Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.38127

Player Retention Prediction Method for Freemium Games

Haoyang Jiang[1], Xintao Nie[2,*]

[1]*School of Artificial Intelligence and Computer Science, North China University of Technology, Beijing, China*
[2]*International School, Beijing University of Posts and Telecommunications, Beijing, China*
**Corresponding Author. Email: 2024213460@bupt.edu.cn*

Abstract. Player retention in freemium games is crucial to the users' lifecycle and operational revenue. This article uses the fixed player queue in Brawl Stars as a case study to construct a reproducible workflow covering battle log cleaning, time window division, player-level feature engineering, and classification model evaluation. This article uses 46 characteristics measured within 24 hours after the player's first baseline observation-including battle count, active span, battle interval, win rate, trophy changes, mode participation, and hero information-to predict whether the players will be active again within the 7th to 8th day window. The latest data includes 166 players and 41,402 post-cleanup battle records. The D7 tag override check passed, with 99 players active in the D7 window and 67 inactive. After player grouping, random forest achieved the highest ROC-AUC of 0.661, accuracy of 0.706, recall of 0.900, and F1 value of 0.783. The results show that early combat behavior has some D7 retention differentiation ability, but the current sample size and a single game source limit the external validity of the conclusions.

Keywords: Freemium games, player retention, early behavior, machine learning, Brawl Stars

1. Introduction

Freemium games attract players with low entry barriers, then realize commercial value through ongoing engagement, in-app purchases, and advertising [1]. Existing work mainly explains churn and retention from the perspectives of early interactions, playtime patterns, social relationships, and player lifecycles. Drachen et al. analyzed early player behavior and retention predictions based on freemium mobile games, demonstrating that early interaction records can support rapid recognition [2]. Yang and others further transformed playtime patterns into entropy and cross-entropy features, and verified the regularity of regularity over the gain from basic gameplay features in six free online games [3]. Kristensen

and Burelli compared aggregated data, sequence data, and models combining both, and found that combining player history aggregated information with behavioral sequences can improve churn prediction effectiveness [4]. Roohi and others, based on automated tryouts, represent player engagement as level completion rate and churn rate, demonstrating that difficulty assessment and engagement prediction can jointly serve game design and player modeling [5]. Óskarsdóttir and others input players' social network information into churn predictions for mobile freemium games

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.38127

and found that social relationships can provide additional predictive information beyond the player's own behavioral traits, indicating that player retention in games is not only related to individual behavior but also influenced by social factors [6]. The above work shows that retention prediction requires attention to both the behavior window, temporal patterns, and the stage in which the player is in. However, most existing studies analyze all players or new players as a whole, and their behavior window settings are often uniformly set based on data availability, with little consideration of the heterogeneity and evolutionary characteristics of fixed player queues in early combat behavior. Existing models mostly focus on identifying long term retention status, lacking research on the dynamic changes in early combat behavior over days and how this differentially affects subsequent retention processes.

Based on these shortcomings, this paper focuses on early combat behavior in fixed player queues, adopts unified D7 retention labels, and groups players based on entry time to empirically test the predictive ability of early combat behavior on retention status, thereby providing a more targeted and repeatable analytical path for retention prediction. The purpose of this article is to answer three questions: First, can early combat behavior predict subsequent retention? Second, among characteristics such as activity level, participation frequency, competitive performance, and content diversity, which variables deserve more attention from operations? Furthermore, with limited sample sizes, battle logs as scrolling windows, and incomplete label observations, how to reduce data leakage risks through strict time window design and player grouping? Based on real Brawl Stars data, this paper constructs a reproducible data processing flow, transforms original single game records into multidimensional behavioral features, conducts experiments on three types of classification models, and focuses on examining the impact of play characteristics on classification tasks, providing a foundation for future research.

2. Data set and data quality control

This paper selects Brawl Stars as the dataset mainly based on two aspects: first, the game experienced significant retention fluctuations, with a 7-day retention rate of about 30%, and publicly available operational cases show it has been practically applied to retention optimization decisions, providing analysis scenarios closely related to the business for this study. Second, its battle logs can meticulously record player combinations, character usage, and win/loss details in a single battle, supporting the construction of early behavioral characteristics from multiple dimensions such as activity level, participation frequency, competitive performance, and content diversity, meeting the core needs of this article focusing on combat behavior analysis. Based on these characteristics, the Brawl Stars dataset is an appropriate choice for examining the framework between early combat behavior and retention in this paper.

For Brawl Stars, the collection script continuously requests player profiles and battle logs from the same batch of players, and retains different collection batches through the collection time field. The data conversion module reads the player information table and battle log table, using player identifiers as the connection key, and later aggregates records at the player or player snapshot level. The advantages of this data source lie in clear field structure, timestamped records, continuous re-extraction, and longitudinal tracking by players, making it suitable both for early retention modeling and for reconstructive empirical analysis.

Table 1 summarizes the player queue, collection batch, number of combat records, observation window, and other information to illustrate the sample size and data coverage.

DOI: 10.54254/2755-2721/2026.38127

Table 1. Dataset size and actual use

Data items	Current state	Practical role
Player queue	166 players	Suitable for longitudinal observation in the same cohort
Collection batch	13 batches	Traceable batch time and coverage
Cleaning battle record	41402 records	Used to construct early combat behavior characteristics
Observation window	24 hours	Construct features only use information from earlier windows
Strict D7 modeling table	166 lines	D7 label has been completed, and model training is available

Data quality control includes four actions. First, the combat time field is parsed and unified to UTC to avoid time zone deviation during cross-batch comparison. Secondly, the identity, time, duplicate records, and unrecognized abnormal records of the missing players were deleted, and 41402 valid combat records were retained in this operation. At the player level, this study randomly generated 2000 player IDS according to the tag coding rules published in the game. After data quality screening and battle record matching, 166 players with at least one valid battle record were finally retained for analysis. Thirdly, the nested fields of battle teams are processed, and the information of heroes, hero levels, and hero trophies used by players themselves is extracted. Finally, before generating the D7 label, check whether the subsequent collection covers the label window. The baseline time is 07:56 UTC on July 23, 2026. The D7 label window is from the 7th day to the 8th day after the baseline. The coverage deadline is 07:56 UTC on July 31, 2026. The latest collection time is 00:17 UTC on August 6, 2026. The quality inspection result has passed. The data source is the player files and battle logs returned from the Brawl Stars public interface. In this paper, the interface fields, collection time, and cleaning rules are all retained in the original CSV and conversion report to ensure that the results are traceable. It should be noted that the public interface does not provide stable real recharge amount, full session duration, and plot task progress, so these variables are not the current empirical characteristics.

3. Data processing and modeling table construction

In the reading phase, players' profiles, battle logs, and batch information are obtained. In the cleaning phase, time resolution, missing data processing, duplicate removal, and result field standardization are completed. In the aggregation phase, a single battle is converted into player-level statistics according to the player ID. In the tag phase, it is checked whether the player is active again in the future time window. In the output phase, the modeling table, cleaning details, player characteristics table, and data quality report are saved. Each stage can be traced in the project main program, data conversion module, and converted CSV file.

$$X_i = Aggregate\{b : a_i \le t_b < a_i + 24h\} \tag{1}$$

Where a_i is the baseline time of player I and t_b is the combat record time. This formula indicates that the player level feature X_i is only aggregated from the combat records in the 24-hour observation window after the baseline.

DOI: 10.54254/2755-2721/2026.38127

$$y_i = 1\{\exists b : a_i + 7d \le t_b < a_i + 8d\} \tag{2}$$

When player I has at least one activity record in the window from day 7 to day 8, the D7 tag y_i takes 1, otherwise it takes 0, but only if the window has been completely observed.

4. Early behavior characteristic engineering

In the aspect of early behavior feature engineering, Bonometti and others built indicators based on players' early game interaction behavior to estimate players' retention time and loss probability [7]. This is consistent with the idea of using early game behavior of players to build D7 retention prediction features in this study. This paper further analyzes the early game behavior of players from the dimensions of the number of games, failure, and combat interval, in order to analyze the predictive value of different behavior characteristics on short-term retention.

The current code has 46 player level features, which can be divided into three groups according to the method framework: activity, participation frequency, competitive performance and content diversity. The feature does not directly copy the original log, but converts the single battle record into a statistical index that can explain the player's input and participation. The current feature group is divided into three categories according to keywords: activity, frequency, performance and progress. Some variables may belong to multiple key phrases at the same time. Therefore, the number of groups is used to describe the engineering structure, which is not equal to the number of mutually exclusive variables.

4.1. Activity and participation frequency

$$ActiveSpan_i = \frac{(t_{last} - t_{first})}{3600} \tag{3}$$

Where, t_{first} represents the time of the first battle in the observation window, t_{last} represents the time of the last battle, and 3600 is used to convert seconds to hours. This indicator depicts the continuous active span of the player in the early window. The larger the value, the longer the player remains active during the observation period.

$$BattleDensity_i = \frac{N_{battle}}{N_{activeday}} \tag{4}$$

Where, N_{battle} represents the total number of battles in the observation window, and $N_{activeady}$ represents the number of active days in which battles have occurred. This indicator reflects the average number of battles participated in every day and can be used to describe the input intensity.

$$AvgInterval_i = mean(t_j - t_{j-1}) \tag{5}$$

Where, t_j is the time of the jth battle, t_{j-1} is the time of the previous battle, and mean is the average of all adjacent time differences. This indicator is used to describe the frequency of participation and the regularity of behavior. The higher the number, the more intense the battle.

4.2. Competitive performance and content diversity

$$WinRate_i = \frac{N_{win}}{N_{knownresult}} \tag{6}$$

DOI: 10.54254/2755-2721/2026.38127

Where, N_{win} represents the number of battles won in the known results, $N_{knownresult}$ represents the total number of battles with known results, including victory, failure and peace. This indicator is only calculated on battles with clear results and is used to measure competitive performance.

$$TrophyDelta_i = \sum battle_{trophychange} \tag{7}$$

Among them, $battle_{trophychange}$ refers to the change of trophies brought by a single battle. This indicator is used to indicate the progress change or income change of players in the early window.

$$Diversity_i = (\#modes, \#maps, \#brawlers) \tag{8}$$

Where, $\#modes$ 、 $\#maps$ and $\#brawlers$ respectively represent the number of modes, maps and heroes that players have contacted. This set of variables is used as a proxy indicator of the breadth of players' content participation. The larger the value, the wider the contact area of the game.

Player portrait variables such as total trophies, experience levels, and club information can be used to describe samples, but formal early prediction models need to be used with caution. The reason is that these portrait fields may be updated during subsequent collection. If the latest snapshot is directly used, future information may be brought into the feature window. The recharge amount, plot progress, and full session duration are currently not stable in the Brawl Stars public interface, so they remain the direction of subsequent data expansion.

5. Model construction and performance evaluation

In terms of model selection, Laia and Fibriani employed Random Forest to categorize the engagement levels of online game players, achieving commendable classification performance. This demonstrates that Random Forest possesses certain advantages in state recognition based on player behavior characteristics [8]. Therefore, this paper adopts Random Forest as one of the candidate core models and further validates its performance in the D7 retention prediction task through experiments.

$$p_i = \sigma(\beta_0 + \beta^T X_i) \tag{9}$$

At the model level, three complementary methods are retained. The logical regression model is used as an interpretable linear baseline, the decision tree model is used to capture simple nonlinear rules, and the random forest model processes the interaction and combination of features through the integration of multiple decision trees.

$$\widehat{p}(x) = (\tfrac{1}{B}) \sum T_{b(x)}, B = 300 \tag{10}$$

Where $\widehat{p}(x)$ represents the prediction probability of the input sample x, x represents the eigenvector of the player to be predicted, B represents the total number of trees, where 300 is taken, $T_{b(x)}$ represents the prediction result of the b decision tree on sample x, b represents the number of the tree. This formula shows that the random forest forms the final judgment by averaging the outputs of multiple trees.

In the operation scenario, the recall rate has a particularly important explanation, because missing potential lost players may mean missing the intervention window. The formal experiment should report the precision, accuracy, recall, F1 score and ROC-AUC at the same time, and give the

DOI: 10.54254/2755-2721/2026.38127

confusion matrix to avoid exaggerating the model performance with a single index. The formal experiment should employ player grouping segmentation and, when sample size permits, be supplemented by repeated cross-validation or bootstrap confidence intervals.

6. Stage experiment results and conclusion boundary

The current D7 empirical operation results include 166 lines of modeling data, of which 99 players are D7 active positive and 67 players are inactive negative, with a positive proportion of 59.64%. Table 2 shows the evaluation indices of the three models, Table 3 summarizes the corresponding confusion matrix, and Figure 1 shows the ranking of feature importance, which is used to compare the recognition effect of the model and analyze the main variables affecting the prediction.

Table 2. Model performance comparison

Model Name	ROC-AUC	Accuracy	Precision	Recall	F1
Random Forest	0.661	0.706	0.692	0.900	0.783
Logistic Regression	0.611	0.647	0.700	0.700	0.700
Decision Tree	0.593	0.559	0.647	0.550	0.595

Table 3. Comprehensive comparison of the confusion matrices of three models

Model Name	TP	FN	FP	TN
Random Forest	18	2	8	6
Logistic Regression	14	6	6	8
Decision Tree	11	9	6	8

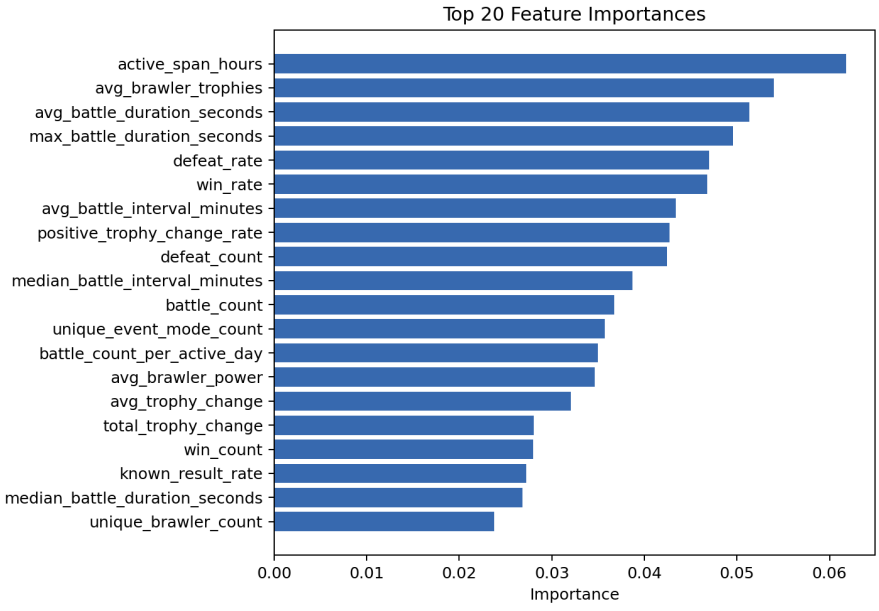


Figure 1. Feature importance ranking (photo/picture credit: original)

DOI: 10.54254/2755-2721/2026.38127

Figure 1 shows the ranking of feature importance. The top variables include active_span_hours, avg_brawler_trophies, avg_battle_duration_seconds, max_battle_duration_seconds, defeat_rate, win_rate, and avg_battle_interval_minutes. A total of 166 lines of modeling data were formed during the official operation of D7, including 99 people in the positive class and 67 people in the negative class. The ROC-AUC of the random forest, logistic regression model, and decision tree model were 0.661, 0.611 and 0.593, respectively. This result is based on the current fixed player queue and a test divided by players, belonging to the phased empirical results. The stability still needs to be tested by repeated cross-validation and confidence intervals.

The above results show that the early combat behavior can provide certain D7 discrimination information, but the results under a single fixed queue and a single grouping segmentation should still be regarded as preliminary evidence, and the subsequent need for repeated cross-validation and confidence intervals. In the division of training tests grouped by players, random forest performed best. It should be clear that the current formal D7 modeling table has generated 166 rows, and the label coverage check has passed. The model results demonstrate that early combat behavior possesses a certain discriminative ability but that the data originate from a single game and a fixed player queue and that the test set comprises only 34 players. Therefore, the current indicators cannot be directly interpreted as a general rule applicable to the general player group, nor can it be inferred that there is a causal relationship between behavior characteristics and retention.

7. Discussion

The focus of this section is to discuss what this set of results actually demonstrates and what it cannot demonstrate. From a modeling perspective, prior research on player churn prediction has shown that machine learning models are capable of effectively utilizing player behavioral features to identify potential churn risk [9]. At this stage, this is more like a

Among these, the significance of the 24-hour observation window is mainly to provide a unified and comparable early benchmark, but it is not naturally sensitive to all types of players. For high-frequency hardcore players, the 24-hour battle, stay, and rhythm changes are often enough to show a stable participation mode. For casual players, the short window may only capture sporadic attempts or intermittent logins, which is difficult to fully reflect their subsequent backflow and viscosity differences. Therefore, it is necessary to test its robustness in larger-scale samples and rolling-window prediction.

From the perspective of characteristic contribution, the most important thing to be explained is not the individual failure rate or average combat interval, but the $ActiveSpan$. $ActiveSpan$ describes the length of time that players continue to participate in the observation window. It emphasizes "how long they have been active continuously", rather than "how many games they have played". In this result, $active_span_hours$ ranked first in the importance of characteristics, indicating that the duration of continuous activity is more able to distinguish D7 state than the isolated match result. The matching avg_battle_interval_minutes, avg_battle_duration_seconds, max_battle_duration_seconds, win_rate and defeat_rate are more like depicting the rhythm, feedback and friction of players' early participation: the longer the interval, the lower the participation frequency and the higher the defeat_rate, which may mean a stronger sense of early frustration and a greater exit risk.

Han and Lee discussed the impact of game mechanisms on continuous participation from the perspective of competence and immersion experience, and pointed out that feedback, goals, and rewards will further affect the willingness to continue playing by changing the player's experience [10]. This is consistent with the results of this article: early behavior characteristics reflect not only

DOI: 10.54254/2755-2721/2026.38127

the degree of activity, but also the differences in rhythm, progress, and experience of players. Based on this, future work should, in addition to enlarging the sample, unifying field mappings, and conducting independent testing, incorporate repeated group cross-validation, bootstrap confidence intervals, observation window comparisons, and stratified analysis by player type to assess the stability and transferability of these features across different queues. It should be particularly emphasized that the current conclusions are still based on 166 players and a single grouped split. Therefore, they are more appropriately regarded as a reference for feature selection and dimensionality reduction in subsequent extensions to larger player cohorts or even rolling-window prediction tasks, rather than being directly generalized as universal conclusions for all games and all player types.

8. Conclusion

In general, the project has completed the unification of problem definition, data set, processing flow, feature engineering, model scheme, and evaluation criteria, and completed the first empirical operation under the strict D7 label. After cleaning and aggregating the original player and the battle log, a player-level modeling table was formed. The early features were constructed with a 24-hour observation window, and the D7 label was constructed with a window from day 7 to day 8, and the comparison of three types of classification models was completed. The current results indicate that random forest achieves the best performance on the player-grouped test set, with an ROC-AUC of 0.661, an accuracy of 0.706, a precision of 0.692, a recall of 0.900, an F1 score of 0.783, and a confusion matrix of TN = 6, FP = 8, FN = 2, TP = 18, all exceeding those of the logistic regression and decision tree models. According to the sample structure of 166 players, 41402 post-cleaning combat records, 99 D7 active players, and 67 inactive players, early combat behavior has a certain ability to distinguish D7 activities.

Furthermore, the value of this paper is not only to get a set of preliminary and effective prediction results, but also to establish a reproducible player retention analysis process: from data collection, cleaning, tag construction to feature engineering and model evaluation, a complete link is formed. Feature importance analysis shows that variables such as active_span_hours, avg_brawler_trophies, avg_battle_duration_seconds, max_battle_duration_seconds, defeat_rate, win_rate, and avg_battle_interval_minutes exhibit strong explanatory power, indicating a stable association between early participation rhythm and competitive performance and subsequent activity status. Although the research is still limited to a single game source and a fixed player queue, this paper has verified the feasibility of the method framework and laid the foundation for the subsequent expansion of samples, supplementary grouping cross-validation, bootstrap confidence intervals, window comparison, and behavior characteristics analysis, and also provided a clear direction for further improving the robustness and external effectiveness of the model.

Author contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Lee, H., Imteyaz, K., & Savage, S. (2025). Playing to pay: Interplay of monetization and retention strategies in Korean mobile gaming. *arXiv preprint arXiv:2504.10714*. https://doi.org/10.48550/arXiv.2504.10714

[2] Drachen, A., Lundquist, E. T., Kung, Y., Rao, P. S., Sifa, R., Runge, J., & Klabjan, D. (2016). Rapid prediction of player retention in free-to-play mobile games. *Proceedings of the AAAI Conference on Artificial Intelligence and*

DOI: 10.54254/2755-2721/2026.38127

Interactive Digital Entertainment, 12(1), 23–29. https://doi.org/10.1609/aiide.v12i1.12856

[3] Yang, W., Huang, T., Zeng, J., Chen, L., Mishra, S., & Liu, Y. E. (2022). Utilizing players' playtime records for churn prediction: Mining playtime regularity. *IEEE Transactions on Games, 14*(2), 153–160. https://doi.org/10.1109/TG.2020.3024829

[4] Kristensen, J. T., & Burelli, P. (2019). Combining sequential and aggregated data for churn prediction in casual freemium games. In *2019 IEEE Conference on Games (CoG)* (pp. 1–8). IEEE. https://doi.org/10.1109/CIG.2019.8848106

[5] Roohi, S., Guckelsberger, C., Relas, A., Heiskanen, H., Takatalo, J., & Hämäläinen, P. (2021). Predicting game difficulty and engagement using AI players. *Proceedings of the ACM on Human-Computer Interaction, 5*(CHIPLAY), Article 231. https://doi.org/10.1145/3474658

[6] Óskarsdóttir, M., Gísladóttir, K. E., Stefánsson, R., Aleman, D., & Sarraute, C. (2022). Social networks for enhanced player churn prediction in mobile free-to-play games. *Applied Network Science, 7*(1), Article 82. https://doi.org/10.1007/s41109-022-00524-5

[7] Bonometti, V., Ringer, C., Hall, M., Wade, A. R., & Drachen, A. (2019). Modelling early user-game interactions for joint estimation of survival time and churn probability. In *2019 IEEE Conference on Games (CoG)* (pp. 1–8). IEEE. https://doi.org/10.1109/CIG.2019.8848038

[8] Laia, F. J., & Fibriani, C. (2026). Classification of online game player engagement levels using the random forest algorithm. *Sistemasi: Jurnal Sistem Informasi, 15*(7), 2638–2653. https://doi.org/10.32520/stmsi.v15i7.6514

[9] Peng, L. (2024). Research on machine learning models for predicting player churn. In *Proceedings of the 1st International Conference on Modern Logistics and Supply Chain Management (MLSCM 2024)* (pp. 112–121). https://doi.org/10.5220/0013234700004558

[10] Han, H., & Lee, C. (2026). Stimulating flow in MOBA games: A moderated serial mediation model of gamification mechanics, perceived competence, and player retention. *BMC Psychology, 14*, Article 1152. https://doi.org/10.1186/s40359-026-04853-3