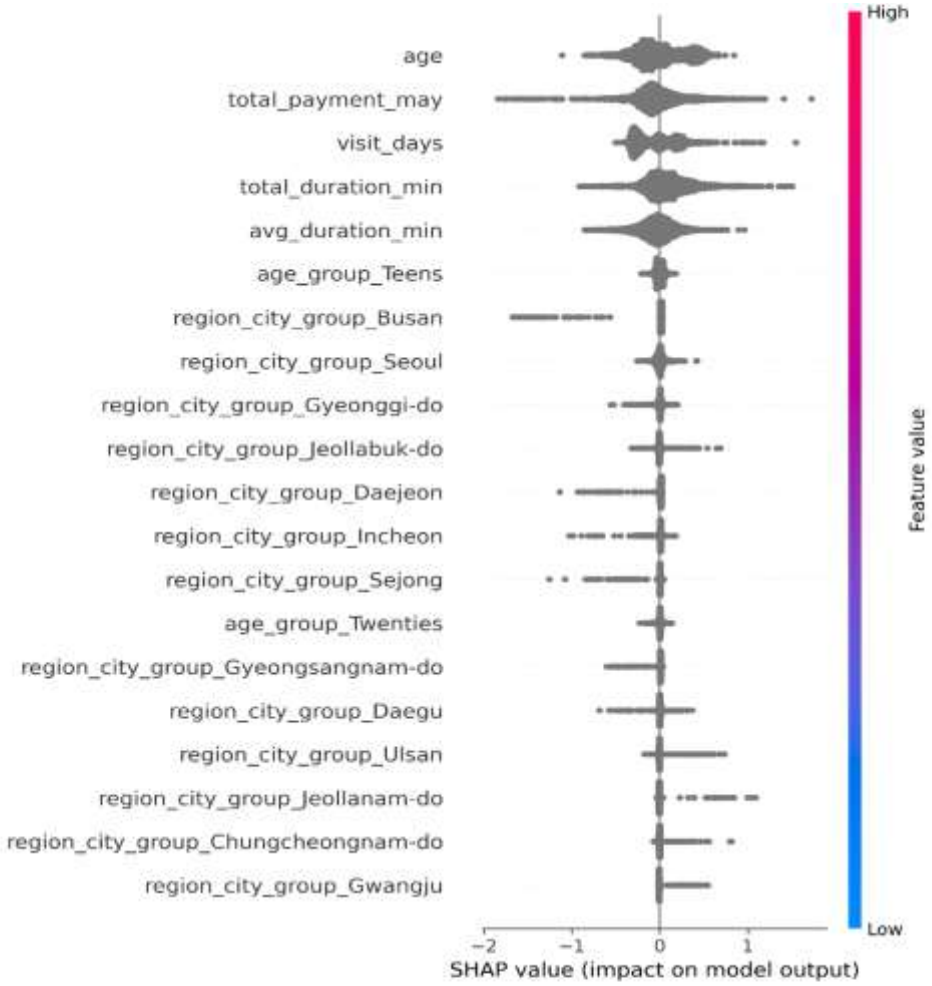

Fig. 10. SHAP Summary Plot (XGBoost). The SHAP summary plot illustrates the direction and relative contribution of each behavioral variable to the XGBoost prediction model.

Taken together, the feature importance and SHAP analyses suggested that the number of visit_days emerged as one of the most influential variables in predicting repurchase, whereas total payment in May and age had negative effects. Although regional variables varied slightly across models, they generally supported regional differences in repurchase (H6). Moreover, the variation in variable importance across models and the complementary insights provided by SHAP analysis suggest that results should be interpreted in terms of model dependence and consistent directional effects, rather than relying solely on the absolute ranking of variable importance. From a managerial perspective, these findings provide actionable insights for customer retention management. The consistently positive contribution of visit frequency suggests that attendance-based incentives, such as stamp rewards or weekly mission programs, may effectively encourage repeat visits. Conversely, the negative contribution of total payment indicates that high-spending customers may require differentiated retention strategies, such as personalized promotions or flexible membership benefits, rather than uniform marketing campaigns. In addition, the observed effects of age and regional variables support customer segmentation strategies tailored to demographic and regional characteristics, enabling more targeted and effective retention management.

Interim Summary

This study empirically analyzed repurchase outcomes based on service usage records and payment data from study cafe users. The results of the hypothesis testing related to behavioral predictors are summarized in Table 15.

Table 15. Summary of Hypothesis Testing Results

Hypothesis	Statement	Result
H1	More visit days in May increase the likelihood of repurchase (June–August, 90 days).	Supported
H2	Total stay duration in May has a positive but limited effect on repurchase.	Partially supported
H3	Average stay duration in May positively affects repurchase.	Partially supported
H4	Higher total payment in May decreases the likelihood of subsequent repurchase.	Supported
H5	Older customers show lower repurchase rates.	Supported
H6	Repurchase rates differ by region.	Supported

DISCUSSION & CONCLUSION

Academic Implications

This study addressed the limitations of prior survey-based research by leveraging service usage records and payment data. The findings are generally consistent with previous attendance-based studies reporting positive relationships between visit frequency and retention-related outcomes. However, unlike most previous studies that primarily focused on educational achievement or academic retention, this study extends the analysis to a commercial study café environment by integrating behavioral log data, transactional payment records, and explainable AI techniques to predict actual repurchase behavior. The findings also extend the literature on customer retention by demonstrating that objective behavioral log data can complement traditional survey-based approaches for explaining and predicting repurchase behavior in attendance-based commercial service environments. Visit frequency was consistently validated across multiple analytical approaches—correlation analysis, logistic regression, quartile analysis, and explainable AI (XAI)—as one of the primary behavioral factors associated with repurchase behavior, thereby supporting the established relationship between attendance frequency and retention. Furthermore, by comparing Random Forest, XGBoost, and Neural Networks with logistic regression, the study demonstrated that AI-based models can provide comparatively improved predictive performance under behavioral log-based service environments, with XGBoost outperforming all others across evaluation metrics and underscoring the utility of AI models in customer retention analysis. Finally, through the application of XAI techniques (SHAP and feature importance), the study elucidated the contribution of each variable by model, offering managerial insights beyond mere performance comparisons. Collectively, these contributions enhance the transparency, interpretability, and credibility of customer behavior research.

Managerial Implications

First, customer retention strategies should prioritize promoting frequent visits, even if they are brief. Initiatives such as stamp rewards, weekly missions, and attendance-based coupons can effectively enhance repurchase rates. Second, strategies focusing solely on extending stay duration provide limited value; therefore, management should emphasize visit frequency rather than seating arrangements or facility upgrades. Third, high-spending customers were found to be less likely to repurchase later, underscoring the need for tailored management programs such as alternative or rollover benefits. Fourth, observed differences across age groups and regions highlight the necessity of segmented marketing. Strengthening incentives for adolescents and young adults and adopting region-specific differentiation strategies may be particularly effective. Finally, implementing XGBoost-based machine learning models can facilitate the early identification of at-risk customers and support targeted marketing efforts.

Limitations and Future Research Directions

This study is limited to a single industry (study cafes) and a short observation window (May–August), which may restrict the generalizability of the findings. Moreover, the analysis relied primarily on service usage records, without incorporating psychological or socio-economic factors. In addition, because the dataset was collected from a single commercial study café platform, the observed behavioral patterns may reflect platform-specific customer characteristics and operational policies, which may limit the broader applicability of the findings. The AI-based models were confined to RF, XGB, and NN, without adequate comparison with other deep learning or time-series–based approaches. Future research should therefore: (1) improve generalizability by applying the framework to diverse service industries and long-term panel datasets, (2) develop hybrid models combining attitudinal factors with service usage records, and (3) enhance predictive performance through advanced deep learning and AutoML techniques. In addition, the present study focused on monthly behavioral aggregates and did not explicitly model within-month churn or short-term engagement fluctuations during May. Future research may utilize finer-grained temporal sequences to capture intra-month behavioral dynamics, investigate interaction effects among behavioral and demographic variables, and further improve model robustness and generalizability through additional validation strategies such as rolling-window validation or K-Fold Cross Validation.

CONCLUSION

This study empirically analyzed customer repurchase outcomes using service usage records and payment data from study cafe users, comparing the predictive performance of traditional regression and machine learning models. The results consistently identified visit frequency as one of the most influential behavioral factors associated with repurchase likelihood, whereas stay-duration variables had only limited effects. This finding suggests that encouraging frequent customer visits may be more effective for retention than promoting longer but less frequent stays, providing actionable managerial implications for study café operators and other attendance-based service businesses. Payment amount and age were negatively associated with repurchase, while regional variables revealed meaningful differences across groups. AI-based models demonstrated comparatively stronger predictive performance than traditional regression, with XGBoost achieving the best performance and demonstrating comparatively strong predictive capability for customer retention prediction. Moreover, the application of explainable AI (XAI) techniques enabled visualization of both the direction and magnitude of variable contributions, thereby enhancing model interpretability and practical value. Overall, this study provides empirical evidence based on behavioral logs to contribute to the growing literature on customer retention and behavioral analytics. The findings also offer practical guidance for developing data-driven customer retention strategies and support the design of AI-driven decision support systems for customer relationship management in attendance-based commercial service environments.

Funding Information

This research received no external funding.

Author's Contributions

Joonghyun Park: Conceptualization, methodology, formal analysis, writing—original draft preparation. Seungchan Lee: Dataset preparation and validation. Hoon Ko: Supervision, project administration, and writing—review and editing.

Ethics

This study involved a secondary analysis of fully anonymized service us- age and payment records originally collected for operational purposes. No personally identifiable information was accessed by the researchers, and no intervention or interaction with human participants was conducted. In accordance with institutional and national guidelines, this study was exempt from Institutional Review Board (IRB) review.

REFERENCES

1. Oliver, R.L., 1999. Whence consumer loyalty? Journal of Marketing, 63(4): 33-44. DOI: 10.1177/00222429990634S105.
2. Zeithaml, V.A., Berry, L.L. and Parasuraman, A., 1996. The behavioral consequences of service quality. Journal of Marketing, 60(2): 31-46. DOI: 10.1177/002224299606000203.
3. Yi, S., Lee, Y.W., Connerton, T. and Park, C.-Y., 2021. Should I stay or should I go? Visit frequency as fitness centre retention strategy. Managing Sport and Leisure, 26(4): 268-286. DOI: 10.1080/23750472.2020.1763829.
4. Wright, L.B., 2021. Assessing library instruction: A study of the relationship between attendance, retention, and student success. The Journal of Academic Librarianship, 47(5): 102431. DOI: 10.1016/j.acalib.2021.102431.
5. Imani, M., Beikmohammadi, A. and Arabnia, H.R., 2025. Comprehensive analysis of random forest and XGBoost performance with SMOTE, ADASYN, and GNUS under varying imbalance levels. Technologies, 13(3): 88. DOI: 10.3390/technologies13030088.
6. Scoulas, J.M., De Groote, S.L., Shotick, K. and Osorio, N.L., 2024. A holistic approach to understanding undergraduates: Campus engagement, library use and psychological factors. The Journal of Academic Librarianship, 50(5): 102936. DOI: 10.1016/j.acalib.2024.102936.
7. O'Kelly, M.K., Jeffryes, J., Hobscheid, M. and Passarelli, R., 2023. Correlation between library instruction and student retention: Methods and implications. College & Research Libraries, 84(1): 85. DOI: 10.5860/crl.84.1.85.
8. Scoulas, J.M. and De Groote, S.L., 2022. Impact of undergraduate students' library use on their learning beyond GPA: Mixed-methods approach. College & Research Libraries, 83(3): 452. DOI: 10.5860/crl.83.3.452.
9. Haugom, E., 2022. The effect of changing from campus-based to digital teaching on student attendance: A case study of Norwegian business students. Heliyon, 8(11): e11307. DOI: 10.1016/j.heliyon.2022.e11307.
10. Matz, S.C., Bukow, C.S., Peters, H., Deacons, C., Dinu, A. and Stachl, C., 2023. Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. Scientific Reports, 13(1): 5705. DOI: 10.1038/s41598-023-32484-w.
11. Gonzalez-Nucamendi, A., Noguez, J., Neri, L., Robledo-Rella, V. and García-Castelán, R.M.G., 2023. Predictive analytics study to determine undergraduate students at risk of dropout. Frontiers in Education, 8: 1244686. DOI: 10.3389/feduc.2023.1244686.
12. Kok, C.L., Ho, C.K., Chen, L., Koh, Y.Y. and Tian, B., 2024. A novel predictive modeling for student attrition utilizing machine learning and sustainable big data analytics. Applied Sciences, 14(21): 9633. DOI: 10.3390/app14219633.
13. Delena, R.D., Dia, N.J., Sacayan, R.R., Sieras, J.C., Khalid, S.A., Macatotong, A.H.T. and Gulam, S.B., 2025. Predicting student retention: A comparative study of machine learning approach utilizing sociodemographic and academic factors. Systems and

Soft Computing, 200352.
DOI: 10.1016/j.sasc.2025.200352.

14. Rabelo, A.M. and Zárate, L.E., 2025. A model for predicting dropout of higher education students. Data Science and Management, 8(1): 72-85. DOI: 10.1016/j.dsm.2024.07.001.
15. Witherspoon, R.L. and Taber, P.O.L., 2021. Increasing student attendance at library workshops: What the data tells us. College & Research Libraries, 82(1): 113. DOI: 10.5860/crl.82.1.113.
16. Boozary, P., Sheykhan, S., GhorbanTanhaei, H. and Magazzino, C., 2025. Enhancing customer retention with machine learning: A comparative analysis of ensemble models for accurate churn prediction. International Journal of Information Management Data Insights, 5(1): 100331. DOI: 10.1016/j.jjimei.2025.100331.
17. Hoca, S. and Dimililer, N., 2025. A machine learning framework for student retention policy development: A case study. Applied Sciences, 15(6): 2989. DOI: 10.3390/app15062989.
18. Iwagami, M. et al., 2024. Comparison of machine-learning and logistic regression models for prediction of 30-day unplanned readmission in electronic health records: A development and validation study. PLOS Digital Health, 3(8): e0000578. DOI: 10.1371/journal.pdig.0000578.
19. Mustofa, S., Emon, Y.R., Mamun, S.B., Akhy, S.A. and Ahad, M.T., 2025. A novel AI-driven model for student dropout risk analysis with explainable AI insights. Computers and Education: Artificial Intelligence, 8: 100352. DOI: 10.1016/j.caeai.2024.100352.
20. Hadden, J., Tiwari, A., Roy, R. and Ruta, D., 2007. Computer assisted customer churn management: State-of-the-art and future trends. Computers & Operations Research, 34(10): 2902-2917. DOI: 10.1016/j.cor.2005.11.007.
21. Buckinx, W. and Van den Poel, D., 2005. Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. European Journal of Operational Research, 164(1): 252-268. DOI: 10.1016/j.ejor.2003.12.010.
22. Neslin, S.A., Gupta, S., Kamakura, W., Lu, J. and Mason, C.H., 2006. Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research, 43(2): 204-211. DOI: 10.1509/jmkr.43.2.204.
23. Lundberg, S.M. and Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NeurIPS 30), Curran Associates, Inc., pp. 4765-4774. DOI: 10.48550/arXiv.1705.07874.