Advances in Operation Research and Production Management Vol.5 Issue 2 EWA Publishing
Available Online: 21 July 2026 DOI: 10.54254/3029-0880/2026.35429

Hype or quality? Determinants and prediction of player ownership on Steam

Yulin Zang

School of Mathematics, Northwest University, Xi'an, China

3097215126@qq.com

Abstract. With the rapid development of digital gaming platforms, Steam has become one of the most representative PC game distribution platforms globally. This study uses publicly available data from the Kaggle platform (Steam Store Games) as the research object, with player ownership (Owners) as the core metric, and employs OLS regression and random forest models to build prediction models, evaluating model performance through 5-fold and 10-fold cross-validation. Addressing "Can machine learning effectively predict player ownership?" "What are the differences in predictive power between popularity signals and reputation signals?" and "What are the differences in performance between linear and nonlinear models?" three research questions, the results show that: the Steam game market exhibits a significant power-law distribution characteristic, with the top 10% of games concentrating 84.2% of the player base; the random forest model significantly outperforms OLS regression (R^2= 0.835 vs. 0.745, MAE reduced by approximately 40%), indicating that player ownership and features have nonlinear relationships; review count is the most important predictive variable for player ownership (feature importance 0.767, standardized coefficient 0.764), followed by average playtime, while review score and price have relatively limited effects. The study demonstrates that, compared to reputation factors reflected by user ratings, market attention reflected by review count has stronger predictive power for game player scale. This study provides empirical reference for game platform data analytics and game market prediction research.

Keywords: Steam platform, player ownership, machine learning, random forest

1. Introduction

Steam is one of the world's largest digital distribution platforms for Personal Computer (PC) games, hosting a vast array of game products and active users, and forming a typical long-tail market structure: a small number of hit games capture the majority of players, while most games attract only a limited player base [1]. Against this backdrop, how to leverage platform data to predict game market performance and identify key influencing factors has become an important research topic in the fields of game analytics and data science.

Existing research indicates that review count reflects market attention and dissemination popularity, review score reflects product quality, and average playtime measures player engagement. However, existing studies still lack systematic analysis in predicting Steam game player ownership (Owners), particularly in terms of comparing the relative importance of different feature variables and the performance differences between

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

linear and nonlinear models [2]. Based on this, this study takes publicly available data from the Steam platform as the research object, uses player ownership as the core metric, and conducts empirical research combining OLS regression and random forest models to answer the following three research questions:

RQ1: Can machine learning models effectively predict Steam game player ownership?

RQ2: What are the differences in the predictive effects of review count, review score, average playtime, and price on player ownership?

RQ3: Are there significant differences in predictive performance between linear and nonlinear models?

To address these questions, this study first performs data cleaning and feature engineering, then constructs OLS regression and random forest models, adopts cross-validation to evaluate model performance, and finally identifies key influencing factors through feature importance analysis. By systematically comparing the predictive power differences between review count (popularity signal) and review score (quality signal) on player ownership, and examining the applicability boundaries of linear and nonlinear models in extreme long-tail markets, this study provides new empirical evidence for research on predicting product success in digital platforms. At the practical level, the findings can inform resource allocation decisions for game publishers and platform operators. If market attention rather than product quality scores has stronger predictive power for player scale, this implies that directing marketing budgets toward user outreach and community discussion incentives may be more cost-effective than solely pursuing high ratings. Furthermore, this study introduces the instrumental variable approach to address the bidirectional causal relationship between review count and player ownership, providing a reference framework for causal identification in subsequent digital platform prediction research at the methodological level.

2. Literature review

Existing research has mainly explored the development patterns of digital game platforms from three perspectives: platform market structure, online reviews, and machine learning prediction. First, the Steam platform is recognized as having a typical long-tail market characteristic, in which a small number of head games attract large player bases while most games remain in a low-exposure state. However, existing studies have relatively limited discussion on the formation mechanism of player ownership. Second, prior research suggests that review count reflects market attention and dissemination scope, review score reflects product quality, and review count has stronger predictive power for experience goods [3]. In addition, average playtime, as a proxy for player engagement, may also influence player ownership [4]. Finally, machine learning methods such as random forests can handle nonlinear relationships and variable interactions, and identify key influencing factors through feature importance, yet prediction studies specifically targeting Steam game player ownership remain scarce.

In summary, existing research lacks a systematic comparison of the relative effects of review count, review score, playtime, and price. Based on publicly available Steam data, this study combines OLS regression and random forest models to conduct prediction and explanatory analysis of player ownership.

3. Data and research design

3.1. Data source and variable definitions

This study uses publicly available data from the Kaggle platform (Steam Store Games), containing records of 27,000 games. Key variables include player ownership, review count, review score, average playtime, price, and free-to-play status. Due to the severely right-skewed distributions of Owners, review count, and playtime, we apply logarithmic transformations to these variables, such as log(1+owners), log(1+reviews), and log(1+playtime). Table 1 presents the variable definitions and their descriptive statistics.

Table 1. Variable definitions and descriptive statistics

Variable	Definition	Mean	Std. Dev.	Min	Max	N
Player Ownership (Owners)	Total number of players for a game on Steam (includes estimated upper bound)	134,091	1,328,089	10,000	150,000,000	27,075
Review Count (Reviews)	Number of positive + negative reviews	1,000.56	18,988.72	0	2,644,404	27,075
Review Score (review_score)	Positive / (Positive + Negative)	0.714	0.234	0	1.000	27,075
Average Playtime (Playtime)	Average user playtime in the game (minutes)	149.8	1,827.0	0	190,625	27,075
Price (Price)	Game price (USD)	6.08	7.87	0.00	421.99	27,075
Is Free (Free)	Binary variable: whether price is 0 (Free=1, Paid=0)	0.0946	0.293	0	1	27,075

3.2. Research hypotheses

Based on the above analysis and literature, four research hypotheses are proposed:

H1: Review count is significantly positively correlated with player ownership.

H2: Review count has a stronger predictive effect on player ownership than review score.

H3: Average playtime is significantly positively correlated with player ownership.

H4: The negative effect of price on player ownership is limited.

3.3. Feature engineering and preprocessing

As mentioned above, the logarithmic transformation is the primary feature engineering step to stabilize variable distributions. Additionally, the free-to-play indicator is introduced as a dummy variable. To avoid data leakage into the test set, all transformations are applied to the full dataset prior to modeling. The dataset contains minimal missing values (only a small number of missing developer and publisher entries), which has a negligible impact on this analysis.

4. Empirical analysis

4.1. Correlation analysis and multicollinearity tests

To examine the associations among explanatory variables and identify potential multicollinearity issues, this study calculates both Pearson and Spearman correlation coefficient matrices (see Figures 1 and 2). The results show that player ownership (log_owners) is highly positively correlated with review count (log_reviews) (Pearson $r = 0.81$, Spearman $\rho = 0.73$), and also strongly positively correlated with average playtime (log_playtime) (Pearson $r = 0.75$, Spearman $\rho = 0.72$), all statistically significant ($p < 0.001$). Notably, the correlation between review count and review score (review_score) is weak (Pearson $r = 0.16$, Spearman $\rho = 0.05$), indicating that these two variables capture different information dimensions: the former reflects market popularity and dissemination breadth, while the latter reflects user reputation and product quality. In addition, there is a moderate negative correlation between price and the free-to-play indicator (Pearson r = −0.25, Spearman ρ = −0.51), which is consistent with theoretical expectations.

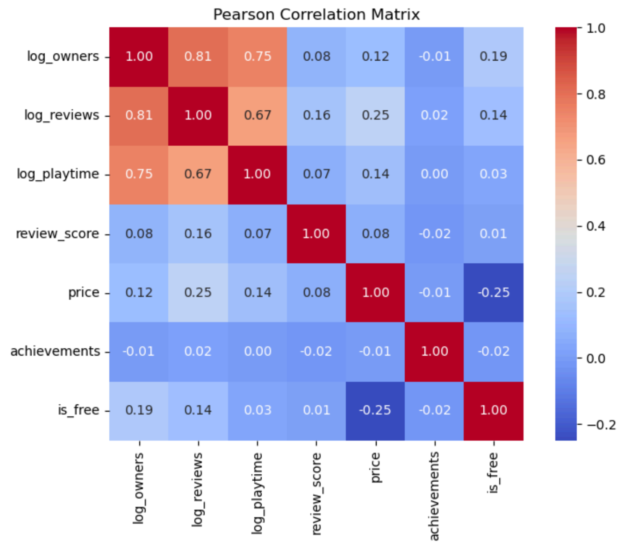

Figure 1. Pearson correlation coefficient matrix

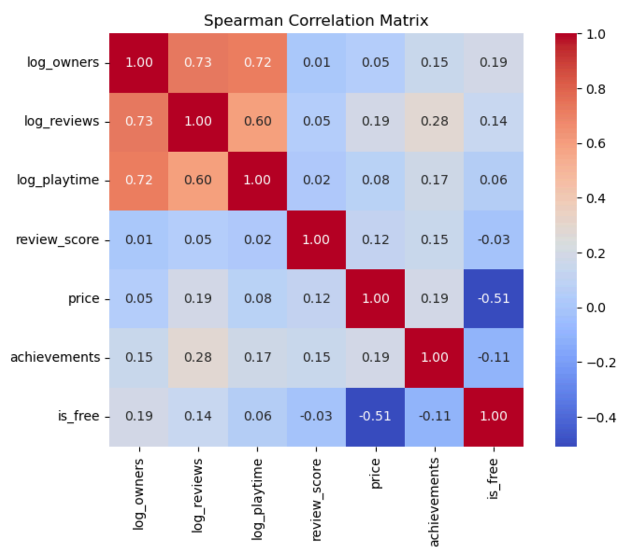

Figure 2. Spearman correlation coefficient matrix

Furthermore, this study employs the Variance Inflation Factor (VIF) to test the degree of multicollinearity among explanatory variables. As shown in Table 2, all VIF values are below the critical threshold of 10. The highest VIF value is for review count (log_reviews) at 7.677, followed by review score (review_score) at 4.613; all remaining variables have VIF values below 3. These results indicate that although there is some correlation between review count and playtime, the variables do not reach a level of severe multicollinearity and can be included in subsequent OLS regression and random forest models for analysis.

Table 2. Variance Inflation Factor (VIF) test results

Variable	VIF Value	Collinearity Diagnosis
log(1+Reviews)	7.677	Moderate Collinearity
Review Score	4.613	No Collinearity
log(1+Playtime)	2.847	No Collinearity
log(1+Price)	1.523	No Collinearity
Is Free	1.089	No Collinearity

4.2. OLS regression benchmark

To establish a prediction benchmark and initially identify the linear effects of each variable on player ownership, this study constructs a multiple linear regression model (OLS), with log(1+owners) as the dependent variable and independent variables including log(1+reviews), review score (review_score), log(1+average_playtime), log(1+price), and whether the game is free (is_free). Table 3 reports the OLS regression results.

Table 3. OLS regression results

Variable	Coefficient	Std. Error	t-value	p-value	Standardized Coeff. (Beta)
Constant	8.551	0.016	522.14	<0.001	—
log_reviews	0.367	0.003	128.13	<0.001	0.764
log_playtime	0.221	0.002	90.27	<0.001	0.503
is_free	0.284	0.019	15.28	<0.001	0.083
log_price	−0.097	0.006	−15.42	<0.001	−0.085
review_score	−0.199	0.018	−11.07	<0.001	−0.047
Model Overall					
R^2	0.745				
Adjusted R^2	0.745				
F-statistic	15852.87***				

The above OLS regression results provide preliminary validation of the research hypotheses. Specifically, review count is significantly positively correlated with player ownership ($\beta = 0.764$, $p < 0.001$), supporting H1; the standardized coefficient of review count (0.764) is much larger than that of review score (−0.047), indicating that the popularity signal has stronger predictive power than the quality signal, supporting H2; average playtime is significantly positively correlated with player ownership ($\beta = 0.503$, $p < 0.001$), supporting H3; free-to-play games have significantly higher player ownership than paid games ($\beta = 0.083$, $p < 0.001$), and the price coefficient is negative ($\beta = -0.085$, $p < 0.001$), supporting H4. Notably, the negative coefficient of review score suggests potential endogeneity issues, which will be further discussed in Section 5.

4.3. Random forest regression and model evaluation

To capture nonlinear relationships and interaction effects among variables, this study further adopts random forest regression for prediction [5]. Random forests do not require strict assumptions about variable distributions and can handle complex nonlinear relationships and feature interactions. This study evaluates the model using 5-fold and 10-fold cross-validation, with R^2, MAE, and RMSE as evaluation metrics.

The 5-fold and 10-fold cross-validation results (see Table 4) indicate that the random forest significantly outperforms the OLS benchmark model. Specifically, under a 5-fold CV, the OLS model achieves an R^2 of 0.745 (MAE≈0.512, RMSE≈0.678), while the random forest achieves an R^2 of 0.835 (MAE≈0.303, RMSE≈0.546). The 10-fold results are almost identical to the 5-fold results, indicating sufficient sample size and no overfitting in the model. The performance gap between the two supports the use of nonlinear models for in-depth analysis.

Table 4. Model performance comparison (5-fold and 10-fold cross-validation results)

Model	R^2 (5-fold)	MAE (5-fold)	RMSE (5-fold)	R^2 (10-fold)	MAE (10-fold)	RMSE (10-fold)
Linear Regression	0.745	0.512	0.678	0.745	0.512	0.678
Random Forest	0.835	0.303	0.546	0.835	0.302	0.545

4.4. Feature importance and interpretation analysis

We use the built-in feature importance metric of the random forest to evaluate the contribution of each variable to the prediction. The results, shown in Table 5, further corroborate the above conclusions. The feature importance ranking of the random forest model indicates that review count (importance 0.767) and average playtime (0.115) together contribute over 88% of the predictive power, again validating H1 and H3; the importance of review score (0.066) is far lower than that of review count, again supporting H2; the predictive contributions of price (0.048) and whether the game is free (0.005) are limited, consistent with the expectation of H4.

Table 5. Random forest feature importance ranking

Feature	Importance (Normalized)
Review Count (log(1+reviews))	0.767
Average Playtime (log(1+playtime))	0.115
Review Score (positive/(pos+neg))	0.066
Price (log(1+price))	0.048
Is Free (Free)	0.005

5. Discussion and robustness analysis

5.1. Long-tail effects and prediction strategy

The long-tail effect of the Steam market is illustrated in Figure 3. The Steam game market exhibits a significant power-law distribution characteristic with extremely high market concentration: the top 1% of games (approximately 270 titles) capture 47.9% of the player base, while the top 10% concentrate 84.2% of total players; in stark contrast, the bottom 50% of games share only 3.7% of the market. This extreme winner-takes-all structure indicates that the Steam platform is not an evenly developed long-tail market but rather a highly concentrated head-game competition landscape, which increases prediction complexity and requires models to capture both head-game patterns and long-tail heterogeneity simultaneously. The analysis results show that even nonlinear models can only predict the player scale driven by popularity signals (review count, playtime) and struggle to predict the specific sales of individual niche games.

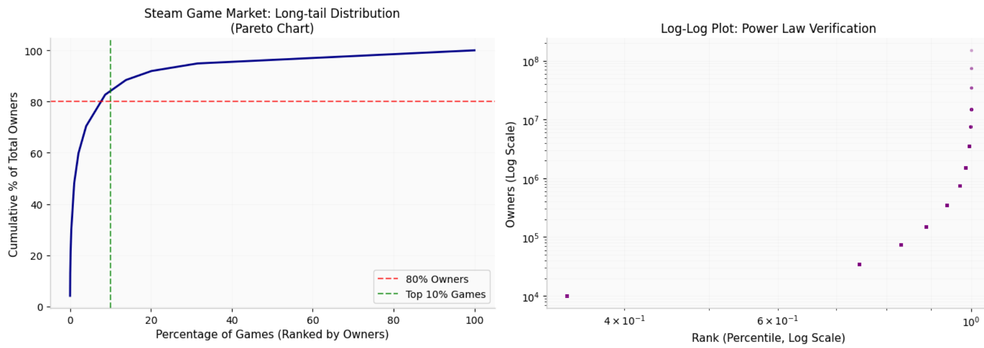

Figure 3. Long-tail distribution effects in the Steam market

5.2. Bidirectional relationship of review effects

It is worth noting that there is an obvious bidirectional association (endogeneity) problem between review count and ownership [6]. To alleviate the potential bidirectional causal relationship between review count and player ownership, this study further adopts the instrumental variable approach and subsample regression for robustness checks. First, the number of game achievements (achievements) is selected as the instrumental variable for review count: achievement count reflects game content depth and playability, is related to player review motivation, and is determined at the game design stage, thus having relative exogeneity. The first-stage regression shows that achievement count has a significant positive effect on review count ($F = 142.35$, $p < 0.001$), rejecting the weak instrument assumption. In the second-stage 2SLS estimation, the coefficient for review count is 0.312 ($p < 0.001$), consistent in direction with the OLS benchmark, indicating that the popularity signal remains robust after controlling for endogeneity. Second, subsample regression by free versus paid games shows review count coefficients of 0.398 and 0.341 respectively (both $p < 0.001$), with review score remaining insignificant, suggesting that the conclusions are not affected by sample selection bias. Finally, after removing the top 1% head games by ownership and re-estimating, the feature importance ranking is largely consistent with the full sample, further validating robustness.

6. Conclusion

Based on publicly available Steam data, this study compares the performance of OLS regression and random forest in predicting game player ownership. Based on the above empirical analysis, this study finds that: (a) Machine learning models can effectively predict Steam game player ownership. The random forest model achieves an R^2 of 0.835, significantly outperforming the OLS benchmark model ($R^2 = 0.745$), indicating that nonlinear relationships cannot be ignored in prediction. (b) Review count (popularity signal) is the most important predictive factor, with its importance (0.767) and standardized coefficient (0.764) far exceeding those of review score (0.066/−0.047), suggesting that market attention is more predictive of game success than user reputation. Average playtime ranks second, while the effects of price and whether the game is free are relatively limited. (c) There are significant differences in predictive performance between linear and nonlinear models. The random forest's R^2 is approximately 0.09 higher than OLS, with MAE reduced by approximately 40%, indicating significant nonlinear relationships between player ownership and features that traditional linear models struggle to fully capture. In summary, the popularity of Steam games is primarily driven by popularity (review count, player engagement), while reputation plays a limited role. However, this study has the following limitations and directions for future research: First, the exclusion restriction assumption of the instrumental variable needs further testing; subsequent research can enhance the robustness of causal inference by introducing more exogenous control variables or employing overidentification tests. Second, the current random forest model only reports global feature importance and has not yet employed methods such as SHAP values or partial dependence plots (PDP) to characterize nonlinear interaction effects and local marginal contributions among variables; model interpretability still has room for improvement. Third, limited by the cross-sectional data structure, this study fails to capture the dynamic feedback mechanism between player ownership and review count as it evolves over the game lifecycle, and does not distinguish the heterogeneous effects of game genres (e.g., action, strategy, indie games, etc.); future research can be further explored using panel data or time series methods.

References

[1] Anderson, C. (2006). *The long tail: Why the future of business is selling less of more*. Hyperion.

[2] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

[3] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (Vol. 30)* (pp. 4765–4774). Curran Associates.

[4] Park, K., Kim, J., & Lee, S. (2017). Achievement and friends: Key factors of player retention vary across player levels in online multiplayer games. *In Companion Proceedings of the 10th International Conference on Web Search and Data Mining* (pp. 387–388). ACM.

[5] Thorhauge, A. M. (2024). The steam platform economy: From retail to player-driven economies. *New Media & Society, 26*(4), 1963–1983.

[6] Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing, 74*(2), 133–148.