Proceedings of CONF-SEML 2026 Symposium: Importance of Machine Learning Methods and Analysis in Engineering
DOI: 10.54254/2755-2721/2026.AS31576

Predicting Online Gaming Player Behavior and Engagement Based on Machine Learning

Longyi Yuan

College of Science, Mathematics, and Technology, Wenzhou-Kean University, Wenzhou, China
yuanlo@kean.edu

Abstract. Online gaming industry has expanded rather rapidly in the last 10 years, which has intensified competition between the developers and the single games have become harder to distinguish on the market. Although a substantial amount of current literature analyzes the behavior of players, including their persistence and attendance, very little research is devoted to the features of the very games. The given study is devoted to game titles and examines the aspects of their design and features concerning the market in reference to the degree of popularity. Based on the data gathered on 27,075 games on the Steam platform, each one of the titles is classified as Low, Medium, or High popularity using a composite engagement measure. The data is explored and converted into information about the pricing, content area, achievement and activity of players before creating the model. Random Forest model is then used and reaches an accuracy of 66.24 which is greater in comparison to the one achieved using the logistic regression baseline. Its achievement is more accurate on the high-popularity category games where F1-score is 0.780. The achievement-related variables and pricing measures, game age, content scope are likely to be near the top in the feature importance rankings. This tendency suggests that popularity is not a characteristic that is linked to any one defining factor, instead, it is a complex of factors that are combined.

Keywords: Game Popularity Prediction, Machine Learning, Random Forest, Steam Platform

1. Introduction

Online games are the usual entertainment action among the players in disparate regions and age brackets. As the number of players grows, games provide very specific data of in-game play, the purchasing patterns, and the varying activity of players over time. The real-world implications to developers and publishers would be in determining why a particular game would be played and remain played over the years or why a particular game would be played and become less popular a short time after being released. This is a particular concern in the case of free-to-play games, where their revenue is derived more on continued play, in-game purchases and advertisements than on initial sales.

It is difficult to analyze the behavior of players in this case. Gameplay data are frequently of high dimension and heterogeneous and there is no correlation between various actions, which are rarely linear. The interaction patterns can be missed in the traditional statistical techniques due to this level

DOI: 10.54254/2755-2721/2026.AS31576

of intricacy. The machine learning techniques provide more flexibility in dealing with such data and have thus been growing popular in the game analytics studies.

The available literature is mostly dedicated to knowing the behavior of players, particularly, the churn. Indicatively, Lee et al. also introduced a data imputation method which can differentiate between various non-Login periods, such as dormant periods, voluntary breaks and access disruption [1]. This approach to rebuilding behavioral sequences better and enhancing churn prediction performance by modeling these situations either directly or indirectly (so that all missing activity is not treated equally). The given line of work emphasizes the significance of the temporal structure when it comes to analyzing the engagement of players.

Mulla et al. researched the method of predicting retention using a combination of social interaction features, purchase history in the game, and historical activity [2]. They discovered that the performance of the ensemble learning models especially the random forest is better than the simple prediction models. Their findings imply that no other types of social behavior and transaction behavior should be analyzed separately but rather they should be viewed in a combined form.

Mustač et al. investigated the behavior at the stage of players and demonstrated that controlled learning models trained on the data regarding the behavior during the first week of the game can forecast the short-term churn with a sufficient level of accuracy [3]. Their results suggest that EPT frequency and rate of advancement are worthy of prediction of whether the player will be particularly active. Wu et al. also expanded this concept by introducing the variability of behavior to the churn prediction models [4]. They did not only find out the aggregated statistics, but also employed the measures of temporal variance to separate between stable and unstable playing pattern, which enhanced better prediction ability in both monetizing and non-monetizing players.

Besides churn prediction, other studies have been oriented on game design techniques which seek to ensure player engagement is preserved. Romero-Mendez et al. used deep learning models to dynamically modify the difficulty of the game according to the real-time player performance [5]. Their strategy resulted in elevated sustained engagement and lower abstention rates by categorizing skill levels in their players and serving them challenges respectively. This paper demonstrates that personalization would be significant in ensuring that players are kept entertained.

Using the clustering methods, Fernandez-Arias et al. analyzed the gaming behavior of children and provided an array of behavior profiles [6]. Their analysis indicates that excessive gaming is not homogenous, however it depends on personal traits and situational issues. This observation gives rise to the fact that there might be players who do not even respond well to the same retention approach. Kępka and Strzelecki examined the relationship between motivational variables (i.e.) challenge, social interaction, and rewards and further gameplay and recommendation behaviors [7]. The result of their study leads towards psychological satisfaction being a critical factor in sustaining long term participation.

Combined, the current studies have given comprehensive information on the mechanics of player behavior and engagement. Although an extensive amount of literature has analyzed the behavior of the players in online games, the majority of available literature prevails with personal level concerns, like predicting churning, psychological motivations, or adaptive intervention methods. In comparison, the level of research at which the game title is conducted is limited. To be more exact, comparatively low amounts of research examine the relationship between the features of an intrinsic game, which include either pricing structure, genre composition, achievement systems, or long-term engagement design and the popularity of a particular game in the market.

There are a few previous researches that have started to attract an interest on the aspects on the level of the game itself. As an example, Hamari et al. observed that a purely behavior-centered approach would be insufficient to observe any significant aspect of the game, and proposed to look at the qualities of searches, experience qualities, and other features of the structure related to commercial performance [8]. Their discussion demonstrates that the visibility and uptake of a game in the market are associated with practical elements of the quantity of available content and the ease with which a game can be discovered. Nonetheless, a great deal of the available research in this field is founded on qualitative discourse or fairly small samples. The literature that investigates game-title features based on large-scale data is uncommon.

On this background, the current work will be analyzing a dataset of 27,075 games on the Steam platform. The point of the analysis does not look at separate players, but rather at traits of the games themselves. These attributes are the pricing-related actions, composition of genres and contents, achievement systems, and the indicators of engagement. These factors are then studied with the use of machine learning to determine their differences based on popularity categories. According to the modeling outcome, games can be classified into Low, Medium and High popularity categories with the help of supervised classification technique.

2. Method

2.1. Data source and description

The data employed in this study was collected in the Steam Store API and SteamSpy API and has 27,075 titles of games. In each game, basic descriptive data is captured in the dataset which included price, release date, genres, categories, number of achievements, user ratings, number of players estimates, and platform availability. A composite engagement score was computed based on the ownership statistics and community rating information and applied to several games to categorize them into one of the three's popularity group namely Low, Medium, and High.

Some preprocessings were done before the modeling. Entries of no-games were eliminated, and missing or unreasonable entries were imputed by median or mode on numerical and categorical variables, respectively. Besides that, a series of derived features was developed in order to describe in more detail the temporal, content-related and value-based features. These are variables related to time like game age, content complexity variables like number of genres and number of categories, and interaction characteristics variables like price per genre and rating efficiency.

$$PricePerGenre = BasePrice / GenreCount \tag{1}$$

Figure 1 provides an overview of the full preprocessing workflow from raw data to the modeling dataset.

DOI: 10.54254/2755-2721/2026.AS31576

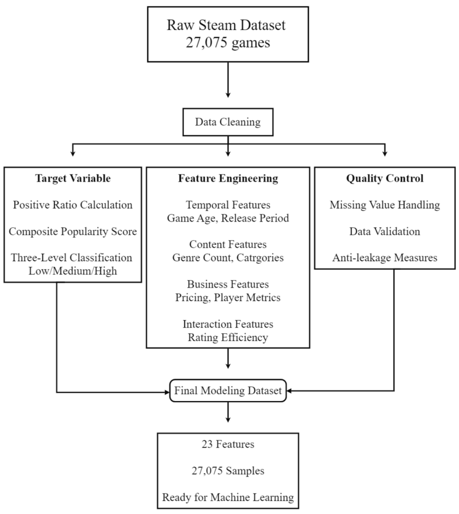

Figure 1. Data preprocessing and feature engineering workflow

2.2. Indicators selection and description

The choice of variables in this research was made out of some pragmatic considerations and not an ultimate preset framework. Variables of price were added to indicate how games are monetized, and achievement information and time-based variables were adopted to indicate the extent of engagement of a game by the player. In order to identify variations among content scopes, the list of genres and categories was documented to each title. The information relating to platforms was also featured as this influences the level to which a game can be widely accessed by the players.

These basic variables, besides, various composite measures were developed so as to further explain value-related relationships. The indicators built show the comparison of price to the content provided or player feedback among some of them. These variables were retained following the preceding research in the field of game analytics and verification of whether the same patterns could be observed at the initial stages of studying the data. An overview of the final set of features employed in the modeling stage is given in Table 1.

Table 1. Summary of engineered features

Feature	Description
GameAge	Number of years since the game's initial release (2025 − ReleaseYear).
GenreCount	Total number of genres assigned to the game.
CategoryCount	Total number of content categories describing gameplay or functionality.
PricePerGenre	Base price normalized by genre count; reflects perceived value.
RatingEfficiency	Ratio of total ratings to player ownership; indicates engagement intensity.
MedianPlaytime	Median playtime in minutes among all players.
AveragePlaytime	Average (mean) playtime; complementary measure of engagement depth.
PlatformSupport	Platforms supported (Windows, macOS, Linux).
DevGameCount	Number of other games released by the same developer.

DOI: 10.54254/2755-2721/2026.AS31576

2.3. Methodology

This paper assumes a workflow that may be characterized as the combination of the exploratory research and the predictive modeling. Prior to the construction of the model, Exploratory Data Analysis (EDA) was conducted to get the preliminary sense of the information. The aim of this step was to test the distributions of variables, to test outliers, and to test the relationship between features as this helped guide further choices regarding feature selection in the model and to guide the model development.

In the case of prediction, a Random Forest (RF) was employed. The reason why RF was selected in the first place was due to its practical benefits in dealing with complication of data sets involving nonlinear relationships, mixed types of features and a relatively high number of variables. The model also utilizes multiple classes of decision trees in training on bootstrap samples of the data and averages the result using majority voting which aids to enhance stability and minimize sensitivity to overfitting. The other benefit of RF is its ability to give guide on the importance of features, such that one can look at which game-related features are more significant in predicting popularity. It has been demonstrated in previous study by Bénard et al. that the Random Forest models can be transformed into interpretable rule-based representations and at the same time retain a strong predictive accuracy [9].

The figure 2 presents the general outline of the model of the Random Forest which had been applied in this research (they can be constructed in parallel) and how the predictions are consolidated.

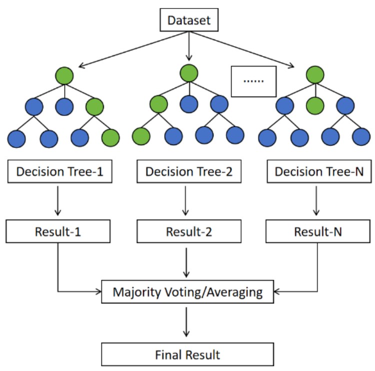

Figure 2. Random forest diagram

Random Forest is suitable in game analytics contexts where the game popularity may often oppose a number of variables such as pricing, content variety, engagement degree, and time influence, interacting with one another. As a matter of fact, these relationships are not straight and may change once some points are passed and as such simple linear models cannot be used. It is also consistent in case features are correlated and variables are of one or more forms. This allows it to be used on large and heterogenous datasets such as those attracted by the Steam platform without having to make strong assumptions regarding data distributions.

In order to offer a current comparison baseline, Logistic Regression (LR) was applied as a linear model that is interpretable. LR model predicts the likelihood of an occurrence to belong to a specific group as shown by the sigmoid function:

$$P\left(y=1x\right)=\frac{1}{1+e^{-\left(w^{T}x+b\right)}} \tag{2}$$

DOI: 10.54254/2755-2721/2026.AS31576

where:
x = input feature vector (e.g., price, genre count, rating metrics)
w = weight vector learned during model training, representing each feature's contribution
b = bias term shifting the decision boundary
$w^T x + b$ = linear combination of features before nonlinear transformation
e = natural exponential base

Both RF and LR were trained using stratified 5-fold cross-validation to ensure balanced class representation and reduce the risk of overfitting. Hyperparameters for the RF model, including the number of trees and maximum depth, were optimized using grid search. In addition, recent studies indicate that Random Forest–based methods perform well in multi-class and ordered classification settings, which supports their suitability for the popularity-tier prediction task considered in this study [10].

Figure 3 presents an illustrative example of the decision boundary produced by the Logistic Regression model, providing a visual reference for its linear separation of the feature space.

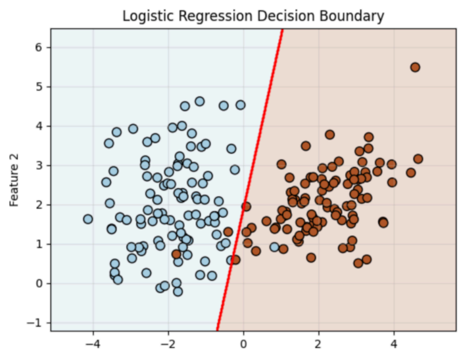

Figure 3. Logistic regression decision boundary

All analyses were conducted in Python 3.10 using Pandas, NumPy, Scikit-learn, and Matplotlib. Model performance was evaluated using accuracy, precision, recall, and F1-score across the three popularity classes.

2.4. Evaluation metrics

Predictions of the models were analyzed by looking at the distribution of the games to the three groups of popularity. Instead of concentrating on one summative measure, the concentration was directed at variations in the manner, with which outcomes were attributed, and points of discrepancy usually located. This allowed creating a more detailed picture of the consistency of the models in their behavior when it came to separating various levels of popularity.

Practically, special attention was paid to the consideration of games which received maximum popularity. When this group is put under lower impact titles, subsequent interpretation becomes skewed, whilst potentially missed valuable patterns are found in one or more games that perform excellently. Due to this reason, Low, Medium and High group results were discussed individually then an overall picture was drawn.

DOI: 10.54254/2755-2721/2026.AS31576

3. Results and discussion

Using the method presented in Section II, the analysis begins by a simple check of the distribution of the target variable. The three popularity ranges (Low, Medium, and High groups) have the same amount of games (n=9,025 in each group) and do not introduce obvious imbalance and do not require more manipulations in order to make comparisons between the groups. The results below are aimed at noticeable associations between variables, variation of model behavior and tendencies linked to the game popularity.

Considering the interrelationship between the features, it is possible to note some consistent tendencies. Most indicators that are associated with player activity tend to vary similarly. Games that have higher owners also get higher ratings and better rating activity is normally seen in games that have more players (r = 0.74 ownership and r = 0.79 rating activity). This trend means that much of the engagement measures are exploring the same features of participation by the players.

Variations are more in pricing related variables. Price has a high relationship with price per genre (r = 0.82), implying that this derived indicator conforms to the same pricing strategy as the underlying price. Meanwhile, games that were linked to more genres have lower values of price per genre (r = 0.32), indicating a potentially higher value compared to the cost. The same trend is observed in the case of players per dollar which is positively correlated with total ratings (r= 0.65) and ownership (r= 0.68). On the one hand, it can be seen that there is a definite correlation between features related to values and the popularity of a game, which points to the relationship when considered together.

Those measures associated with content indicate weaker yet significant effects. Games with a greater diversity of content categories reveal a light connection with the rating efficiency (r = 0.25), which implies that broader content is typically followed by even more effective engagement. An overview of these associations can be seen in Figure 4.

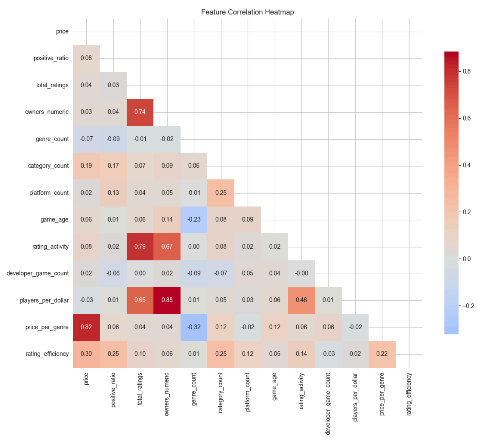

Figure 4. Feature correlation heatmap of games

The Random Forest model is used to classify game popularity into three levels. Under this setup, the Random Forest model reaches an accuracy of 66.24%, whereas logistic regression records a value of 60.41%. To check whether this result is consistent, stratified 5-fold cross-validation is applied. Across the different folds, the Random Forest results remain fairly consistent, with accuracy values clustering around 65%. Prediction quality is not the same across popularity groups. Outcomes for highly popular games are comparatively more stable, with precision and recall values close to 80% and an F1-score of 0.780. By contrast, games in the medium-popularity range are less clearly

DOI: 10.54254/2755-2721/2026.AS31576

separated, as indicated by a lower F1-score of 0.527. This suggests that games in this category share characteristics with both lower- and higher-popularity titles. A summary of class-level performance metrics is reported in Table 2.

Table 2. Classification metrics by popularity class

Popularity Class	Precision (%)	Recall (%)	F1-score
High	79.6	76.5	0.780
Medium	—	—	0.527
Low	—	—	—

Note: Medium- and Low-popularity classes only report available F1-scores because their class-specific precision and recall values were not the primary focus of analysis and are therefore omitted for brevity.

Looking at the feature importance rankings, a small group of variables tends to appear repeatedly near the top. Among the features, achievement-related variables appear near the top, contributing about 9.98% of the overall importance. Several pricing-related measures also rank highly, including price per genre (9.66%) and base price (8.18%), together with game age (7.86%). Rather than pointing to a single dominant driver, these rankings reflect a mix of progression design, pricing decisions, and time on the market that often coincide with higher popularity levels.

Features linked to content scope and development background also appear near the top of the rankings. Category count (7.63%) and developer game count (6.61%) recur in the results, a pattern more commonly seen among games with wider content scope and a longer development background.

There is also another view of the results given by playtime indicators. Playtime (mean playtime of 5.45) and median playtime (6.46) are both on the upper half of the importance value ranking, indicating that the games with more and lasting play cycles are more likely to exhibit more visible trends of popularity than just the initial adoption. Taken collectively, the findings indicate that popularity is not hinged on one element of design, but rather on a composite of the depth of progression, given values, extent of content, and further interaction. The summary of the relative significance of the leading features that have been generated by the Random Forest model is presented in Figure 5.

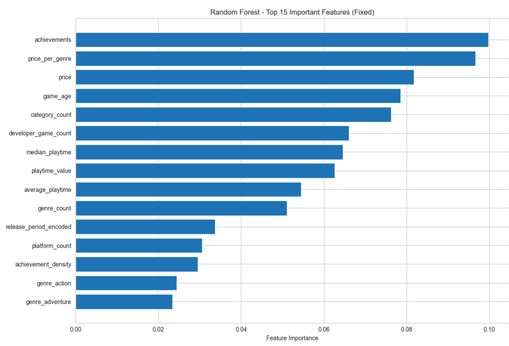

Figure 5. Feature importance ranking from random forest

On the whole, these observations can be related to the research questions provided in the introduction. The analysis reveals that there are a few recurring factors including achievement systems, value-oriented pricing, and content diversity that are always tied to the increased level of popularity. Simultaneously, the model based on the Random Forest shows that these features could prove effective to draw a line between levels of popularity, particularly the highly successful games. This justifies the application of data-driven analysis as a complement applicable in the current methods of game design and marketing.

4. Conclusion

This paper uses machine learning to analyze trends on the popularity of games in the Steam platform. There is an analysis that reveals that, more than one cause exists, but the factors coexist and most often come in play simultaneously. Specifically, the features associated with progression creation, costs, and player retention were replicated in more successful games. The Random Forest model could be used to classify games to various popularity segments with relatively predictive behavior particularly the high-popularity segment. Comprehensively, the discussions indicate that visible aspects of the game (achievements systems, price-related indicators, scope of content and measures of engagement) are strongly related to the variation in commercial performance.

The scope of the current analysis is one of the aspects that should be remembered. The data is a reflection of the data on the one platform at a certain time and therefore, fluctuations in popularity across time are not literally reflected. Subsequently, the case of short-term fluctuation, season, and long-term shifts in the interest of players are beyond the focus of this analysis.

The study is based on the organized numerical data of the prices, content and interaction patterns. Consequently, the player opinion and other contextual data - e.g. written feedback or community discussion - will not be available in the dataset. The discussion is narrowed down to Steam platform and thus, the trends here will not necessarily represent the gaming industry in general but this specific ecosystem.

References

[1] J. Lee, P. Rerkjirattikal, and S. Nam, "A data imputation strategy to enhance online game churn prediction, considering non-login periods, " Data, vol. 10, no. 7, p. 96, 2025.
[2] R. Mulla, S. Potharaju, S. N. Tambe, S. Joshi, K. Kale, P. Bandishti, and R. Patre, "Predicting player churn in the gaming industry: A machine learning framework for enhanced retention strategies, " Journal of Current Science and Technology, vol. 15, no. 2, Art. 103, 2025.
[3] K. Mustač, K. Bačić, L. Skorin-Kapov, and M. Sužnjević, "Predicting player churn of a free-to-play mobile video game using supervised machine learning, " Applied Sciences, vol. 12, no. 6, p. 2795, 2022.
[4] R.-Y. Wu, Y.-H. Hu, and E.-Y. Chou, "Predicting the churn patterns of monetizers and non-monetizers: Exploring the influence of behavioral variability in churn prediction, " Internet Research, in press, 2024/2025.
[5] E. A. Romero-Mendez, P. C. Santana-Mancilla, M. Garcia-Ruiz, O. A. Montesinos-López, and L. E. Anido-Rifón, "The use of deep learning to improve player engagement in a video game through a dynamic difficulty adjustment based on skills classification, " Applied Sciences, vol. 13, no. 14, p. 8249, 2023.
[6] I. Fernández-Arias, M. Labrador, M. Bernaldo-de-Quirós, F. J. Estupiñá, M. Vallejo-Achón, I. Sanchez-Iglesias, M. González-Álvarez, and F. J. Labrador, "Cognitive patterns and problematic use of video games in adolescents: A cluster analysis, " International Journal of Environmental Research and Public Health, vol. 20, no. 24, p. 7194, 2023.
[7] P. Kępka and A. Strzelecki, "Exploring the factors influencing motivation and satisfaction of video game players, " Journal of Electronic Gaming & Esports, vol. 2, no. 1, pp. 1–14, 2024.
[8] J. Hamari, J. Kinnunen, and M. Sjöblom, "Flawless victory! Investigating search and experience qualities as antecedent predictors of video game success, " Electronic Markets, 2023.
[9] C. Bénard, G. Biau, S. da Veiga, and E. Scornet, "Interpretable random forests via rule extraction, " in Proc. 24th Int. Conf. Artificial Intelligence and Statistics (AISTATS), pp. 937–945, 2021.
[10] M. Lechner and G. Okasa, "Random forest estimation of the ordered choice model, " Empirical Economics, vol. 68, pp. 1–106, 2024.