

Comprehension of Health Apps' Retention Rates Employing the Means of Machine Learning Models

Ruisi LIU

2655804216@qq.com

University of Nottingham

Yining Xu

Taizhou University

Youjia Gu

Hohai University

Research Article

Keywords: Health apps, User retention, Machine learning, XGBoost, Predictive modelling, User satisfaction

Posted Date: October 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10932103/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Comprehension of Health Apps' Retention Rates Employing the Means of Machine Learning Models

Ruisi LIU^{1*}, Yining Xu² and Youjia Gu³

^{1*}Faculty of Human Social Science, University of Nottingham , Ningbo,
315100, China.

²School of Economics and Management, Taizhou University, Taizhou,
225300, China.

³Business School of Hohai University, Hohai University, Nanjing,
211100, China.

*Corresponding author(s). E-mail(s): 2655804216@qq.com ;
Contributing authors: 978039987@qq.com ; youjiagu713@gmail.com ;

Abstract

Background:Health apps for multiple purposes have been widely adopted as the global health burden has increased. With an increase of their popularity, the economic significance of the health app market has become substantial. However, health apps continue to face the challenge of low user retention rates, which hinders their sustainable development and requires urgent attention. Therefore, this study develops solutions to this situation, identifying the determinants of retention and building a machine learning model to help predict the retention rate of health apps.

Methods:In this study, three machine learning models were constructed based on linear regression, gradient boosting and XGBoost. Data preprocessing included checking missing values, encoding categorical variables, feature scaling, and feature engineering. Model performance was evaluated using R^2 , MAE, MSE, RMSE, and Spearman's rank correlation coefficient. Feature importance was analysed using Pareto diagrams, partial dependence plots, and ablation studies.

Results:The results indicate that linear regression fails to capture nonlinear user psychological interactions and yields weak fitting results; XGBoost outperforms the other models via second-order optimization and regularization, with superior rank preservation critical for ordinal 1–5 rating data. It achieves the highest R^2 and Spearman's ρ , confirming that usability and medical communication are core retention drivers.

Conclusion: This study provides possible prediction models for the retention rates of health apps, which developers can use as diagnostic tools to evaluate and optimize their app design for long-term user engagement.

Trial Registration: *Not applicable.*

Keywords: Health apps , User retention , Machine learning , XGBoost, Predictive modelling, User satisfaction

1 Introduction

With rising global healthcare burdens such as chronic disease [1, 2] and mental health [3, 4] crises, digital health interventions have become a promising solutions to help manage health conditions more effectively [5]. Despite the growing adoption of mHealth apps for chronic disease management, user retention remains low [6]. More than two-thirds of users abandon mental health apps after their initial use, with only a median retention rate of 3.3% at 30 days [7]. This alarming attrition rate underscores the need to identify the factors that drive sustained engagement, which is essential to achieve the desired health outcomes. The retention rate predicts the long-term market viability and development of health apps, which requires research to identify its drivers and formulate improvement strategies.

However, existing studies provide limited insight into how specific factors interact to shape user continuance. Previous studies have identified several key retention drivers, for instance, self-monitoring features significantly improve retention. Apps with frequent self-monitoring capabilities were estimated to have 80% retention at 40 weeks, compared with 60% of the control group [6]. Studies have also reported that app quality metrics such as MARS scores are positively related to popularity but negatively correlated with user retention, which indicates that high-quality apps attract downloads but fail to ensure long-term use [8]. This indicates a disconnect between initial appeal and ongoing value.

How these factors can be jointly interpreted to inform design decisions remains unclear. Most of the previous studies rely on observational designs, such as cross-sectional surveys or cohort analyses, but lack interventional experiments to test and optimize retention. There are few models for predicting retention rates, and these models fail to enhance the applicability of the research findings. There is also a limited understanding of how determinants interact. While determinants such as perceived usefulness and health consciousness are often examined in isolation, their potential interactions, such as synergistic or moderating effects, remain overlooked.

To address this challenge, predictive modelling offers a promising approach. In particular, machine learning has proven effective in uncovering complex patterns from user data. The ability of machine learning in health-related prediction tasks has been validated. For example, a previous study demonstrated the application of XGBoost and random forest on a substantial dataset of 70,000 patient records to predict cardiovascular disease risk, achieving a predictive accuracy of 70% [9]. This research not only highlights the ability of machine learning approaches to handle high-dimensional

medical data but also provides a valuable methodological framework for predicting the retention rates of health apps. As a result, this study adopts a similar paradigm, applying and comparing three machine learning models: Linear Regression is employed as an interpretable baseline, while Gradient Boosting and XGBoost are chosen for their superior ability to capture complex and nonlinear relationships between app features and user retention.

This study integrates two theoretical models from information systems research: the Unified Theory of Acceptance and Use of Technology (UTAUT) and the Expectation-Confirmation Model of IS Continuance (ECM-IS). This combination provides a comprehensive perspective for analysing the entire user journey from initial adoption to continuous use, which is important for understanding the user retention rate of health apps.

UTAUT suggests that behavioral intention determines the actual use of technology. Performance expectancy, effort expectancy, social influence, and facilitating conditions are the four key contributing factors to the likelihood of adopting the technology, and their effects are affected by age, gender, experience, and voluntariness of use. These key factors account for 70% of the variance in use intention, offering stronger predictive power than other models that examine technology acceptance do[10, 11]. Therefore, UTAUT is especially effective in explaining the initial user acceptance or adoption phase of new technology.

The ECM-IS was developed on the basis of the expectation-confirmation theory (ECT). ECT was originally developed to explain the process by which users form their satisfaction with products or services and eventually have the intention to repurchase [12]. Bhattacharjee emphasized the distinction between initial user acceptance and continuance behaviors. ECM-IS indicates that continued usage intention is predicated upon users' satisfaction with information systems (IS) and the perceived usefulness of continued IS use [13]. Thus, the ECM-IS is important for explaining the retention rate, as it was specifically designed to explain users' continuous usage behavior.

The adoption of this integrated framework is significant for this study, because it provides a comprehensive theoretical explanation for the entire process of investigation, from users' initial adoption to long-term retention of health apps. UTAUT and ECM-IS theoretically justify some specific features of the machine learning models adopted in this study, interpreting machine learning results and demonstrating the justification of the selection of predictor variables. For instance, the 'Expert' variable (communications with medical workers) corresponds to performance expectancy in UTAUT, 'Subscription Fees' relates to perceived cost-benefit in ECM-IS, and 'Usage Level' (individual average level of using apps) can be an indicator of confirmation or satisfaction in ECM-IS. This framework bridges theory and practice by translating model insights into actionable business strategies (e.g., boosting performance expectancy via expert content enrichment or reducing Effort Expectancy via UI simplification).

Building on this methodological and theoretical foundation, this study investigates the determinants influencing the retention rates of health apps. To this end, it develops a predictive model that serves as a tool for developers to evaluate and optimize their apps. Three machine learning models were developed, and XGBoost achieved the

highest predictive accuracy, demonstrating a strong link between user behavior, app features and the retention rate. All the models (especially XGBoost and gradient boosting) confirmed the influence of key features (user behavior, subscription fees, incentives, communication, and usage level) on app retention, providing insights for optimizing app design and commercial strategy. For app developers and designers, the predictive model can reduce development costs by prioritizing resources for features with the highest ROI. In addition, this model benefits users by improving their user experience and providing a reliable and sustainable health app to track and improve their health condition.

The remainder of this paper is organized as follows: Section two reviews related work on health app retention, including its current situation and influencing factors. Section three provides an introduction to the methodology of this study, which lists the three machine learning models employed, including linear regression, gradient boosting and XGBoost, and evaluates their performance. Section four provides the data source and describes the data used in this study, including data processing and visualization. Section five describes the model-building process and results. This section illustrates the XGBoost tuning and feature interaction capturing process. The performance is compared across three models, with the latter two minimizing loss and maximizing accuracy. The main findings are discussed in the results section by comparing the performance of the three machine learning models on the basis of four metrics: R^2 , RMSE, MSE and MAE. Finally, section six presents the discussion and conclusions of this study. This section reaffirms the main findings and translates these insights into concrete business strategies, such as modifying app attributes, maintaining incentive policies and paying attention to users' OS preferences, while it highlights data limitations and future research needs.

2 Literature review

Health apps with various functions have been developed and widely applied, but there is still a gap between app popularity and sustained retention [7]. A substantial body of literature has investigated health app engagement and retention. These studies analyse the issue primarily from three perspectives: app function design, user experience, and user psychology and behavioral patterns. The following synthesis of the key findings from each of these perspectives provides a comprehensive backdrop for the current study.

Upgrading some key functions of apps can increase the retention rate. Research has shown that the regular use of self-monitoring features is key to long-term mHealth app engagement [6]. By regularly recording health data such as steps and sleep quality, users develop a reliance on these monitoring functions. This reliance, in turn, reduces their likelihood of abandoning the app. Another study revealed that self-monitoring was the most widely employed behavior-changing technique [14], demonstrating the significant effectiveness of this function. Unlike functions such as breathing exercises that users abandon shortly after they acquire the relevant skills [7], self-monitoring requires continuous data accumulation, which is suitable for long-term usage. Thus, the retention rate is expected to increase. Previous studies emphasize the effectiveness

of the self-monitoring function and identify a key area for the design of health apps. The problem is that these studies focus on functions and overlook user experience and psychological factors.

Beyond specific functions, the overall user experience plays an equally pivotal role in retaining users. Increasing automated data input, providing more historical analysis data, regularly updating information, and offering feedback on physical activities can enhance users' long-term engagement and effectively intervene in their health behavior [15]. Human feedback, appropriate reminders, and in-app support from peers or coaches promote the retention of mHealth apps, while a lack of support features, technical difficulties, and the usefulness of the app obstructs retention [16]. Low subscription fees, positive user ecosystems, and personalization were classified as factors that influence the uptake of and engagement with health apps [17], and long-term user engagement enhances retention rates. Contrary to expectations, chatbots can improve user engagement but do not necessarily improve retention [18]. Research has shown that app functions such as right-fit guidance, real-time feedback, and expert consultation are highly accepted and of maximal utility among users, which in turn could result in higher retention rates and enhance intervention effectiveness [19]. In addition, simplifying the interface and reducing the difficulty of obtaining information can save time and improve the user experience, thus promoting user retention [19, 20]. It was found that users overlook security issues, but users develop a strong demand for reliable personal data protection systems when services become paid. In addition, compatibility with other devices is another key factor in user satisfaction [21]. Related works pay much attention to various health app features that contribute to user engagement, experience and satisfaction, and the content is rich and comprehensive. Although user experience affects long-term engagement, it should not be fully relied upon, because Kaveladze's research showed that a good experience does not necessarily mean a high retention rate.

This suggests that factors are internal to the user, such as their psychological predispositions and ingrained habits. In addition, these factors should be considered to fully understand retention dynamics. User psychology and behavioral patterns are also worthy of attention. Previous studies have shown that individuals' health consciousness affects their use of health apps directly [20], and this is correlated with retention. Health consciousness moderates the effect of perceived usefulness and ease on user satisfaction, whereas perceived usefulness and ease are predictors of CI [22]. It should be emphasized that the case of "happy abandonment" is a positive phenomenon and should be distinguished from the traditional case of user attrition. Social and gamification features were also identified as facilitators of sustained app engagement in multiple health contexts [23]. Social features enable users to interact with each other and promote their mutual assistance, while gamification features integrate game elements into nongame contexts, such as setting up challenge tasks and rewards. This finding indicates that the added fun elements are also related to higher retention rates.

Notably, the poor rates of long-term engagement in health apps may be due to the "happy abandonment" of users when they have finally solved their current health problems and no longer need to use those apps consistently [7, 8]. However, in addition

to this unique situation, the retention rates of health apps are jointly determined by a variety of objective factors and can be improved in most cases.

3 Methodology

3.1 Definition and Operation of the Target Variable

The main goal of this study is to identify the factors that influence the retention rate of health apps. Ideally, the retention rate can be defined as the proportion of users who remain active in apps after a certain time window [24]. However, the retention rate in its classical sense cannot be directly calculated since the cross-sectional survey dataset employed in this study (source: Kaggle, Health Apps Evaluation Dataset) lacks time-series data on user behavior.

This theory indicates that users' satisfaction is the key antecedent of the intention to continue using. To address the data limitations, this study draws on the expectation-confirmation model of IS continuance (ECM-IS) [13]. In this study, the overall user rating is used as a proxy for retention intention. Specifically, the target variable (proxy retention variable) is the standard value from the "non-clinician reviewer rating" field in the dataset after operations and definition. This field was selected because of its superior data completeness compared with clinician ratings and its relatively uniform distribution (Figure 1).

Rating reflects users' immediate satisfaction or perceived values rather than actual long-term behaviors. However, according to the ECM-IS, there is a significant positive correlation between user satisfaction and long-term existence [13]. Thus, predicting factors that influence user satisfaction is valuable for understanding and addressing potential existing problems.

To achieve this predictive objective, three machine learning models with varying levels of complexity and interpretability were selected and compared. There are three different machine learning models for predicting the retention rate:

3.2 Linear Regression

Linear regression represents a linear relationship between independent variables X and a dependent variable y .

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \quad (1)$$

where β_0 is the intercept, $\beta_1 \cdots \beta_n$ are coefficients, and ϵ is the error term.

This interpretable model serves as the baseline for measuring linear patterns. Its performance is used to evaluate whether more complex models (gradient Boosting and XGBoost) perform better.

3.3 Gradient Boosting

Gradient boosting is a powerful machine learning technique that views function estimation as a numerical optimization problem in function space, rather than in traditional

parameter space [25]. This method builds a strong predictive model by combining multiple weak learners, such as decision trees, through stagewise additive modelling. The model is initialized, and each weak learner is sequentially fitted to the data. Each subsequent learner is trained on the pseudoresiduals (negative gradients) of the previous learner.

The algorithm iteratively reduces the residuals, which are the differences between the current predictions and observed values. Through this process, the fit of the ensemble model gradually improves to the data.

$$F_0(x) = \arg \min_y \left\{ \sum_{i=1}^N L(y_i, y) \right\} \quad (2)$$

For iteration $m = 1$ to M :

The core of this algorithm is that, at each iteration, the new model learns from the negative gradient direction of the current model's predictions, rather than from the original response variable.

Compute pseudoresiduals:

$$\gamma_{im} = - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)} \quad (3)$$

The pseudoresidual γ_{im} represents the negative gradient of the loss function about the current model output $F_{m-1}(x)$. This gradient points to the direction where the loss falls the fastest, guiding the next learner $h_m(x)$ to correct the error.

Fit a decision tree $h_m(x)$ to the residuals γ_{im} .

Update model:

$$F_m(x) = F_{m-1}(x) + \gamma_{im} h_m(x) \quad (4)$$

where

$$\gamma_m = \arg \min_{\gamma} \left\{ \sum_{i=1}^N L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \right\} \quad (5)$$

is the optimal step size, and ν (learning rate) controls shrinkage, which is usually set $(0, 1]$ to prevent overfitting.

The advantage of gradient boosting is its versatility: it can handle any differentiable loss function, such as squared error, absolute error, or Huber loss, making it ideal for both regression and classification tasks. Additionally, when regression trees are used as base learners, the algorithm demonstrates strong resilience to noisy data and outliers, which makes it well-suited for real-world datasets that are often incomplete or imperfect [25]. By incorporating regularization through the learning rate ν and the number of iterations M , the algorithm effectively controls overfitting and improves generalization. Through tools such as variable importance analysis and partial dependence plots, it can uncover relationships between features and the target variable, offering valuable insights for analysing retention rates in health applications.

3.4 XGBoost

XGBoost made several key improvements based on the traditional gradient boosting framework. XGBoost incorporates a regularized learning objective to effectively control model complexity and prevent overfitting, and it uses a second-order approximation to quickly optimize the objective, accelerating the model's convergence rate and improving the accuracy of the parameter estimates [26]. In this study, variables such as user behavior, app subscription fees, and individual average level of app use include both numerical and categorical data types. XGBoost can automatically handle sparse data and missing values, reducing the need for extensive preprocessing [26]. Predicting retention rates requires a model with strong expressive power and generalizability, and XGBoost ensures prediction accuracy with its regularized objective function and boosting mechanism. Additionally, its feature importance scoring mechanism facilitates the identification of factors with the greatest impact on user retention rates.

Unlike standard gradient boosting, XGBoost introduces key enhancements such as regularization and second-order optimization to improve performance and prevent overfitting.

Its objective function is as follows:

$$\text{Obj} = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (6)$$

where

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (7)$$

T is the number of leaves and ω is the leaf weight.

XGBoost optimizes the objective function via second-order Taylor expansion. Let

$$g_i = \partial_{\hat{y}_i^{(t-1)}} L(y_i, \hat{y}_i^{(t-1)}) \quad (8)$$

and

$$h_i = \partial_{\hat{y}_i^{(t-1)}}^2 L(y_i, \hat{y}_i^{(t-1)}) \quad (9)$$

be first/second-order gradients. The objective at step t simplifies to the following:

$$\text{Obj}^{(t)} \approx \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (10)$$

The optimal weight ω_j^* for a leaf j and the resulting gain are as follows:

$$\omega_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}, \quad \text{Gain} = \frac{1}{2} \left[\frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} - \gamma \right] \quad (11)$$

Where I_j is the instance set in leaf j . A node split is executed only if the corresponding gain value is positive; otherwise, the split stops.

Regularization constraints:

$$\gamma \geq 0 \text{ (minimum loss reduction to split a node)}$$

$$\lambda \geq 0 \text{ (L2 regularization term on leaf weights)}$$

3.5 Evaluation Metrics

3.5.1 R-squared

A numeric value for variance, which aims to explain how well the independent variables can illustrate the dependent variable in a statistical model, ranging from 0 to 1, where 1 indicates a perfect fit of the model to the data. It can be negative when the model performs worse than the horizontal mean line does.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (12)$$

3.5.2 MSE

The average of the squared differences between the predicted and actual values. Theoretical Range: $MSE \geq 0$, with 0 indicating perfect prediction. The value is unbounded above and scale dependent.

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (13)$$

The relationship between MSE them R^2 is derived from (12) and (13):

$$R^2 = 1 - \frac{n \cdot MSE}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (14)$$

The RMSE provides error the magnitude of the error in target units, whereas the MAE is the L1-norm counterpart of the MSE (L2-norm).

3.5.3 RMSE

The square root of MSE. $RMSE \geq 0$, the same scale as the target variable. $RMSE \geq MAE$ by Jensen's inequality.

$$\frac{RMSE}{MAE} \geq 1$$

with equal values when all the errors are equal. It can be used to penalize large prediction errors.

$$RMSE = \sqrt{MSE} \quad (15)$$

3.5.4 MAE

The average of absolute difference between the predicted and actual values. The mathematical property of linear loss makes its applications robust to outlier influences. Theoretical Range: $\text{MAE} \geq 0$, with 0 indicating perfect prediction. The partial derivative with respect to prediction is the sign function:

$$\frac{\partial \text{MAE}}{\partial \hat{y}_i} = \text{sign}(\hat{y}_i - y_i) \quad (16)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (17)$$

Error bound:

$$0 \leq \text{MAE} \leq \max_i |y_i - \hat{y}_i|$$

3.5.5 Spearman's Rank Correlation Coefficient (ρ)

A nonparametric measure of rank correlation that assesses the strength and direction of the monotonic relationship between two variables. Unlike Pearson correlation, Spearman's ρ does not assume linearity and is robust to outliers, making it particularly suitable for ordinal data such as user ratings.

It is defined as follows:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (18)$$

where d_i is the difference between the ranks of the i th observation in the two variables, where n is the number of observations.

In this study considers both numerical accuracy and rank preservation are considered when in evaluating model performance. Four numerical metrics are employed. The R^2 measures the degree of variance matching and goodness of fit between the predicted values and actual values. On the basis of the requirement to tightly control scoring bias, the mean squared error (MSE) is sensitive to large errors. The root mean squared error (RMSE) produces results in units that are consistent with the original data, enabling intuitive representation of the error magnitude. By squaring errors, the RMSE amplifies the influence of large deviations, which helps detect extreme outliers. Given the presence of outliers, the mean absolute error (MAE) is less sensitive to them (such as extreme errors) and avoids causing overall evaluation bias due to individual extreme values. Given that the target rate has moderate variance, both the RMSE and the MAE are applicable. In addition, Spearman's ρ is used to evaluate whether the models preserve the ordinal ranking of the user ratings (1-5 scale). A value close to 1 indicates strong positive rank concordance, whereas a value near 0 suggests no monotonic relationship. Statistical significance is assessed at the $\alpha = 0.05$ level.

4 Methods

4.1 Data Source

The dataset used in this study is the Health Apps Evaluation Dataset, which is publicly available on Kaggle[27]. The raw data were collected through a survey on user's

satisfaction with mobile health applications, including variables such as users' behavior, subscription fees and in-app functions. In addition, the training set contains 137 records (participants) with 50 features each.

The key research subjects are summarised as follows: user retention rate, expert, communication with medical workers, app compatibility and motivation, app subscription fees and ratings from users. There are two aspects that explain the user's retention rate: the recommendation of clinical reviewers and the usage frequency of nonclinical reviewers. The determinant of the rate is the highest evaluation of all the reviewers. The variable for expert features indicates whether the frequency might increase when the expert clinic function is added, which also indicates the apps' medical treatment, influencing users' behaviors, coded as 1 for yes and 0 for no. Communication with medical workers is among the most important factors, as it offers a more convenient and effective treatment service. In this research, medical workers were categorized as nurses or doctors. To identify whether an app involves the relevant application, the relevant data are encoded as 1 for yes and 0 for no. In addition, there are three major types of app subscription fees, namely, the price to download, in-app purchases and annual subscription fees, providing an intuitive perspective for understanding the distribution of the purchase price, which is represented by boxplots. Meanwhile, the varieties are classified to recognize the features of subscription fees. This study also uses proxy retention rate variables and dependent variables, which are only aimed at the nonclinical reviewer's field. This dataset comprise rich numerical and behavioral data. This abundant dataset not only facilitates exploratory data analysis (EDA) of user behavior and app structures but also reveals user preferences and evaluations of health apps.

4.2 Data Preprocessing and Exploratory Analysis

4.2.1 Data Preprocessing

The raw data require preprocessing before being used for machine learning models. With the goal of addressing inconsistencies and improving data quality, this process enables the construction of a more accurate model. The main steps include the following. Handling missing values is the first step to prevent them from adversely affecting the model training process. Next, the context variables are transformed into numeric depictions in making them suitable for machine learning. After encoding categorical variables, feature engineering is important for feature scaling to improve performance, which assigns numerical features such as the user agreement level and subscription fee variety.

4.2.2 Data Visualization

The histograms and distributions of users' retention, which include clinical recommendations and nonclinical frequency, are shown in Figure 1. The x-axis represents the level of clinical recommendation data, nonclinical frequency data are clustered in a higher score range. Given the higher concentration of scores in the nonclinical frequency group, this study focuses on nonclinicians who are investigating drivers of sustained engagement.

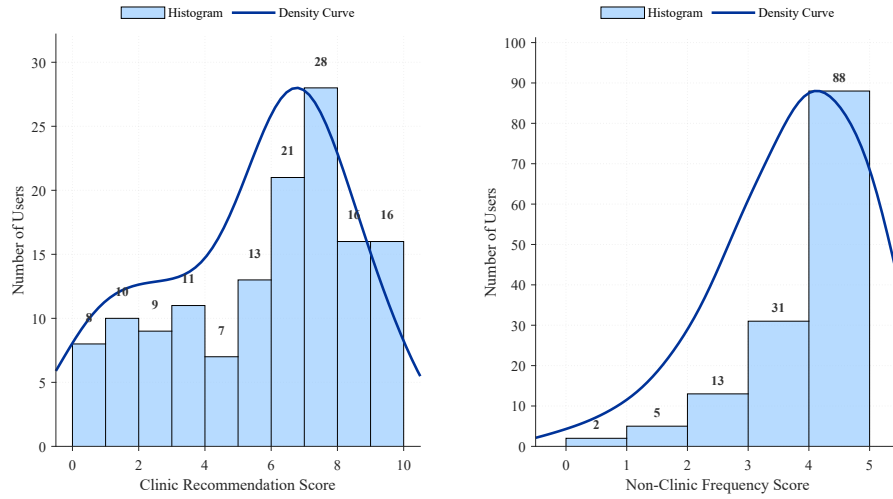


Fig. 1 Histograms and distributions of users' retention rate aspects.

The count plots for the categorical features are shown in Figure 2, which displays the distribution of communications with medical workers, in-app function types, and the professionalism of the app. As shown in Figure 2a, communication with medical staff, whether nurses or doctors, is regarded as an important factor in health apps. However, the data in Figures 2b and 2c reveal two key shortcomings. First, many health apps lack both performance compatibility and features that support internal motivation. Second, despite enabling user communication with medical workers, expert involvement remains limited. This lack of medical professionalism may undermine consumer trust and contribute to user attrition.

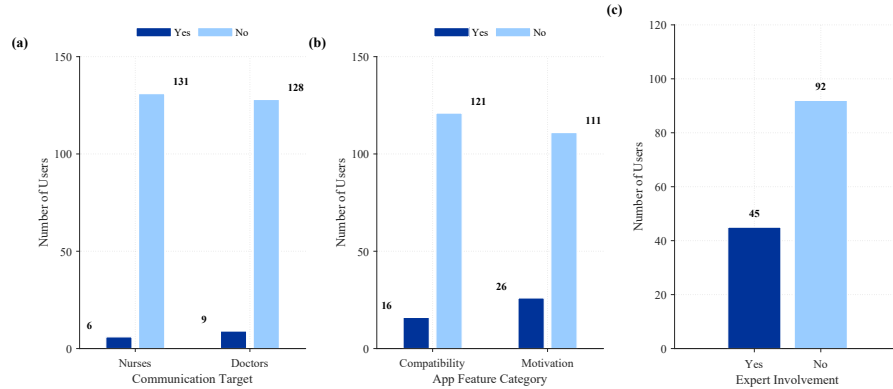


Fig. 2 Histograms and boxplots of numerical features.

The grouped boxplots in Figure 3 and Figure 4 illustrate the distributions of in-app prices and download prices, each segmented by two different grouping variables. One grouping is based on the operating system (Android, iOS, or both), whereas the other

is based on the reviewer type (clinician versus non- clinician). For both in-app and download prices, compared with iOS, apps available on Android tend to have lower median prices, Notably, the prices of the apps that support both operating systems are higher in both categories, which may reflect the increased costs associated with cross-platform development and maintenance. The price distribution patterns also differ by reviewer type. Non- clinicians exhibit a wider range of in-app purchase prices than clinicians do, suggesting greater variability in their willingness to pay. In contrast, clinicians show a wider range of download prices, which may indicate higher sensitivity to upfront costs.

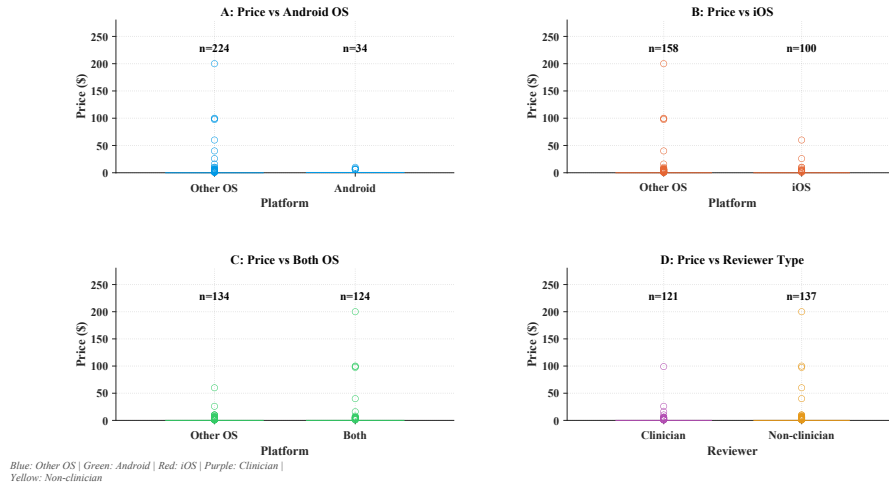


Fig. 3 Grouping boxplot of price in-app.

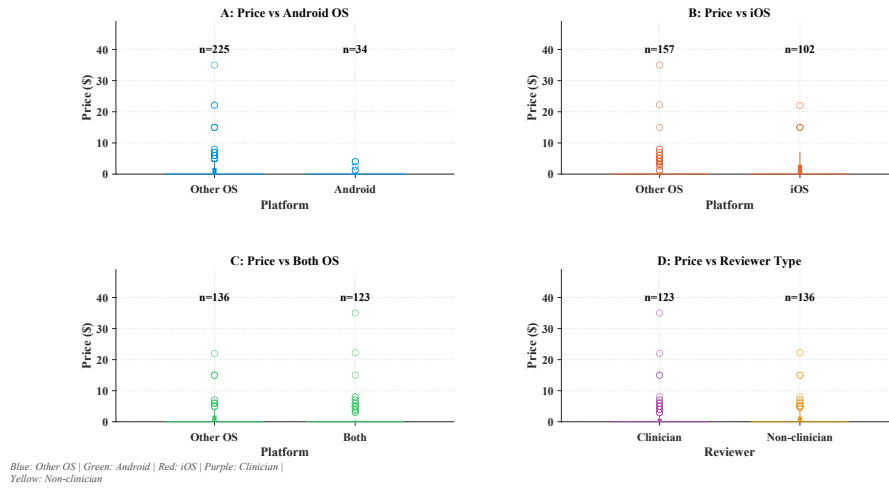


Fig. 4 Grouping boxplot for price downloads.

The correlation heat map in Figure 5 emphasizes the relationships between various natures in the dataset. This visualization assists us in classifying the impacts on each variable and provides valuable insights for feature engineering.

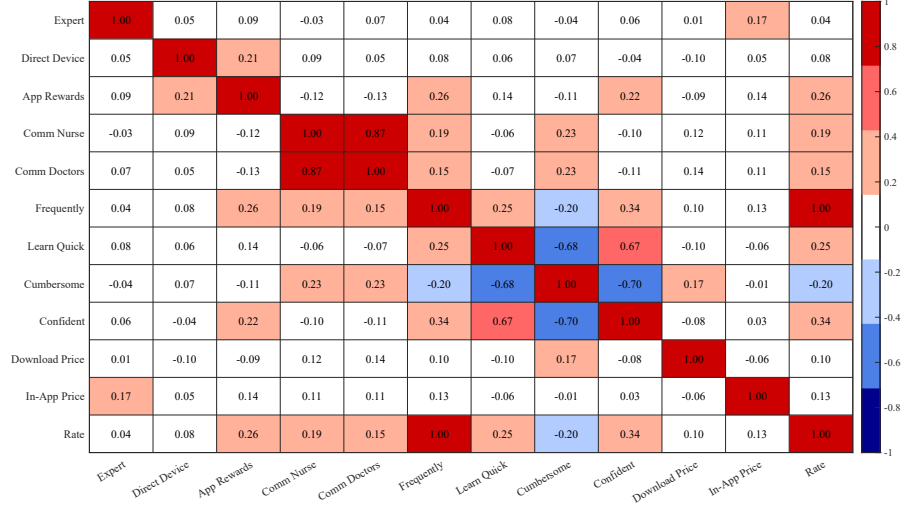


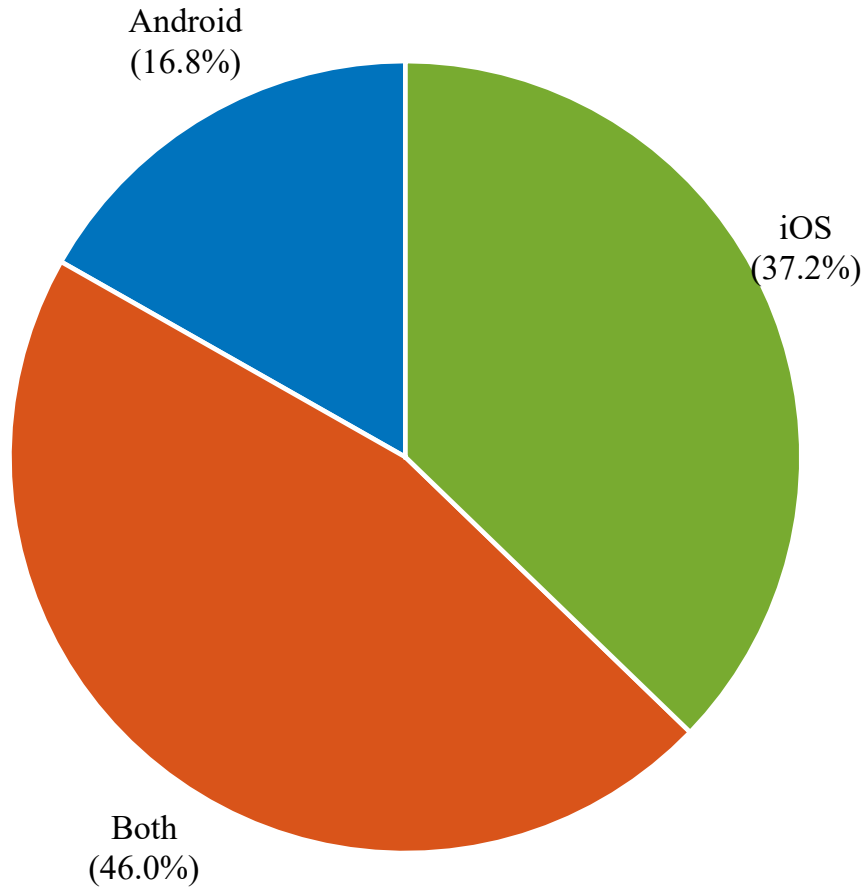
Fig. 5 Correlation heatmap.

4.2.3 Feature Engineering

Two major parts of feature engineering exist: creating new features and transforming existing ones to enhance the predictive power of the models. This research combines the theoretical framework, the unified theory of acceptance and use of technology (UTAUT) and the expectation–confirmation model of IS continuance (ECM-IS), contributing to the understanding of the relationships between feature engineering and the target variables. In this study, we derive three further features:

User Grouping is necessary to understand the user behavior associated with their operating system (OS) preferences. To this end, the distribution of app users was analysed, leading to the grouping of users into 'Android', 'iOS', and 'Both' categories (Figure 6). This grouping provides developers with clear insights into user preferences. This grouping directly reflects users' existing Facilitating Conditions [16, 17], a key construct in UTAUT. Facilitating conditions indicates that the existence of organizational technical infrastructures can support the degree to which the system is used personally. If users continue using a type of operating system and its products, such as iPhone and Apple Watch, this could construct a facilitating condition to lower application adoption and the threshold for sustained use [16], affecting the retention rate.

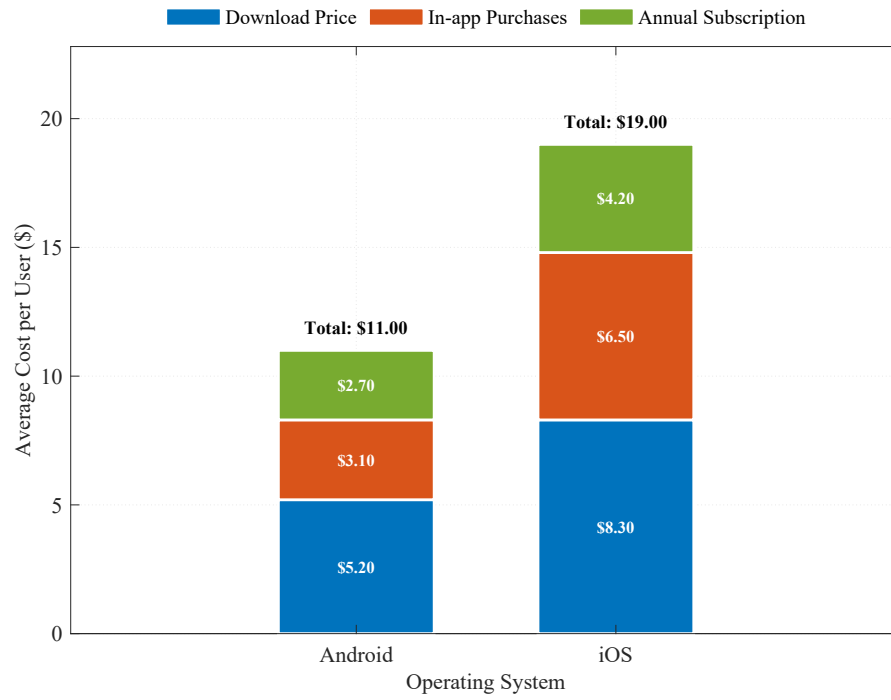
After discovering that the relative costs of subscription fees for each reviewer in apps differ because of distinct operating systems, and it can be categorized into three types, this research classifies subscription fees as 'Annual fee to purchase', 'In-app



Note: Total $N = 274$. Values represent percentage of users per platform.

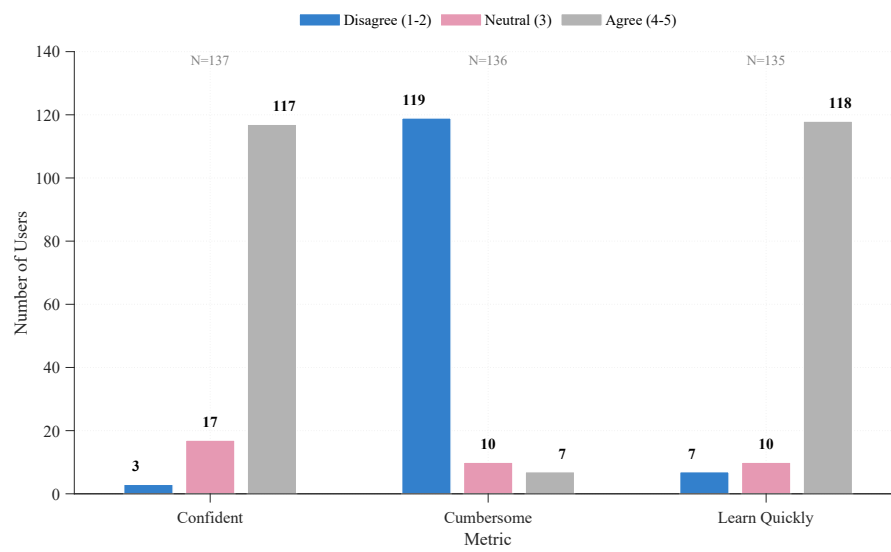
Fig. 6 Distribution of types of app users .

spend', and 'average download price' on the basis of the numeric values of subscription fees. This segment assists in understanding the various costs of using the apps (Figure 7). According to the ECM-IS, the effectiveness of using a certain system continuously from the users' perspective is significantly affected by the involved cost assessments [12]. This price segment allows models to capture the nonlinear relationship between price sensitivity and user satisfaction, which could indirectly influence the retention rate. For example, price sensitivity suggests that reasonable in-app costs can enhance the user experience, thereby improving retention. Thus, reasonable in-apps costs can enhance the consumer experience so that users' willingness to use the application can be improved. To determine whether the difficulty level of using apps influences the app retention rate, observing the distribution of the level of agreement with the



Note: Values represent average costs per user. Android: n=XX, iOS: n=YY

Fig. 7 Distribution of subscription fee variations.



Note: Sample sizes differ due to missing responses. Confident: N=137, Cumbersome: N=136, Learn Quickly: N=135. Total possible respondents: 137.

Fig. 8 Distribution of user agreement levels about the difficulty level of using the apps .

use of related apps is necessary. Therefore, this study categorizes the difficulty level of using the apps into three standards: average level, easy level, middle level, and challenging level, which are determined by the users' agreement level. The visualization reveals clear distinctions between the levels, offering a clearer understanding of user satisfaction (Figure 8). In addition, this feature operationalizes the effort expectancy construct from UTAUT, which is defined as the degree of ease associated with using a system [10]. High effort expectancy, and low ease of use, are a known barriers to initial adoption and can lead to dissatisfaction during continued use, negatively impacting retention [10, 13]. By segmenting the difficulty level, the models can more effectively assess usability. Aspects of usability that impact the user experience thereby indirectly influence the intention to retain the app.

4.3 Model

In this adopts, three distinct machine learning models are adopted to predict the app retention rate: Simple linear regression, gradient boosting, XGBoost.

These three models use the preprocessed dataset. To compare the performances and screen the optimal model, the standard of evaluation includes four metrics: R-squared, root mean squared error (RMSE), mean squared error (MSE), and mean absolute error (MAE). The R-squared, which is the coefficient of determination of the regression variables, and specifies the proportion of variance related to the dependent variable, should ideally be close to 1, indicating a strong fit. These error metrics quantify the discrepancies between the actual values and the model predictions. Therefore, lower values of the MAE, MSE, and RMSE suggest better accuracy.

4.3.1 Model Building

The dataset was initially split into a training set (80%) and a testing set (20%). The model-building process commenced with hyperparameter tuning for XGBoost. A parameter grid was defined, and K-fold cross-validation was employed in conjunction with methods such as GridSearchCV and RandomizedSearchCV to identify the optimal hyperparameter combination within the designated search space.

To capture potential synergistic effects, interaction terms were subsequently created between key features. These primarily included four interactions: the difficulty level of using apps and communicating with medical workers, the types of operating systems (OSs) used and the difficulty level, difficulty level and incentive policies, and communications with medical workers and incentive policies.

A preliminary analysis was subsequently conducted to investigate specific relationships, such as the potential link between OS preference and perceived usability. This involved processing the dataset containing products and renewal variables. The variance inflation factor (VIF) was computed to examine multicollinearity before a simple linear regression model was fitted to quantify the influence of the explanatory variables.

Finally, this study employed a performance comparison. The gradient boosting and XGBoost models were trained on the training set to minimize the loss function and improve the predictive accuracy.

4.3.2 Model Results

Model performance was evaluated using four metrics: R-squared (R^2), mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). Detailed definitions of these metrics are provided in Section 3.5.

5 Results

5.1 Evaluations of the optimal model

This section reports the performance of the three machine learning models under their optimized configurations (i.e., with hyperparameter tuning and the engineered features/interaction terms described in Section V-A). These results serve as the benchmark for selecting the optimal predictive model. The performance metrics of three models, namely linear regression (LR), gradient boosting (GB), and XGBoost, are systematically evaluated across multiple statistical dimensions. The results are shown in Table 1 and Table 2.

Table 1 Results of the machine learning models

Models Metrics	Linear Regression	Gradient Boosting	XGBoost
R^2	0.786	0.869	0.872
RMSE	0.1034	0.058	0.057
MSE	0.0107	0.003	0.003
MAE	0.0826	0.029	0.028

Table 2 Spearman’s rank correlation for model prediction

Model	Spearman’s Correlation Coefficient (ρ)	*p*-value
Linear Regression	0.6061	0.0006*
Gradient Boosting	0.3729	0.0506
XGBoost	0.6142	0.0005*

*Note: Spearman’s rank correlation assesses whether models preserve the ordinal ranking of user ratings. $p < 0.05$ indicates statistically significant rank concordance. *

From the perspective of numerical prediction accuracy, the ensemble learning paradigms (gradient boosting and XGBoost) demonstrate a foundational advantage over the traditional linear regression baseline. Linear regression yields an R^2 of 0.786, with an RMSE of 0.1034, an MSE of 0.0107, and an MAE of 0.0826. While this finding indicates that a linear combination of app attributes can explain a substantial proportion of the variance in nonclinician reviewer ratings, it fails to accommodate the intricate, nonlinear interactions inherent in user psychological evaluations.

In contrast, XGBoost achieves the best predictive performance, delivering the highest goodness-of-fit, with an R^2 score of 0.872, paired with the lowest error magnitudes (RMSE = 0.057, MSE = 0.003, and MAE = 0.028). Gradient boosting follows closely, generating an R^2 of 0.869 and an RMSE of 0.058. The marginally but statistically significant superiority of XGBoost over gradient boosting can be attributed to its implementation of second-order Taylor expansions for the objective function and its integrated L_1/L_2 regularization mechanisms, which effectively suppress model overfitting in a relatively compact sample size ($n = 137$).

Although both Gradient gradient boosting and XGBoost outperform Linear linear regression, the performance gap between these two advanced models is minimal. This makes it difficult to definitively determine the superior model and raises the question of whether the rank order of predictions aligns with the numerical metrics. To address this, this study introduces Spearman’s correlation coefficient and its accompanying statistical test to evaluate the consistency between the numerical values and the rank order of the dependent variable (user ratings). The results are presented in Table 2.

Both XGBoost and linear regression exhibit highly significant Spearman correlations ($p < 0.05$), with nearly identical coefficients (0.6142 vs. 0.6061). This statistical validation provides a robust foundation for selecting XGBoost as the optimal model, moving beyond a mere comparison of metric values. In contrast, gradient boosting was not significantly different ($p > 0.05$) and was significant at the margin ($p = 0.0506$). Furthermore, its Spearman correlation coefficient ($\rho = 0.3729$) is the lowest among the three models.

The gradient boosting algorithm has relatively high numerical prediction accuracy (coefficient of determination $R^2 = 0.869$), but its Spearman correlation coefficient is rather low, and the two are not contradictory. This phenomenon highlights the key difference between single-point prediction accuracy and ranking preservation ability. Although gradient boosting can output predictions close to the true values numerically, it may mix up the order for samples with similar scores—for example, predicting 4.8 for a user whose true rating is 5.0, and predicting 4.9 for a user whose true rating is 4.0. This results in a high R^2 but a low Spearman correlation coefficient.

In the actual context of health app ratings—which inherently represent an ordinal scale (1–5 discrete scores)—the model’s capacity for rank preservation carries is highly practical utility and often supersedes coordinate proximity. XGBoost’s built-in regularization helps it maintain correct ordinal relationships, even if its numerical precision is closely matched by alternative methods. This unique ability to balance stringent numerical accuracy with robust rank preservation provides a mathematically sound foundation for selecting XGBoost as the optimal predictive architecture. The low mean errors across the ensemble approaches further suggest that features spanning user behavioral frequency, communication channels with medical professionals, and price vectors constitute a highly stable system that continuously shapes long-term user satisfaction and retention intent.

5.2 Feature importance analysis

XGBoost was selected as the optimal model because of its superior numerical performance and statistically significant Spearman’s correlation coefficient. Under these

stable conditions, identifying the key determinants of user ratings becomes essential. However, the dataset contains only 137 samples, which limits the reliability of feature importance analysis. The restricted range and imbalanced data may lead to unstable feature rankings. To address this limitation, this research first employed bootstrap analysis with 500 iterations to expand the effective sample size and assess the stability of feature importance across different sample compositions.

5.2.1 Bootstrap validation of feature rankings

Table 3 Bootstrap validation of feature importance rankings

Feature	Median Rank	Rank IQR	Top1 (%)	Top3 (%)	Stability
individual average level	1	[1,2]	65.2%	97.8%	High
communication with nurses	2	[2,3]	14.0%	90.6%	Moderate
expert	3	[2,3]	14.8%	75.8%	Moderate
in-app incentive policy	4	[3,4]	6.0%	37.8%	Low
sync directly with a peripheral device	5	[4,5]	0.0%	2.0%	Low

Note: Based on 500 bootstrap iterations. The median Rank indicates the median ranking across all iterations. The rank IQR shows the interquartile range (25th-75th percentile). The top 1 (%) and top 3 (%) represent the percentage of iterations where the feature ranked first or among the top three, respectively. Stability classification: High ($\geq 50\%$ Top 1), moderate ($\geq 70\%$ Top 3), low ($< 70\%$ Top 3).

Table 3 presents the validation results, identifying one feature with high stability and two with moderate stability. These three features consistently rank among the top three in more than 80% of the bootstrap samples. The individual average level emerges as the most important and robust determinant of user ratings (median rank = 1, rank IQR = [1,2], top 1 = 65.2%, top 3 = 97.8%). Communication with nurses (median rank = 2, rank IQR = [2,3], top 1 = 14.0%, top 3 = 90.6%) and expert (median rank = 3, rank IQR = [2,3], top 1 = 14.8%, top 3 = 75.8%) exhibited moderate stability, with some variation between ranks 2 and 3. This provides strong evidence that these features are genuinely important determinants of user ratings, rather than artifacts of the specific dataset. In contrast, in-app incentive policy and direct synchronization with a peripheral device have low stability, suggesting that their influence may be sample-specific rather than generalizable. These findings confirm that the top three features represent stable, generalizable determinants of user satisfaction.

5.2.2 Pareto diagram of key determinants

After feature importance is examined through bootstrap validation, a Pareto diagram and partial dependence plots (PDPs) are introduced to further elucidate the model’s decision-making process. After confirming the stability of the top features, an XGBoost model was trained on the full dataset to obtain precise importance scores. The resulting

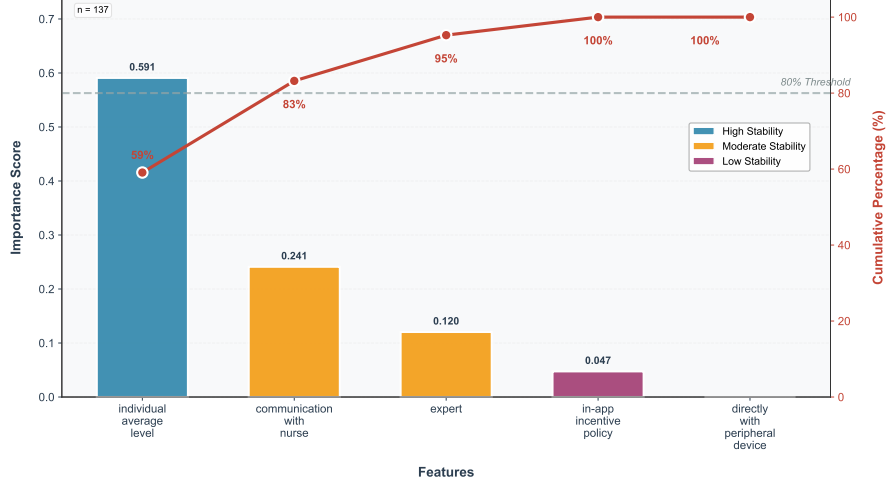


Fig. 9 Pareto diagrams of feature importance with bootstrap stability validation

Pareto diagram is shown in Figure 9, with colors indicating the stability levels derived from Table 3.

The Pareto chart illustrates the cumulative contribution of features to the XGBoost model’s predictions. “Difficulty of use” and “communication with medical workers” were identified as the two most influential predictors, together accounting for more than 80% of the total feature importance. These results highlight the dominant role of usability and motivational design in determining user satisfaction, which serves as a proxy for retention intention.

In detail, the diagram confirms that the three high -stability features account for approximately 95% of the total predictive power, with “Difficulty of use” alone contributing 59%. These findings align with the Pareto principle, suggesting that developers should prioritize these features when optimizing app design to enhance user retention.

5.2.3 Ablation Study: validation of the Pareto principle via feature pruning

To scrutinize the isolated marginal contributions of the predictor variables and validate the empirical stability of the identified core features, an ablation study was systemically conducted on a baseline XGBoost configuration. By evaluating four distinct feature subsets under default hyperparameter constraints, we explicitly isolated the predictive power inherent in the feature architectures from hyperparameter tuning artifacts. The statistical trajectory of the ablation metrics is summarized in Table 4.

A cross-model comparative analysis yields several critical insights into the underlying data dynamics. Most notably, Model 2 (Only the top 3 features), which restricts the training input to the highest-ranking stable features (“Individual average level”,

Table 4 Ablation study results across four feature configurations (RMSE and Spearman’s Correlation Coefficient (ρ))

Model	RMSE	Spearman’s Correlation Coefficient (ρ)
Model 1 (Full 10)	0.2677	0.1890
Model 2 (Only Top3)	0.2513	0.1794
Model 3 (Only Bottom7)	0.2666	0.1858
Model 4 (Full-Top3)	0.2666	0.1858

”Communication with nurses”, and ”Expert involvement”), achieves the lowest absolute error profile, with an RMSE of 0.2513. This surpasses the accuracy of Model 1 (full 10 features), which has an RMSE of 0.2677.

From an information-theoretic perspective, this phenomenon strongly demonstrates that the bottom 7 features introduce substantial stochastic noise rather than constructive signals. In restricted sample regimes ($n = 137$), the inclusion of these low-stability features forces the unregularized tree structure to fit spurious correlations, leading to localized overfitting and generalized error inflation.

This interpretation is further verified by the performance equivalence observed between Model 3 (Only Bottom 7 Features) and Model 4 (Full Minus Top 3). Both models exhibit identical diagnostic footprints (RMSE = 0.2666, $\rho = 0.1858$). This absolute numerical convergence establishes that removing the top 3 core features completely strips the model of its primary predictive architecture, rendering the presence or absence of remaining variables computationally inconsequential.

However, it is vital to note that while Model 2 minimizes the numerical distance (RMSE), its rank preservation capability ($\rho = 0.1794$) decreases slightly compared with that of Model 1 ($\rho = 0.1890$). This finding implies that although the bottom-tier features largely act as predictive noise, they carry weak monotonic signals that assist the model in resolving fine-grained ranking thresholds among closely tied user responses.

Ultimately, the ablation matrix reinforces the Pareto principle established in the feature stability analysis: the top 3 variables contain the vast majority of the core predictive signals, and concentrating engineering and optimization resources on these key dimensions provides the highest return on investment for app retention modelling.

5.2.4 Partial dependence of higher rank determinants

To understand how the most stable features influence user satisfaction, this research also uses partial dependence plots (PDPs) with bootstrap confidence intervals. Each feature identified as having ”high” or ”moderate” stability in Table 3 (which is consistently ranked among the top three stable determinants) was subjected to 100 bootstrap iterations; then each feature was retrained with the XGBoost model and computed its partial dependence curve was computed.

The PDP for the ”individual average level” is shown in Figure 10a. The solid blue line shows the mean partial dependence across bootstrap samples, while the shaded area represents the 95% confidence interval. The positive monotonic relationship is clearly visible and statistically robust, as the confidence interval remains entirely above

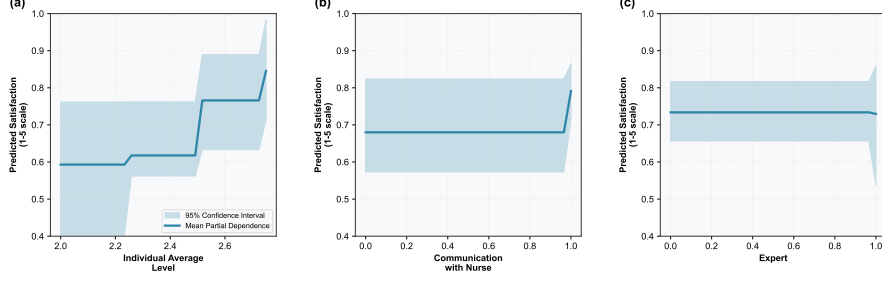


Fig. 10 Partial dependence plots of XGBoost

the baseline across the feature range. This finding indicates that users with higher average engagement consistently rate apps more favourably.

The PDP for “communication with nurses” is shown in Figure 10b. As the feature shifts from 0 (absence) to 1 (presence), the predicted satisfaction score increases by approximately 0.15 on the 1-5 scale. The narrow confidence intervals across the baseline confirm the stability of this positive empirical effect.

The partial dependence plot for the “expert” feature is shown in Figure 10c. The predicted satisfaction remains constant across different expert scores, with the 95% confidence interval confirming the stability of this nearly null effect. This suggests that the presence of expert-related features within the application system may not be a primary driver of user satisfaction.

A comparison between Figure 10b and Figure 10c reveals an important distinction: users appear to value effective communication more than the passive absorption of expert knowledge. This finding implies that interactive consulting services should be prioritized over static expert content when optimizing internal application features for user satisfaction.

In summary, XGBoost demonstrated superior predictive ability over both LR and GB. Through bootstrap validation, Pareto analysis, ablation experiments, and PDPs, the framework consistently identified the individual average level and communication with nurses as the primary drivers of user satisfaction, with the ablation study empirically confirming that the bottom- seven features introduce stochastic noise under sample constraints. These findings provide a data-centric foundation for subsequent discussion.

6 Discussion

6.1 Conceptual scope and proxy justification

This study examines perceived user satisfaction as a psychological construct related to app continuance, rather than directly observed retention behavior. As defined in the Methodology section, this research operationalizes perceived satisfaction related to continuance. In the context of mobile health applications, app retention-related evaluations reflect users’ psychological evaluation of apps. Prior information systems

research has confirmed the significant association between user satisfaction and continuance intentions in postacceptance usage models [28]. Users perceive psychological satisfaction as an indicator for evaluating whether an app is worth continuous use. This perception influences users’ long-term willingness to continue using the same app or other apps under the same brand. More specifically, the psychological satisfaction of apps embodies users’ overall experience and promotes greater trust in the brand and reliance on its products. Psychological evaluation is formed by individuals’ internal thoughts and complex emotions. Thus, it is difficult for us to directly quantify users’ inherently subjective recognition. However, in the absence of detailed behavioral logs, the recommendation rating scores can capture users’ degree of psychological satisfaction. Prior research has shown that people are more likely to recommend satisfying products to others because of their continuous willingness to use products. Thus, clinician recommendation rating scores represent users’ perceptual satisfaction and can be used as an operationalization of perceived satisfaction related to continuance.

6.2 Model interpretation boundary conditions

The analysis of these predictive models captures features that are potentially related to users’ perceptual satisfaction in this study. High R-squared values indicate a consistent relationship between internal app design features and users’ subjective satisfaction. Low mean errors suggest that these features form a relatively stable system and continue influencing users’ perceived satisfaction throughout the usage process. Statistical tests further support the stability of the identified influential mechanisms within health applications. The superior performance of the ensemble models reveals the complexity of users’ psychological evaluation mechanisms. This suggests that users’ perceived satisfaction is not determined by a single feature but is formed through interactions among multiple features. Partial Dependence Plots illustrate how internal app features influence users’ perceptual satisfaction and reflect the underlying psychological processes involved in its formation.

6.3 Reconciling ablation findings with model optimization

To examine the importance of key determinants from the best predictive model (XGBoost), an ablation study was conducted on a baseline model without hyperparameter tuning, while the original optimal XGBoost model achieved higher numerical accuracy.

Unlike the baseline model, GridSearchCV and RandomizedSearchCV serve as tools to identify optimal hyperparameters. These include the tree depth and regularization strength. The tuning process minimizes the validation error by searching for the best combination. Although these procedures amplified the numerical value of R^2 , they did not alter the individual features themselves. This means that the relative importance of the intrinsic features did not change with hyperparameter adjustments, and the signal-to-noise ratio among the variables remained stable. Moreover, small samples may intensify the model’s sensitivity to features with high noise levels. Additionally, insufficient regularization may cause models to overprocess features with high noise

levels. These features often exhibit spurious correlations with training data. This type of overprocessing can easily go unnoticed.

In addition, compared with the model using only the top three features, the full-feature baseline has a higher RMSE. Thus, tuning enhances overall predictive accuracy but does not overturn the fundamental ranking of feature importance. Crucially, the results of the Pareto analysis, bootstrap validation, and ablation study confirmed that the top three features remained dominant regardless of the model configuration.

Therefore, the stable importance ranking of the top three core determinants reflects genuine underlying relationships rather than artifacts of hyperparameter optimization. Tuning contributes to numerical refinement but does not affect the theoretical interpretation of which features drive user satisfaction.

6.4 Applicability and practical implications

Within the defined scope of perceived user satisfaction, the findings offer several practical implications for health app developers and service providers.

From a strategic perspective, these findings provide valuable insights for organizations involved in health app development. Specifically, the findings suggest a two-phase strategy for optimization: First, efforts should focus on enhancing app attributes, including usability, compatibility, and communication; subsequently, the focus should shift to maintaining incentive policies and emphasizing user OS preferences. This integrated approach may help improve resource allocation efficiency and support sustainable service development.

These implications are derived from patterns associated with users' perceived satisfaction rather than directly observed usage behavior. At the operational level, the enterprise can focus on the following key resource areas on the basis of the model analysis: First, prioritize the optimization of the application. Since 'difficulty of use' is identified as a key influential factor, simplifying the user interface and designing guidance procedures for beginners are crucial for minimizing users' learning costs and avoiding initial user loss. Second, systematic incentive mechanisms should be established. The model confirms the effectiveness of the "incentive policy" and suggests the introduction of game elements such as goal setting, achievement badges, and progress feedback, thereby transforming health management into a more engaging and enjoyable experience to enhance long-term user engagement. Finally, efforts should focus on enhancing the specialized medical characteristics of the application. The feature importance of "communication with medical workers" indicates that the core competitiveness of the application relies on its professional credibility. Integrating lightweight expert content such as popular science articles and Q&As or collaborating with professional institutions, can significantly enhance users' sense of trust and reliance.

Overall, implementing personalized and compatibility-focused strategies is necessary after users' subjective factors such as OS preferences and the need for convenience are considered. Optimizing the interface for different operating systems and ensuring that intelligent wearable devices such as wristbands and smartwatches can connect to health data platforms can provide users with a comprehensive experience.

7 Conclusion

This study validates the effectiveness of machine learning models in predicting user satisfaction, which serves as an immediate factor influencing app retention rates. Linear regression is limited to capturing linear correlations between features and user satisfaction and cannot model complex nonlinear interactions within users' psychological evaluation logic, resulting in inferior predictive performance. Among the two ensemble tree algorithms, XGBoost is identified as the optimal model: it leverages second-order Taylor expansion to accelerate objective optimization and embeds $L1/L2$ regularization terms to restrain overfitting, presenting stronger generalization ability under the small sample ($n=137$) condition of this research, and it precisely determines the key drivers of perceived satisfaction. With respect to health app user ratings measured on a discrete 1–5 ordinal scale, the model's rank preservation capacity reflected by the Spearman's rank correlation coefficient is more than pure numerical fitting accuracy is. The resulting insights offer a robust, data-driven foundation for app design optimization and engagement strategy refinement. Future research should directly investigate additional determinants and more accurately estimate the coefficients of app retention rates to support more comprehensive optimization recommendations.

Declarations

- Ethics approval and consent to participate: Not applicable. This study used a publicly available, anonymized dataset from Kaggle (Health Apps Evaluation Dataset). The dataset does not contain any personally identifiable information, human biological samples, or clinical intervention data. Therefore, ethical approval and participant consent were not needed.
- Consent for publication: Not applicable.
- Funding
- Availability of data and materials: The dataset analysed during the current study is publicly available in the Kaggle repository: <https://www.kaggle.com/datasets/subhamsarkar2002/mobile-health-apps-evaluation-dataset>
- Competing interests: The authors declare no known competing interests that could have appeared to influence the work reported in this paper.
- Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
- Authors' contribution: Ruisi Liu, Yining Xu and Youjia Gu contributed equally to this work and should be considered co-first authors.
 - Ruisi Liu: Exploratory data analysis, data visualization, discussion drafting, implementation of feature importance analysis results, and manuscript format adjustment.
 - Yining Xu: Methodology development, implementation of model evaluations (including optimal model selection and comparative performance analysis), and results presentation.
 - Youjia Gu: Introduction and literature review, conclusion drafting, and reference compilation and formatting.

- All the authors reviewed and approved the final manuscript.
- Acknowledgements: Not applicable.

References

- [1] Soriano, J.B., *et al.*: Prevalence and attributable health burden of chronic respiratory diseases, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet Respiratory Medicine* **8**(6), 585–596 (2020) [https://doi.org/10.1016/S2213-2600\(20\)30105-3](https://doi.org/10.1016/S2213-2600(20)30105-3)
- [2] Lv, J.-C., Zhang, L.-X.: Prevalence and disease burden of chronic kidney disease. In: *Advances in Experimental Medicine and Biology* vol. 1165, pp. 3–15 (2019)
- [3] Arias, D., Saxena, S., Verguet, S.: Quantifying the global burden of mental disorders and their economic value. *eClinicalMedicine* **54**, 101675 (2022) <https://doi.org/10.1016/j.eclinm.2022.101675>
- [4] Rehm, J., Shield, K.D.: Global burden of disease and the impact of mental and additive disorders. *Current Psychiatry Reports* **21**(2) (2019) <https://doi.org/10.1007/s11920-019-0997-0>
- [5] Chong, S.O.K., Predron, S., Abdelmalak, N., Laxy, M., Stephan, A.J.: An umbrella review of effectiveness and efficacy trials for app-based health interventions. *npj Digital Medicine* **6**(1), 1–11 (2023) <https://doi.org/10.1038/s41746-023-00981-x>
- [6] Lee, K., *et al.*: Effect of self-monitoring on long-term patient engagement with mobile health applications. *PLOS ONE* **13**(7), 0201166 (2018) <https://doi.org/10.1371/journal.pone.0201166>
- [7] Baumel, A., Muench, F., Edan, S., Kane, J.M.: Objective user engagement with mental health apps: Systematic search and panel-based usage analysis. *Journal of Medical Internet Research* **21**(9), 14567 (2019) <https://doi.org/10.2196/14567>
- [8] Kaveladze, B.T., Wasil, A.R., Bunyi, J.B., Ramirez, V., Schueller, S.M.: User experience, engagement, and popularity in mental health apps: Secondary analysis of app analytics and expert app reviews. *JMIR Human Factors* **9**(1), 30766 (2022) <https://doi.org/10.2196/30766>
- [9] Slime, K., Maizte, A., Hassouni, L., Najat, N.: Toward a model to predict cardiovascular disease risk using a machine learning approach. *LAENG International Journal of Computer Science* **51**(5), 519–527 (2024)
- [10] Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D.: User acceptance of information technology: toward a unified view. *MIS Quarterly* **27**(3), 425–478 (2003) <https://doi.org/10.2307/30036540>

- [11] Marikyan, D., Papagiannidis, S.: Unified Theory of Acceptance and Use of Technology: A Review. <https://open.ncl.ac.uk/theories/2/unified-theory-of-acceptance-and-use-of-technology/> Accessed 2025-07-17
- [12] Oliver, R.L.: A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research* **17**(4), 460–469 (1980) <https://doi.org/10.1177/002224378001700405>
- [13] Bhattacharjee, A.: Understanding information systems continuance an expectation-confirmation model. *MIS Quarterly* **25**(3), 351–370 (2001) <https://doi.org/10.2307/3250921>
- [14] Zhao, J., Freeman, B., Li, M.: Can mobile phone apps influence people’s health behavior change? an evidence review. *Journal of Medical Internet Research* **18**(11), 287 (2016) <https://doi.org/10.2196/jmir.5692>
- [15] Adu, M.D., Malabu, U.H., Malau-Aduli, A.E., Drovandi, A., Malau-Aduli, B.S.: User retention and engagement with a mobile app intervention to support self-management in australians with type 1 or type 2 diabetes (My Care Hub): Mixed methods study. *JMIR mHealth and uHealth* **8**(6), 17802 (2020) <https://doi.org/10.2196/17802>
- [16] Amagai, S., Pila, S., Kaat, A.J., Nowinski, C.J., Gershon, R.C.: Challenges in participant engagement and retention using mobile health apps: A literature review. *Journal of Medical Internet Research* **4** (2022) <https://doi.org/10.2196/35120>
- [17] Szinay, D., Jones, A., Chadborn, T., Brown, J., Naughton, F.: Influences on the uptake of and engagement with health and well-being smartphone apps: Systematic review. *Journal of Medical Internet Research* **22**(5) (2020) <https://doi.org/10.2196/17572>
- [18] Cevasco, K.E., Morrison, R.E., Woldelessie, R., Kaplan, S.: Patient engagement with conversational agents in health applications 2016-2022: A systematic review and meta-analysis. *Journal of Medical Systems* **48**(1) (2024) <https://doi.org/10.1007/s10916-024-02059-x>
- [19] Mangalji, A., *et al.*: Uncovering care barriers experienced by patients and loved ones in the emergency department: Can a health app help? In: *Studies in Health Technology and Informatics* (2020). <https://doi.org/10.3233/shti200470>
- [20] Cho, J., Park, D., Lee, H.E.: Cognitive factors of using health apps: Systematic analysis of relationships among health consciousness, health information orientation, ehealth literacy, and health app use efficacy. *Journal of Medical Internet Research* **16**(5), 125 (2014) <https://doi.org/10.2196/jmir.3283>
- [21] Kwon, J.-Y., Lee, J.-S., Park, T.-S.: Analysis of strategies to increase user

- retention of fitness mobile apps during and after the COVID-19 pandemic. *International Journal of Environmental Research and Public Health* **19**(17), 10814 (2022) <https://doi.org/10.3390/ijerph191710814>
- [22] Yan, M., Filieri, R., Raguso, E., Gorton, M.: Mobile apps for healthy living: Factors influencing continuance intention for health apps. *Technological Forecasting and Social Change* **166**, 120644 (2021) <https://doi.org/10.1016/j.techfore.2021.120644>
 - [23] Jakob, R., *et al.*: Factors influencing adherence to mhealth apps for prevention or management of noncommunicable diseases: Systematic review. *Journal of Medical Internet Research* **24**(5), 35371 (2022) <https://doi.org/10.2196/35371>
 - [24] Cho, J.: The impact of post-adoption beliefs on the continued use of health apps. *International Journal of Medical Informatics* **87**, 75–83 (2016) <https://doi.org/10.1016/j.ijmedinf.2015.12.016>
 - [25] Friedman, J.H.: Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29**(5), 1189–1232 (2001) <https://doi.org/10.1214/aos/1013203451>
 - [26] Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794 (2016). <https://doi.org/10.1145/2939672.2939785>
 - [27] Sarkar, S.: Mobile Health Apps Evaluation Dataset. Kaggle (2025). <https://www.kaggle.com/datasets/subhamsarkar2002/mobile-health-apps-evaluation-dataset>
 - [28] Mishra, A., Shukla, A., Rana, N.P., Currie, W.L., Dwivedi, Y.K.: Re-examining post-acceptance model of information systems continuance: A revised theoretical model using MASEM approach. *International Journal of Information Management* **68**, 102571 (2023) <https://doi.org/10.1016/j.ijinfomgt.2022.102571>