Proceedings of the 3rd International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2025.DL25672

Analysis of Factors Influencing Board Game Ownership Based on A Gradient Boosting Model

Thomas Xiao

Irvine Valley College, Arnold O. Beckman High School, Irvine, USA
sxiao7@ivc.edu

Abstract. This study investigates the factors influencing board game ownership using a comprehensive dataset from BoardGameGeek. By applying statistical techniques such as correlation analysis, Lasso regression, Random Forest, and a Gradient Boosting Model (GBM), the paper explores how variables such as user ratings, complexity, playtime, age rating, and year of publication affect ownership. The analysis reveals that user engagement metrics—particularly the number of users who rated a game—are the strongest predictors of ownership. However, when visibility metrics are excluded, intrinsic attributes like game complexity, playtime, and accessibility (minimum age) emerge as significant drivers. The Gradient Boosting Model, trained with log-scaled ownership values, achieves a high score of 0.975 on the validation set, confirming its strong predictive performance. These findings provide actionable insights for developers, highlighting the importance of both marketing efforts and thoughtful design choices. The study contributes to the understanding of consumer behavior in the board game industry and offers a data-driven framework for optimizing game appeal and commercial success.

Keywords: Board games, ownership analysis, complexity, predictive modeling, gradient boosting model.

1. Introduction

A board game consists of any game played on a board, typically one that involves the movement and placing of pieces in different positions to achieve a strategic advantage. Classic examples include chess, checkers, and Monopoly. In recent years, board games have experienced a revival, with players rating and reviewing games on platforms like BoardGameGeek (BGG) [1], especially during times such as COVID, in which people sought at-home entertainment and social interaction. All of this culminated in the world board game market to be valued at $20 billion, a number that is expected to nearly double by 2028 [2]. As the market grows, understanding what drives ownership can help developers optimize game design for broader appeal. Researchers have explored various factors influencing board game ownership and purchase intentions. For example, Kosa and Spronck [3] examined the purchase intentions for modern board games, finding that enjoyment, positive word of mouth, age, and gender were positively associated with purchase decisions, while factors like income, play frequency, prior board gaming experience, and feelings of presence had no significant impact. Building on this, d'Astous [4] concluded that the success of a board game relies

DOI: 10.54254/2753-8818/2025.DL25672

on its ability to deliver a unique and immersive play experience while fostering player interaction. Beyond consumer preferences, market structures also shape ownership trends. Crowdfunding has emerged as a powerful force in board game publishing, especially on platforms like Kickstarter. However, Wachs and Vedres [5] argue that crowdfunding tends to favor projects with incremental innovation rather than those with radical new ideas, meaning that highly novel games may not always gain traction or ownership despite their originality. This finding highlights the importance of market visibility and user familiarity in driving ownership metrics. Additionally, the economics of board game production can influence both pricing and ownership. Bohnsack and Supakkeittikul [6] show that complexity is a major cost driver in board game design and manufacturing. As complexity increases, so do production and retail prices, potentially affecting accessibility and mass-market adoption. This makes the evaluation of complexity a crucial dimension in predicting ownership, particularly in distinguishing between niche, high-involvement games and broadly accessible family games. Likewise, I analyze factors more specific to the board games themselves, such as game complexity and player count. The significance of this research lies in its potential to provide actionable insights for board game designers, publishers, and marketers seeking to appeal to a broader audience in a rapidly growing industry. By identifying which intrinsic game features—such as average playtime, player count, and complexity—correlate most strongly with ownership rates and popularity, developers can make more informed design decisions. This research also contributes to the academic understanding of consumer behavior in the board game sector, a relatively underexplored area. Ultimately, the goal is to bridge quantitative game data with player preferences to better predict commercial success and consumer appeal.

2. Methodology

2.1. Dataset overview

The dataset, sourced from Kaggle [7], contains over 20,000 entries with attributes such as year of publication, player count, playtime, age rating, user ratings, game complexity, and the number of owned users. These features provide a comprehensive basis for analyzing the factors influencing board game ownership, with the number of owned users serving as the target variable.

2.2. Dataset variables

The dataset includes many variables, of which the relevant ones for this paper are:

2.3. Data cleaning and preprocessing

Data cleaning involves handling missing values and ensuring consistency. Missing values in the Year Published column were filled with the median year, while other columns were converted to appropriate data types. Additionally, two separate analyses were conducted: one including rating-related variables (Users Rated, Rating Average, and BGG Ranking) and another excluding them. This approach allows us to evaluate the impact of these variables while also exploring the influence of intrinsic game attributes.

3. Statistical methods

Firstly, correlation heatmaps visually display the strength of relationships between variables. They help identify key interactions, such as the strong correlation between Users Rated and Owned Users,

DOI: 10.54254/2753-8818/2025.DL25672

which highlights the impact of user engagement on game ownership. Secondly, random Forest regression is an ensemble learning method that builds multiple decision trees to predict outcomes. It identifies influential variables even in the presence of nonlinear relationships. Finally, lasso regression applies a penalty term to reduce the impact of less important variables, performing feature selection by setting some coefficients to zero. This method highlights influential variables while minimizing noise from less significant factors.

4. Analysis and results

4.1. Correlation heatmaps

Figure 1 demonstrates a strong positive correlation (0.99) between Users Rated and Owned Users, indicating that visibility and user engagement are extremely critical drivers of ownership. In contrast, Rating Average shows a weaker correlation (0.18), suggesting that game quality actually plays a secondary role compared to visibility under a purchasing context. Interestingly, BGG Ranking exhibits a moderate negative correlation (-0.33) with Owned Users, meaning that games with better (lower on a first to last numerical scale) rankings tend to have higher ownership. However, the correlation is not strong, indicating a difference in opinion between critics and the average consumer. This underscores the impact of perceived reputation and popularity on user acquisition. Beyond Owned Users, other high correlations reveal important dynamics. For instance, Rating Average has a strong negative correlation (-0.74) with BGG Ranking, indicating that games with higher average rating tend to achieve better rankings, likely due to increased visibility and community engagement. In the context of gameplay elements, Complexity Average shows a moderate positive correlation (0.48) with Rating Average, suggesting that more complex games tend to have higher average ratings, while also showing a moderate negative correlation (-0.38) with BGG Rank, indicating a similar conclusion of a preference for more complex games.

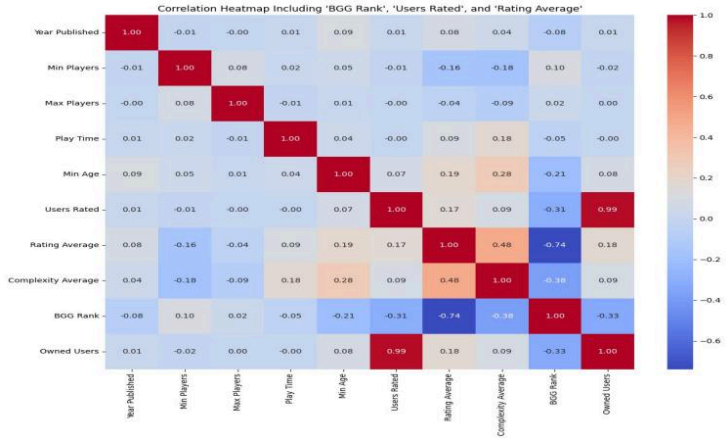

Figure 1. Correlation heatmap including rating variables

When rating-related variables are excluded, as shown in Figure 2, Complexity Average emerges as the most correlated intrinsic feature (0.28). This indicates that while complexity contributes to ownership, its effect is relatively minor compared to engagement-driven variables.

DOI: 10.54254/2753-8818/2025.DL25672

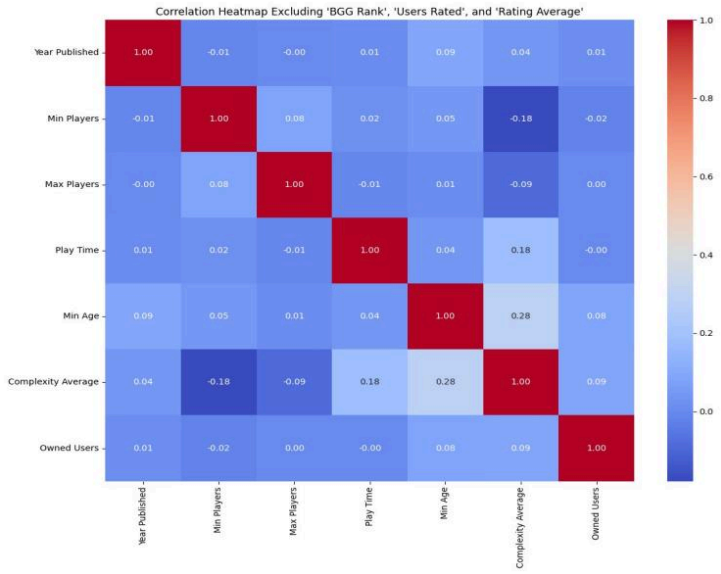

Figure 2. Correlation heatmap excluding rating variables

4.2. Lasso regression results

The Lasso regression coefficients reaffirm the dominance of Users Rated as a predictor of ownership, underscoring the critical role of visibility and user engagement. Including rating-related variables (Figure 3) highlights the overwhelming importance of visibility metrics such as Users Rated. Specifically, Users Rated exhibits the largest coefficient, indicating its direct influence on ownership, while BGG Ranking—though inversely correlated with Owned Users—also plays a role in determining visibility and perceived popularity. This finding emphasizes the need for developers to actively engage with online communities to enhance a game's visibility and encourage rating frequency. Excluding rating-related variables (Figure 4) shifts the focus to intrinsic game attributes. Here, Complexity Average and Minimum Age emerge as the most influential predictors. Complexity Average, which reflects the strategic depth of a game, positively correlates with ownership, particularly for more dedicated audiences who value engaging gameplay experiences. However, the significance of Minimum Age implies that games accessible to younger audiences may have broader appeal, possibly due to their suitability for families or casual gamers. Additionally, Play Time also shows a moderate influence, suggesting that games with shorter or moderate playtimes are more likely to attract a wider range of players, as excessively long games may deter casual audiences. This shift in significant predictors highlights a key insight: while visibility metrics dominate when included, intrinsic game attributes like complexity, accessibility, and playtime become critical in their absence. This implies that for developers, increasing visibility should be a top priority; however, crafting games with an appropriate balance of complexity, accessibility, and playtime is also essential to ensure appeal and long-term retention across different demographics.

DOI: 10.54254/2753-8818/2025.DL25672

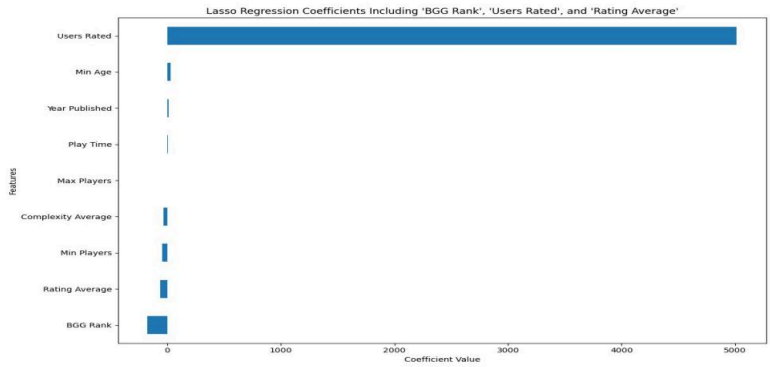

Figure 3. Lasso regression coefficients including rating variables

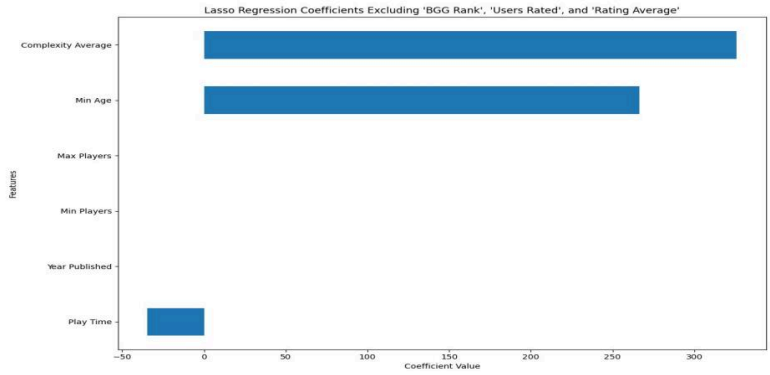

Figure 4. Lasso regression coefficients excluding rating variables

4.3. Random forest feature importance

Random Forest analysis further underscores the overwhelming importance of Users Rated, with an importance score near 1 (Figure 5). This reaffirms that visibility, as captured by the number of user ratings on platforms like BoardGameGeek, is the primary driver of ownership. Users Rated serves as a direct proxy for exposure and community engagement, suggesting that games with higher levels of interaction and user feedback gain greater traction among audiences. Noticeably, BGG Rank is nowhere near considered important. When rating-related variables are excluded (Figure 6), intrinsic features such as Complexity Average and Year Published emerge as significant predictors of ownership. Complexity Average is particularly noteworthy, highlighting that games with strategic depth and engaging mechanics appeal strongly to dedicated gamers who are willing to invest time and effort into understanding more complex gameplay. This aligns with previous insights, where a balance between complexity and accessibility is vital for maximizing appeal across different audience demographics. Interestingly, Year Published also emerges as a key predictor in the absence of rating-related variables. This suggests that modern games benefit from contemporary design trends, innovative mechanics, and active marketing campaigns, which resonate with players seeking fresh and up-to-date gaming experiences. However, the importance of Year Published also indicates that older games with sustained appeal and strong word-of-mouth may still hold a considerable audience base, particularly if they are considered classics, especially with older games such as chess and checkers. Play Time shows moderate importance, reinforcing its role as a secondary yet relevant factor in influencing ownership. Games with shorter or moderate playtimes are generally more accessible to casual players, while excessively long playtimes may cater more specifically to niche

DOI: 10.54254/2753-8818/2025.DL25672

audiences of dedicated gamers. This highlights the need for developers to strike a balance in playtime duration to ensure broad appeal while retaining depth. Overall, the Random Forest analysis provides nuanced insights into the interplay between visibility metrics and intrinsic attributes. While visibility metrics dominate ownership predictions when included, intrinsic factors such as complexity, recency, and accessibility become critical when evaluating games in isolation.

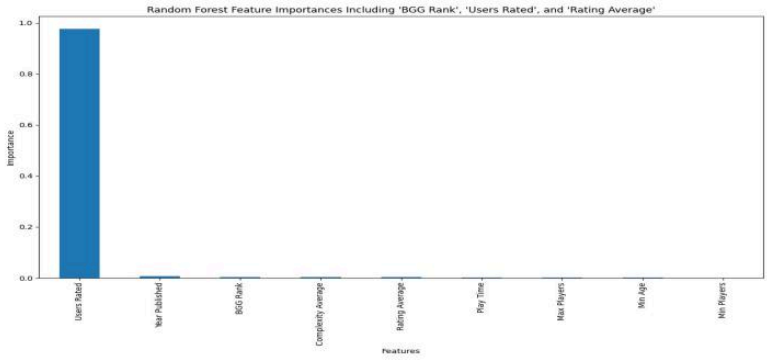

Figure 5. Feature importance (random forest including rating variables)

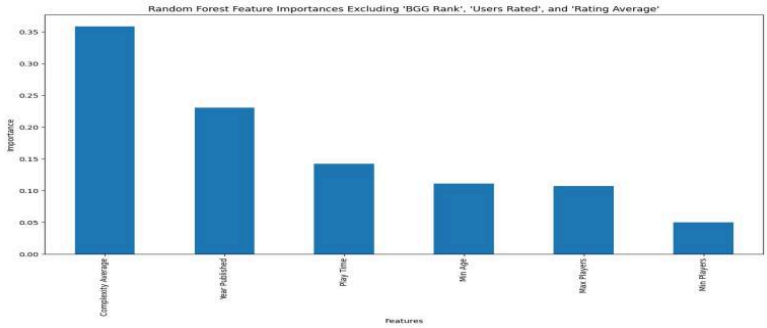

Figure 6. Feature importance (random forest excluding rating variables)

5. Predictive model analysis

5.1. Gradient boosting model development

A Gradient Boosting Model (GBM) was developed to predict the logarithmically scaled number of owned users based on Complexity Average, Play Time, Users Rated, Minimum Age, and Year Published. These variables were selected to capture both intrinsic attributes and engagement-driven metrics. The dataset was split into training and validation sets, with 80% used to train and 20% used to test, and log scaling was applied to the target variable to better account for the wide range of ownership values.

5.2. Gradient boosting model function

The predicted values in the Gradient Boosting Model are computed by combining the outputs of a series of decision trees. The general function used for the predictions is given by:

$$\hat{y} = F_M(x) = \sum_{m=1}^{M} \nu \cdot h_m(x) \tag{1}$$

Where:

DOI: 10.54254/2753-8818/2025.DL25672

$\hat{y}$: The predicted value for a given input x computed in the log-transformed space

$F_M(x)$: The overall prediction after M boosting iterations (the combined result of all trees)

M : The total number of trees (iterations) used in the model

ν : The learning rate, a scaling factor that controls the contribution of each tree to the overall prediction

$h_m(x)$: The prediction made by the m -th decision tree for the input x

5.2.1. The process

Initialization: The model initializes the predicted values to a constant value, typically the mean of the target variable $\overline{y}$ in the log-transformed space:

$$F_0(x) = \log(\overline{y}) \tag{2}$$

Iterative Improvement: At each iteration m the model:

Fits a decision tree $h_m(x)$ to the negative gradient of the loss function (the residuals from the previous iteration). This captures the direction in which predictions should be adjusted to minimize error

Updates the overall prediction by adding the scaled output of the tree:

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x) \tag{3}$$

Final Prediction: After M iterations, the final prediction is the sum of the initial prediction and the weighted outputs of all the trees:

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \nu \cdot h_m(x) \tag{4}$$

The final prediction is then transformed back from the logarithmic space using:

$$\widehat{y_{final}} = \exp(F_M(x)) \tag{5}$$

5.2.2. Application to ownership prediction

In this study, the Gradient Boosting Model predicts the number of owned users $\hat{y}$ using the following input features. They are Complexity Average, Play Time, Users Rated, Minimum Age,Year Published

The model minimizes the loss function (e.g., Mean Squared Error) by iteratively computing residuals and fitting decision trees. The final prediction for ownership is the sum of the contributions from all the trees, scaled by the learning rate ν. The results are shown in the table below:

Table 1. Performance metrics of the gradient boosting model with log scaling

Metric	Training Set	Validation Set
MAE	199.64	241.02
RMSE	532.80	765.20
R^2 Score	0.989	0.975

DOI: 10.54254/2753-8818/2025.DL25672

According to Table 1, Mean Absolute Error (MAE) measures the average absolute difference between the predicted and actual values. For the training set, the MAE is 199.64, indicating the average error in predicting ownership is approximately 200 users. The slightly higher MAE of 241.02 on the validation set suggests good generalization with minimal overfitting. Looking to Root Mean Squared Error (RMSE), the RMSE values of 532.80 (training) and 765.20 (validation) reflect the complexity of real-world ownership patterns and the model's ability to capture them. The R^2 scores of 0.989 (training) and 0.975 (validation) indicate that the model explains most of the variance in ownership, showcasing strong predictive capability.

5.3. Predicted vs actual values

To validate the model's accuracy, predicted ownership values were compared to actual values from the dataset. The close alignment between predicted and actual results demonstrates the reliability of the GBM in capturing relevant ownership trends.

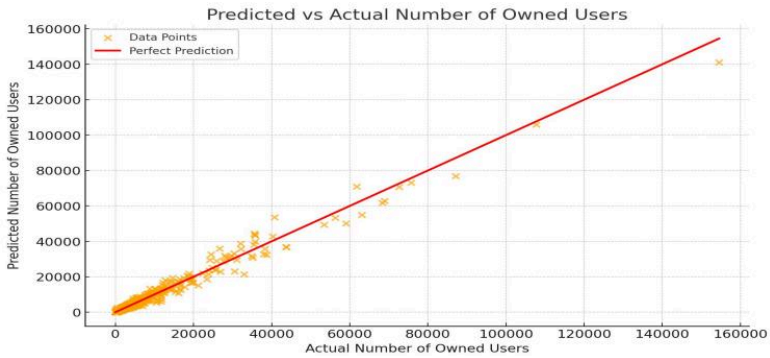

Figure 7. Predicted vs actual number of owned users (log scaling applied)

Discrepancy for Large Values: While the log scaling significantly improves model accuracy and alignment for most data points, discrepancies may still exist for games with extremely large ownership numbers. This is due to the inherent skew in the dataset and the limitations of the log transformation in fully capturing extreme outliers. Developers should interpret predictions for such cases with caution and consider further analysis. Log scaling offers the advantage of reducing skew and improving model performance on typical data ranges, but it can underrepresent extreme values. To improve accuracy for high-ownership games, future work could explore piecewise regression, quantile-based methods, or ensemble models that better capture outlier behavior.

6. Conclusion

This study provides a comprehensive analysis of the factors influencing board game ownership by combining statistical methods, optimization frameworks, and predictive modeling. Through the use of correlation heatmaps, Lasso regression, and Random Forest analysis, the study identified key drivers of ownership, including user engagement metrics (such as Users Rated) and intrinsic game attributes (such as Complexity Average and Play Time). The findings highlight the critical role of visibility in driving ownership, while also emphasizing the importance of balancing gameplay features to appeal to diverse audiences. A Gradient Boosting Model (GBM) was employed, leveraging historical data to predict ownership with high accuracy. The GBM successfully captured non-linear relationships and feature interactions, demonstrating strong predictive capabilities (R^2 = 0.975 on the validation set) and providing actionable insights for game design. The implications of

DOI: 10.54254/2753-8818/2025.DL25672

this study are significant for developers. By prioritizing visibility and user engagement through strategies such as community-driven marketing and user reviews, developers can enhance the appeal of their games. Additionally, the insights into optimal feature ranges, such as maintaining moderate complexity and playtime, enable developers to design games that balance accessibility and depth. The use of advanced modeling techniques, such as GBMs, further empowers developers to test hypothetical designs, optimize feature interactions, and tailor games to specific market segments. However, this study is limited by its reliance on publicly available data from BoardGameGeek, which may reflect selection bias toward more engaged users. Additionally, unmeasured variables such as marketing spend, artwork quality, or brand recognition could also influence ownership but were not included in the analysis. In conclusion, this study bridges the gap between theoretical optimization and practical predictive modeling, offering a robust framework for understanding and maximizing board game ownership. Future research could expand on this work by incorporating additional variables, such as production quality or marketing budget, and exploring new methodologies to refine predictions and enhance game design strategies.

References

[1] Zachary Horton, The Rise of Board Games in Today's Tech-Dominated Culture, Pittwire, University of Pittsburgh, April 10, 2020. Available at: https: //www.pitt.edu/pittwire/features-articles/riseboard-games-today-s-tech-dominated-culture.

[2] John Doe, "Board games are making a comeback. Here's why, " Oak Park Talon, March 10, 2023. Available at: https: //oakparktalon.org/16817/feature/ board-games-are-making-a-comeback-heres-why/.

[3] Kosa, M., and Spronck, P., "An Exploratory Study on the Purchase Intentions of Modern Board Games, " ResearchGate, November 2022. Available at: https: //www.researchgate.net/ publication/365138247_An_Exploratory_Study_on_the_Purchase_Intentions_of_Modern_ Board_Games_Purchase_Intentions_of_Modern_Board_Games.

[4] d'Astous, A., "An Inquiry into the Factors that Impact on Consumer Appreciation of a Board Game, " ResearchGate, January 2012. Available at: https: //www.researchgate.net/publication/ 235265445_An_inquiry_into_the_factors_that_impact_on_consumer_appreciation_of_a_ board_game.

[5] Wachs, J., & Vedres, B., "Does Crowdfunding Really Foster Innovation? Evidence from the Board Game Industry, " arXiv, January 7, 2021. Available at: https: //www.sciencedirect.com/science/ article/pii/S0040162521001797. .

[6] Bohnsack, B., & Supakkeittikul, P., "Why are board and card games so expensive? Complexity is a major factor, " SSRN, October 25, 2024. Available at: https: //ssrn.com/abstract=4999711.

[7] Dilini Samarasinghe, "BoardGameGeek Dataset on Board Games, " IEEE Dataport, July 5, 2021. doi: https: //dx.doi.org/10.21227/9g61-bs59.