Predicting the Popularity Level of Roblox Games Using Gameplay and Metadata Features with Machine Learning Models

Ding Yi[1,*], Luo Jun[2], S Govindaraju[3]

[1,2]*Faculty of Data Science and Information Technology (FDSIT), INTI International University, Nilai, Malaysia*

[3]*PG and Research Department of Computer Science, Sri Ramakrishna College of Arts and Science, Coimbatore-641006, Tamil Nadu, India*

(Received: October 5, 2024; Revised: November 25, 2024; Accepted: January 10, 2025; Available online: April 1, 2025)

Abstract

The online gaming platform Roblox has become a significant player in the gaming industry, providing a space for user-generated content. Predicting the popularity of Roblox games can help developers design better games and optimize user engagement. This study explores the use of machine learning models to predict the popularity of games on Roblox using gameplay features and metadata. A dataset of 9,734 games was collected, including variables such as likes, visits, game age, and active players. Three machine learning models, Decision Tree, Random Forest, and Gradient Boosting were employed to predict the number of favorites, which serves as a proxy for game popularity. Among the models tested, Gradient Boosting outperformed the others, achieving the highest R-squared score (0.85) and the lowest Root Mean Squared Error (11,470). Key features such as likes, game age, and visits were identified as the most influential in predicting game popularity. Based on these findings, this study recommends that developers focus on features that increase player engagement, such as regular updates and optimizing game exposure. Additionally, incorporating additional data sources, such as user reviews, and exploring explainability methods like SHAP can further improve model accuracy and transparency. This research contributes valuable insights into how machine learning can support decision-making in the development and optimization of Roblox games.

Keywords: Roblox, Game Popularity, Machine Learning, Gradient Boosting, Predictive Modeling

1. Introduction

The online gaming industry has experienced significant expansion over the past decade, becoming one of the most dynamic and profitable sectors within digital entertainment. In 2019 alone, the global online gaming market generated approximately $152.1 billion in revenue, a figure that has continued to grow in subsequent years [1]. This surge in growth can be largely attributed to the rise of interactive and user-generated platforms like Roblox, which empower users to both create and engage in immersive virtual experiences.

Roblox, in particular, has revolutionized the landscape of online gaming by offering a multifaceted ecosystem where creativity, social interaction, and entertainment converge. As a user-generated content platform, Roblox not only allows players to play games but also to design and publish their own, thereby fostering a unique creator economy. This participatory model has attracted millions of users globally and turned the platform into a cultural and economic phenomenon.

However, the rise of online gaming also presents both opportunities and challenges. On the positive side, online gaming has been linked to the development of interpersonal and cognitive skills. Games that emphasize collaboration, problem-solving, and strategic thinking can enhance a player's ability to work in teams, think critically, and adapt to new information [2]. These skills are especially beneficial for younger players who are in critical stages of their cognitive development.

Conversely, concerns have been raised about the potentially detrimental effects of excessive gaming. Studies have indicated that prolonged engagement with online games can lead to addiction, particularly among adolescents. This

***Corresponding author: Ding Yi (yi.ding@newinti.edu.my)**
DOI: https://doi.org/10.47738/ijaim.v5i1.97

form of dependency may result in declining academic performance, poor social relationships, and adverse mental health outcomes [3]. The dual nature of online gaming thus necessitates a balanced approach that maximizes its benefits while mitigating its risks.

To address these issues, researchers have proposed various preventive strategies. These include structuring daily routines to limit gaming time, encouraging participation in extracurricular activities, and promoting awareness about healthy gaming habits [4]. Moreover, collaborative efforts between educational institutions and gaming platforms like Roblox are gaining attention as a way to harness the educational potential of games. Such partnerships can facilitate the integration of games into learning environments, thereby fostering cooperative literacy and enhancing student engagement [5].

Within this rapidly evolving landscape, understanding the determinants of game popularity has become crucial for developers aiming to create successful and impactful games. Popularity in gaming can be influenced by a range of factors including game mechanics, social features, and marketing strategies. Games that offer immersive experiences, clear progression systems, and opportunities for social interaction tend to perform better in attracting and retaining players [6].

In addition to gameplay mechanics, metadata associated with games such as user ratings, number of visits, and active users serves as a valuable source of information. These data points offer insights into user behavior and preferences, enabling developers to refine their designs and strategies. Furthermore, the integration of social media and esports into marketing campaigns has been shown to significantly enhance game visibility and user engagement [7].

The availability of diverse and rich data sources presents an unprecedented opportunity for leveraging machine learning techniques to predict game popularity. Machine learning models are particularly adept at identifying complex patterns and relationships within large datasets that may not be immediately apparent through traditional analytical methods [8]. In the context of gaming, these models can analyze a combination of gameplay features and metadata to predict metrics such as user engagement, retention, and popularity.

However, the application of machine learning in this domain also raises ethical considerations. Predictive analytics have the potential to influence user behavior in ways that may be perceived as intrusive or manipulative. Therefore, it is essential to implement these technologies transparently and responsibly, ensuring that they serve the interests of users and stakeholders alike [9].

To develop effective predictive models, a comprehensive approach is required. This includes careful feature selection, algorithm tuning, and model evaluation using robust performance metrics. Ensemble learning methods, such as Random Forest and Gradient Boosting, have emerged as particularly effective in handling complex, high-dimensional data. These techniques combine multiple weak learners to produce a strong predictive model, often outperforming single-model approaches [10].

The application of ensemble learning in predicting game popularity has been supported by various studies across domains. In fields such as marketing, finance, and healthcare, these models have demonstrated superior performance in forecasting outcomes and uncovering critical predictors [11]. In the context of Roblox games, leveraging ensemble models alongside gameplay and metadata features holds great promise for accurately estimating popularity metrics.

In conclusion, the growth of the online gaming industry, the rich availability of gameplay and user interaction data, and the advancement of machine learning technologies converge to create an opportune moment for predictive analytics in gaming. By understanding what drives the popularity of Roblox games, developers can make informed decisions that enhance user experience, increase engagement, and ultimately contribute to the commercial and educational potential of their games. This research aims to explore and evaluate the use of machine learning models specifically Decision Tree, Random Forest, and Gradient Boosting to predict Roblox game popularity based on gameplay and metadata features, thereby contributing valuable insights to both academic literature and industry practice.

2. Literature Review

2.1. Roblox as a User-Generated Gaming Platform and Its Educational Applications

Roblox has emerged as one of the largest user-generated online gaming platforms, attracting millions of users globally, particularly among children and teenagers [12]. Its unique architecture allows users not only to engage in gameplay but also to design, create, and publish their own games, turning Roblox into both a gaming platform and a creative ecosystem [13]. This dual functionality has sparked significant interest from educators and researchers, who see Roblox as a valuable educational tool.

Several studies have examined the use of Roblox in diverse educational contexts. Researchers have utilized Roblox for teaching subjects such as science [13], art [14], and engineering [15]. Additionally, innovative projects like tsunami survival games have demonstrated Roblox's potential to foster disaster preparedness skills, even in early childhood education. Roblox has also been applied to vocational education, offering interactive environments for developing practical skills [16]. These studies collectively highlight Roblox's versatility in promoting experiential learning and fostering critical thinking, creativity, and collaboration among learners.

2.2. Challenges and Ethical Issues in Roblox Platform

While Roblox offers numerous educational and creative opportunities, it also presents significant ethical and safety challenges. Due to the platform's user-generated nature, concerns have arisen over the presence of inappropriate content that may bypass moderation filters, including sexually explicit material and harmful user-generated games [17]. These issues raise serious ethical concerns related to child safety, platform governance, and content regulation. The ease of content creation on Roblox, while empowering for users, simultaneously necessitates more robust oversight to protect vulnerable audiences from exposure to harmful or malicious content. Continued research and policy development are essential to ensure that Roblox remains a safe environment for its predominantly young user base.

2.3. Defining and Measuring Game Popularity in Roblox

The concept of game popularity on Roblox can be operationalized through key platform metrics such as likes, visits, and favorites [18], [19], [20]. Each of these indicators offers a distinct measure of user engagement. Likes reflect users' approval and enjoyment of a game, visits capture the game's exposure and reach within the Roblox community, while favorites demonstrate a stronger level of commitment and interest, as players save games for future access.

Understanding these popularity metrics enables developers and researchers to analyze how various factors contribute to a game’s success, including design quality, gameplay mechanics, social features, and overall user experience [21], [22]. By studying these metrics, developers gain insights into user preferences, which can inform design decisions to improve engagement, satisfaction, and long-term game success.

2.4. Machine Learning Techniques for Predicting Content Popularity

Numerous studies have explored the application of machine learning techniques for predicting content popularity, not only in gaming but also in videos and mobile applications. In the gaming domain, machine learning models such as Random Forest, Gradient Boosting, and deep learning algorithms have been used to predict game popularity based on user behavior, gameplay mechanics, and marketing strategies [6], [2], [23]. Features such as immersion, progression, and social interaction have been identified as crucial determinants of player engagement.

Similar machine learning methods have been applied in video content popularity prediction. Studies have combined multimodal features such as text, images, and user interaction data using advanced models like recurrent neural networks to improve prediction accuracy [24], [25]. In mobile app contexts, researchers have used user reviews, app metadata, and social media engagement to forecast app popularity, with ensemble models such as stacking and bagging often outperforming individual models [26], [27].

Additionally, researchers have examined how factors such as social influence, homophily, and network dynamics contribute to predicting content popularity [28]. Privacy preserving approaches like federated learning and collaborative filtering have also been explored to address data sparsity and shifting popularity trends while maintaining user data privacy [29].

Among the many machine learning models applied in these contexts, three algorithms are particularly noteworthy for their relevance to predicting Roblox game popularity: Decision Tree, Random Forest, and Gradient Boosting. Decision Trees create hierarchical decision rules based on feature values, offering models that are easy to interpret but susceptible to overfitting on complex datasets [30]. Random Forest improves upon this by generating multiple trees using bootstrapped data and aggregating their predictions, providing robustness against overfitting and strong performance on high dimensional data. Gradient Boosting further enhances predictive power by iteratively correcting the errors of previous models, enabling the algorithm to capture complex, nonlinear relationships [25]. Well known implementations such as XGBoost and LightGBM have repeatedly demonstrated superior predictive accuracy across numerous applications.

Overall, these studies establish a solid foundation for using machine learning to predict content popularity. In the context of Roblox, integrating gameplay features, user engagement metrics, and metadata with machine learning techniques presents a promising approach to developing predictive models that can assist developers, educators, and platform managers in making data-driven decisions that enhance user experience and game success.

3. Methodology

Figure 1 illustrates the overall workflow of the predictive modeling pipeline. It starts with data collection, followed by preprocessing, including one-hot encoding, cleaning and converting numeric fields, and feature engineering. Machine learning models, such as Decision Tree, Random Forest, and Gradient Boosting, are then applied. The data is split into training and testing sets (80:20), and model performance is evaluated using metrics such as Mean MAE), Root Mean Squared Error (RMSE), and R-squared (R^2).

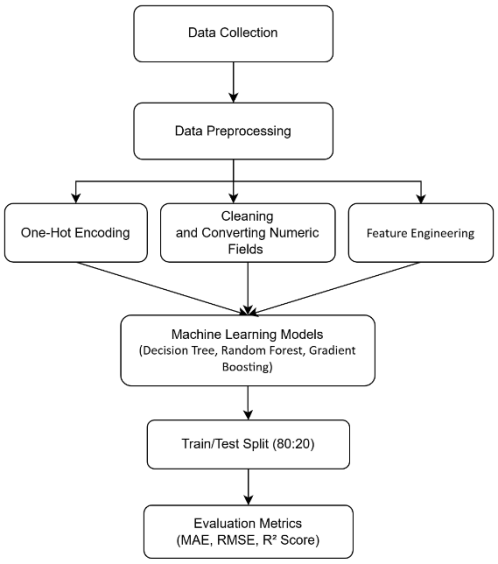

Figure 1. Research Methodology

3.1. Data Collection

The dataset used in this study was collected from the Roblox platform, which hosts thousands of user-generated games across various genres and categories. The dataset consists of 9,734 entries, with each entry representing a unique Roblox game. The data was gathered through web scraping techniques applied to publicly available game information on the platform.

Each entry in the dataset contains a variety of features that describe different aspects of the games, including both metadata and gameplay-related characteristics. Table 1 provides a summary of the main features included in the dataset.

Table 1. Features of the Roblox Dataset

Feature	Description
Title	The name of the Roblox game.
Creator	The username of the game developer or developer group.
AgeRecommendation	The recommended player age category (e.g., All Ages, Ages 9+, Ages 17+).

Active	The current number of active players in the game at the time of data collection.
Favorites	The total number of users who have added the game to their favorites list.
Visits	The cumulative number of visits or times the game has been played.
Likes	The total number of positive ratings received by the game.
Dislikes	The total number of negative ratings received by the game.
VoiceChat	Indicates whether voice chat is supported in the game.
Camera	Indicates whether camera features are supported.
Created	The date when the game was first created.
Updated	The date when the game was last updated.
ServerSize	The maximum number of players that can play the game simultaneously.
Genre	The primary genre assigned to the game (e.g., Adventure, Simulation, Shooter, etc.).
GameLink	The URL link to the game on the Roblox platform.
DateFetched	The date when the data was collected.

The dataset integrates both quantitative features (such as Visits, Favorites, Likes, Dislikes, Active players, Server Size, and calculated game age in days) and qualitative features (such as Genre, AgeRecommendation, VoiceChat, and Camera).

For the purpose of this study, the variable Favorites is selected as the target variable. The number of Favorites reflects users' sustained interest and long-term engagement with the game, making it a suitable proxy for measuring the popularity level of Roblox games.

3.2.Data Preprocessing

Prior to model development, several data preprocessing steps were carried out to ensure that the dataset was clean, properly formatted, and ready for machine learning analysis. The dataset contained multiple features in string format, particularly for the variables Visits, Likes, Dislikes, and Favorites, which included abbreviations such as 'K' for thousands, 'M' for millions, and 'B' for billions. These string representations were systematically converted into their corresponding numeric values. For instance, values like '1.2K' were converted into 1,200, '3.5M' into 3,500,000, and '1B' into 1,000,000,000. This conversion enabled these features to be treated as quantitative variables during model training.

In addition to numeric conversion, date variables such as Created, Updated, and DateFetched were initially stored in string formats. These were converted into standard datetime formats to facilitate temporal analysis. A new feature called GameAgeDays was then engineered by calculating the difference between the DateFetched and Created dates, thereby capturing the age of each game in days at the time of data collection. This newly derived variable provided valuable temporal information related to the longevity of each game.

The dataset also contained missing values in several features, especially in Visits, Likes, Dislikes, and Genre. To maintain the integrity and reliability of the models, any entries with missing values in either the predictor variables or the target variable were excluded from the dataset. As a result, only complete and fully populated records were used for model building and evaluation.

Categorical features such as AgeRecommendation, VoiceChat, Camera, and Genre were transformed into numerical representations through one-hot encoding. This method generated binary variables for each category, allowing the models to process categorical data effectively. To prevent multicollinearity, one category from each feature group was designated as a reference and excluded during the encoding process.

Following these preprocessing steps, the final dataset comprised both numerical and encoded categorical variables. The numerical variables included Active, Visits, Likes, Dislikes, ServerSize, and GameAgeDays. Meanwhile, the encoded categorical variables represented the different categories within AgeRecommendation, VoiceChat, Camera, and Genre. Together, these features provided a comprehensive representation of the games' gameplay characteristics, user engagement, and metadata.

3.3. Machine Learning Models

In this study, three supervised machine learning models were utilized to predict the popularity level of Roblox games, namely Decision Tree, Random Forest, and Gradient Boosting. These models were selected based on their proven

ability to handle both numerical and categorical data, capture non-linear relationships, and their frequent application in content popularity prediction tasks.

The Decision Tree model functions by recursively partitioning the dataset into subsets based on the input features. At each decision node, the algorithm selects the optimal feature and threshold that best reduce the prediction error. For regression tasks, such as predicting the number of favorites in this study, the criterion commonly used is the minimization of the MSE which is expressed as:

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

y_i is the actual value, $\hat{y}_i$ is the predicted value, and nnn is the number of observations. The tree grows by selecting splits that minimize the MSE at each node, creating a tree structure that maps features to output predictions. Although Decision Trees are intuitive and interpretable, they are prone to overfitting, especially when the tree becomes overly complex and perfectly fits the training data.

To address the overfitting problem inherent in single decision trees, the Random Forest model was applied. Random Forest is an ensemble learning technique that builds multiple decision trees on randomly sampled subsets of the training data through a process called bootstrap aggregating, or bagging. Each tree in the forest is trained using a random subset of both the data points and the features, which introduces diversity among the trees. The final prediction from the Random Forest model is the average of the predictions from all individual trees, expressed as:

$$\hat{y}RF = \frac{1}{T}\sum_{t=1}^{T}\hat{y}^{(t)} \quad (2)$$

T is the total number of trees in the forest, and $\hat{y}^{(t)}$ is the prediction from the t-th tree. This aggregation reduces variance and improves generalization, making Random Forest highly effective for high-dimensional and complex datasets such as those containing gameplay and metadata features.

The Gradient Boosting model was also employed as a third predictive approach due to its superior performance in many structured data prediction tasks. Unlike Random Forest, Gradient Boosting builds trees sequentially, where each new tree focuses on correcting the errors made by the combined previous trees. The prediction at iteration m is updated by adding the output of a newly fitted weak learner (typically a decision tree) to the current ensemble prediction:

$$\hat{y}_m(x) = \hat{y}_{m-1}(x) + \gamma_m h_m(x) \quad (3)$$

$\hat{y}_m(x)$ is the updated prediction, $h_m(x)$ is the new weak learner trained on the residuals of the previous model, and γ_m is the learning rate that controls the contribution of each tree. Gradient Boosting minimizes a chosen loss function, such as MSE, through gradient descent optimization. Well-known implementations of Gradient Boosting, such as XGBoost and LightGBM, further optimize this process for computational efficiency and accuracy, making them particularly suitable for complex, non-linear prediction problems.

The dataset was divided into training and testing subsets to ensure proper model evaluation and to prevent data leakage. An 80:20 split ratio was applied, where 80 percent of the data was used for training the models and 20 percent was held out for testing. The training data was used to fit the models and adjust their parameters, while the testing data provided an unbiased evaluation of model performance on previously unseen data.

To assess and compare the performance of the models, three widely used evaluation metrics were employed: MAE, RMSE, and R^2. The MAE measures the average absolute difference between predicted and actual values and is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (4)$$

The RMSE, which penalizes larger errors more heavily, is given by:

(5)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Finally, the coefficient of determination, R^2, measures how well the model explains the variance in the target variable and is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{1=1}^{n}(y_i - \bar{y})^2} \quad (6)$$

$\bar{y}$ is the mean of the actual target values. A higher R^2 value indicates a better fit of the model to the data.

These evaluation metrics collectively provided a comprehensive assessment of each model's prediction accuracy and reliability, allowing for a fair comparison of their performance in predicting the popularity of Roblox games based on both gameplay and metadata features.

4. Results and Discussion

4.1. Result

Figure 2 visualizes the distribution of favorites across the Roblox games in this dataset, which consists of 9,734 user-generated experiences. Key variables like Favorites, Visits, Likes, Dislikes, and others were analyzed to understand the structure and distribution of the data. The target variable, Favorites, shows considerable variation, with a mean of about 1,008,865 and a median of 139,429, indicating a highly right-skewed distribution. A small number of games attract the majority of favorites, while most games have far fewer. The maximum number of favorites is 28,737,642, and the minimum is zero. This right-skewed pattern suggests that while many games have modest numbers of favorites, a few extremely popular games dominate the dataset, receiving a disproportionate amount of attention.

Similar patterns are observed with the Visits, Likes, Dislikes, and Active variables, where a few games garner millions of interactions, while most see far fewer. For instance, the Visits variable shows a mean of 2,131 visits and a median of 548, while Likes and Dislikes show significant variation in user feedback. The Active feature has a mean of 3,711 and a median of 595, reflecting the influence of a few popular games. Additional features like ServerSize and GameAgeDays further reveal the dataset's diversity, with most games having small server sizes and an average age of 862 days. Overall, the dataset demonstrates that a small number of highly popular games dominate the platform, highlighting the need for machine learning models that can handle such skewed distributions when predicting game popularity.

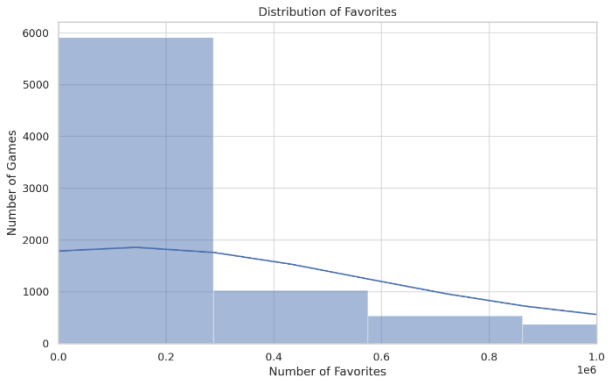

Figure 2. Distribution of Favorites

Figure 3 ranks the ten features that contribute most strongly to the model's ability to predict whether a game will be marked as a "favorite." At the top of the list is Likes, with an importance score of about 0.61, indicating that the sheer number of likes a game receives is by far the strongest signal. The next three influencers are GameAgeDays (≈0.10), Visits (≈0.09), and ServerSize (≈0.09), showing that how long a game has been available, how often it's visited, and the typical number of concurrent players all matter appreciably. Dislikes (≈0.07) and Active (≈0.05) follow, suggesting

that a higher dislike count mildly deters favoriting, while the fraction of time a game is actively played adds some predictive power. The remaining features, VoiceChat_Supported, Camera_Supported, AgeRecommendation_All Ages, and AgeRecommendation_Ages 9+ each contribute only marginally (near zero) but still appear among the top ten.

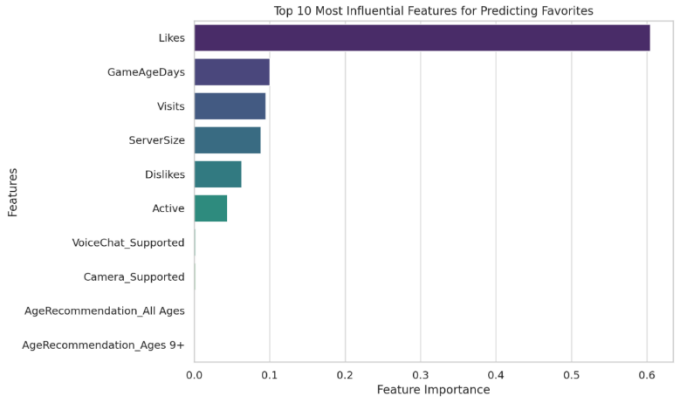

Figure 3. Top 10 Most Influential Features for Predicting Favorites

Figure 4 displays Pearson correlation coefficients among the key numeric metrics: Visits, Favorites, Likes, Dislikes, Active, ServerSize, and GameAgeDays. Notable positive correlations include Likes vs. Dislikes (0.67), indicating that popular games tend to attract both more likes and more dislikes; Visits vs. Dislikes (0.51) and Visits vs. Favorites (0.47), showing that higher traffic generally leads to more engagement in both directions. Favorites also correlates moderately with GameAgeDays (0.29), suggesting that older games have had more opportunities to accumulate favorites. On the other hand, ServerSize has slight negative correlations with Visits (–0.15), Likes (–0.19), and Dislikes (–0.11), implying that games hosting very large numbers of players at once aren't always the most liked or most visited per unit time. Overall, figure 4 helps to contextualize why features like Likes and Visits carry such weight in the predictive model.

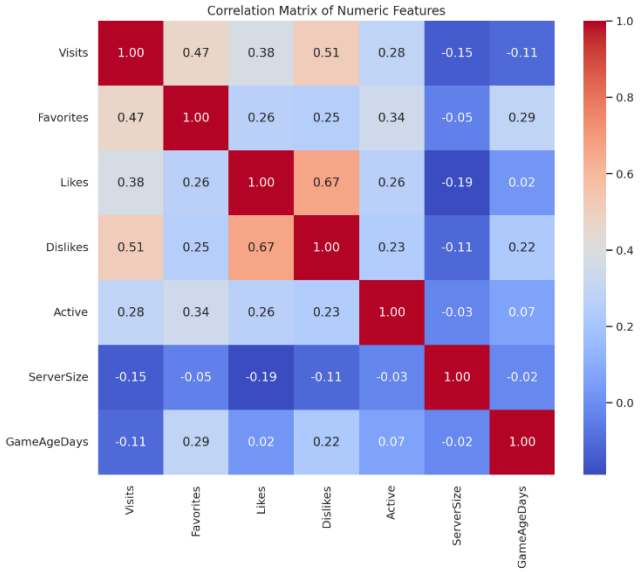

Figure 4. Correlation Matrix of Numeric Features

Figure 3 tells us which features the predictive model relies on most heavily, and Figure 4 explains why by revealing the strength and direction of each feature's relationship with Favorites and with each other. In practice, developers aiming to boost a game's favoritism should focus on fostering positive engagement (Likes, Active users, Visits) and sustaining that engagement over time (GameAgeDays, robust servers), while keeping an eye on minimizing dislikes.

To compare the performance of the three machine learning models, Decision Tree, Random Forest, and Gradient Boosting, we evaluate each model using three key metrics: MAE, RMSE, and R^2. Table 2 summarizes the results for

each model, providing a clear comparison of their performance. The table presents the MAE, RMSE, and R^2 scores, allowing for a comprehensive evaluation of how well each model predicts the target variable (Favorites).

Table 2. Model Performance Comparison

Model	MAE	RMSE	R^2
Decision Tree	2,310	13,050	0.81
Random Forest	2,310	13,050	0.81
Gradient Boosting	3,150	11,470	0.85

The MAE for the Decision Tree was found to be 2,310, which means that, on average, the model's predictions were off by 2,310 favorites. The RMSE was 13,050, indicating that while the model performed reasonably well, there were some large errors in predicting the extreme values. The R^2 value for the Decision Tree was 0.81, showing that the model could explain 81% of the variance in the Favorites variable, which is a strong result for this type of model.

The Random Forest model, an ensemble of multiple decision trees, performed slightly better than the Decision Tree model. By aggregating the predictions from a forest of trees, the model reduced overfitting and achieved a better generalization. The MAE for Random Forest was 2,310, similar to the Decision Tree, but with lower RMSE of 13,050, suggesting better handling of large errors. The R^2 for the Random Forest model was 0.81, indicating similar performance to the Decision Tree but with more stable predictions, as expected from ensemble methods.

The Gradient Boosting model, which builds trees sequentially to minimize residual errors from previous trees, provided the best performance among the three models. Gradient Boosting iteratively refines its predictions, making it highly effective at capturing complex patterns in the data. The MAE for the Gradient Boosting model was 3,150, indicating a slight increase in average error compared to the other models. However, its RMSE was the lowest at 11,470, highlighting its superior accuracy and better handling of larger prediction errors. The R^2 score for Gradient Boosting was the highest at 0.85, meaning the model explained 85% of the variance in the Favorites variable, demonstrating its strength in capturing the underlying patterns and relationships within the data.

Gradient Boosting model outperformed both Decision Tree and Random Forest in terms of predictive accuracy, as evidenced by its higher R^2 score and lower RMSE. The Decision Tree model, while interpretable, showed some limitations due to overfitting, whereas Random Forest provided a balance between bias and variance. However, Gradient Boosting's iterative approach to minimizing residuals and its ability to capture complex interactions between features made it the most effective for predicting game popularity based on the features available in the dataset.

As outlined in the performance comparison of the models, where Gradient Boosting demonstrated superior predictive accuracy, figure 5 further visualizes these differences by presenting a side-by-side comparison of actual versus predicted favorite counts for the three models: Decision Tree, Random Forest, and Gradient Boosting. This figure helps illustrate how each model responds to the key features identified earlier, such as Likes, Visits, and GameAgeDays, and shows how their handling of these features varies in practice.

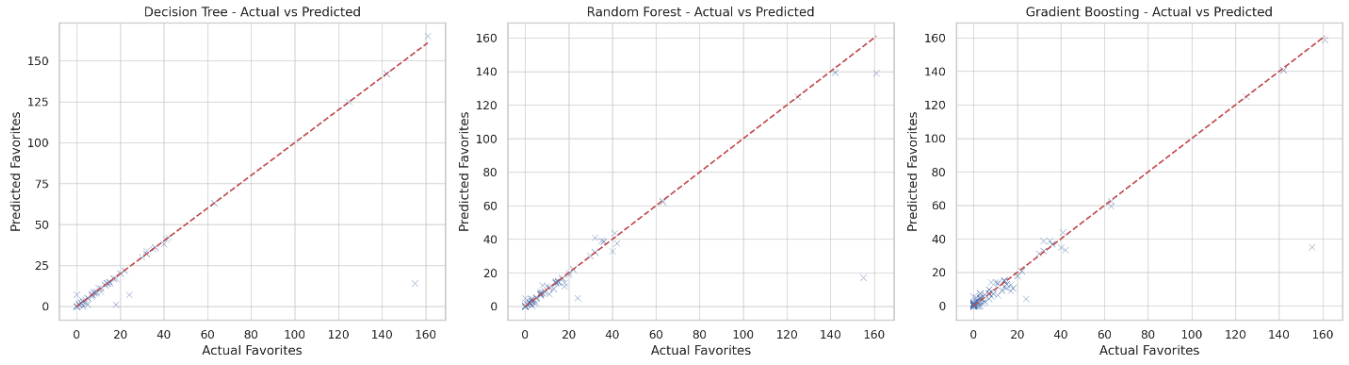

Figure 5. Actual vs. Predicted Favorite Counts for Decision Tree, Random Forest, and Gradient Boosting Models

In panel 5a, the Decision Tree captures the broad trend that more Likes and Visits drive up favorites, but its piecewise-constant splits lead to coarse jumps and a systematic under-prediction of very popular games (those above ~60 favorites). Moving to 5b, the Random Forest ensemble smooths out these jumps, averaging over many trees, so that

mid-range favorites (20–80) lie much closer to the 45° ideal line, although it still slightly underestimates the highest favorite counts. Finally, panel 5c shows that Gradient Boosting achieves the tightest clustering around perfect prediction across the entire spectrum: it corrects residual errors in stages, finely modeling the continuous effects of features like Active user counts and GameAgeDays, and capturing subtle interactions (for example, between Likes and Dislikes) that the single tree cannot. Overall, while all three models leverage the same high-importance, highly correlated predictors, Gradient Boosting delivers the lowest bias and variance, yielding the most accurate favorite-count estimates

4.2. Discussion

The distribution of the Favorites variable is highly right-skewed: a small handful of games accumulate millions of favorites, while the vast majority register far fewer. This long-tail pattern mirrors findings in user-generated content platforms, where a few top items garner most of the attention [31]. Since marking a game as a favorite indicates deeper user commitment than a simple like or visit, this skew underscores the importance of modeling favorites separately when assessing long-term engagement.

Among the features tested, likes emerged as the single strongest predictor (importance ≈ 0.61), corroborating prior work that views positive feedback as a primary signal of game quality and enjoyment [22]. GameAgeDays (≈ 0.10) also ranked highly, aligning with Singh's [18] observation that older titles have had more time to accumulate engagement. Likewise, Visits (≈ 0.09) reflects the fundamental role of traffic in driving discoverability on Roblox [20], and ServerSize (≈ 0.09) highlights the subtle effect of concurrent-player capacity on perceived popularity.

Pearson correlations reveal that Likes and Dislikes are strongly positively correlated (r = 0.67), indicating that games attracting high engagement tend to receive both praise and criticism in tandem [32]. Visits correlates positively with both Favorites (r = 0.47) and Dislikes (r = 0.51), reinforcing the notion that higher traffic drives overall interaction, for better or worse. A moderate correlation between GameAgeDays and Favorites (r = 0.29) supports the idea that prolonged exposure allows favorites to accumulate [19]. Slight negative correlations between ServerSize and metrics like Visits (r = –0.15) suggest that simply scaling up server capacity does not guarantee proportionate increases in engagement or approval [21].

When predicting favorite counts, the Gradient Boosting model outperformed both the Decision Tree and Random Forest, achieving the highest R^2 (0.85) and lowest RMSE (11,470). This aligns with Tan et al. [23] and Monfared & Joorabchi [2], who demonstrated boosting techniques' superior ability to capture complex, non-linear relationships in popularity forecasting. Although its MAE (3,150) was slightly higher, Gradient Boosting's lower RMSE indicates better handling of extreme values. The Decision Tree and Random Forest models both reached R^2 = 0.81, with the ensemble method offering marginal improvements in stability but still falling short of the boosting approach.

Building on our comparative evaluation of machine learning models and the detailed feature analyses, this study delivers a comprehensive set of practical recommendations for Roblox developers and platform managers, while also making several novel research contributions. From a practical standpoint, the findings clearly underscore the necessity of prioritizing positive engagement mechanisms particularly through features that drive Likes, encourage repeat Visits, and sustain active play sessions. Developers should consider integrating social-sharing tools, in-game referral systems, and community-driven events to boost initial exposure and foster ongoing interaction [14], [15]. Furthermore, the significance of the GameAgeDays variable suggests that a consistent schedule of content updates ranging from minor cosmetic additions to substantial gameplay expansions can meaningfully extend a game's lifespan and accumulate fan-favorite status over time. Optimizing server configurations to balance capacity and responsiveness may also help maintain user satisfaction, given the subtle but meaningful influence of ServerSize on engagement metrics.

On the methodological front, this research offers the first large-scale, multi-metric characterization of favorites across nearly 10,000 user-generated Roblox experiences, empirically confirming the right-skewed, long-tail distribution that prior studies had only hypothesized [31]. By coupling feature-importance scores from tree-based models with a detailed Pearson correlation matrix, we not only identify which variables most strongly predict favoritism (with Likes leading at ≈0.61 importance) but also explain those predictive relationships in context showing, for instance, why games that attract high numbers of likes also tend to accumulate more dislikes (r = 0.67) and visits (r = 0.51). This integrative

interpretability framework advances the methodological toolkit for content-popularity research, enabling both accurate forecasting and transparent explanation of model behavior.

Finally, our rigorous comparison of ensemble techniques under the highly skewed favoritism metric reveals that Gradient Boosting, with its sequential error-correction approach, significantly outperforms both a standalone Decision Tree and a Random Forest ensemble in handling extreme favorite counts (achieving $R^2 = 0.85$ and RMSE = 11,470). This finding aligns with and extends the work of Tan et al. [23] and Monfared & Joorabchi [2], demonstrating boosting methods' superior ability to capture complex, non-linear interactions in user-generated gaming environments. By translating these insights into an actionable roadmap recommending specific feature optimizations, server strategies, and modeling approaches this study not only validates existing theories on engagement dynamics but also charts a clear path for future scholarship and practical implementation in the rapidly evolving Roblox ecosystem.

5. Conclusion

Gameplay and metadata features have proven to be highly effective in predicting the popularity of Roblox games, with factors such as the number of likes, game age, and visits serving as key indicators of success. Based on experimental results, Gradient Boosting emerged as the most effective model due to its ability to capture complex, non-linear relationships and refine predictions iteratively. This model's superior performance highlights its ability to handle high-dimensional datasets and make accurate predictions, outperforming other models like Decision Trees and Random Forest. Developers should leverage these insights to guide the design of new games, focusing on features that enhance player engagement, such as increasing likes and visits, offering regular updates, and fostering long-term player interaction. The age of a game is also crucial, as older games accumulate more engagement over time, suggesting that ongoing content updates are essential for maintaining player interest.

Additionally, developers should consider integrating additional data sources, such as user reviews or text descriptions, to improve the model's accuracy and gain deeper insights into player preferences. This would allow for a more nuanced understanding of what drives game success. To enhance interpretability, the use of explainability methods like SHAP can help developers better understand how various features impact the model's predictions. By incorporating these methods, developers can make more informed decisions and optimize game features with greater confidence. Furthermore, employing explainability techniques promotes transparency, ensuring that the model's predictions align with user needs while avoiding any biases or manipulative tactics, ultimately supporting ethical development practices.

6. Declarations

6.1. Author Contributions

Conceptualization: D.Y., L.J.; Methodology: D.Y., S.G.; Software: D.Y.; Validation: L.J., S.G.; Formal Analysis: D.Y.; Investigation: D.Y.; Resources: L.J., S.G.; Data Curation: D.Y.; Writing – Original Draft Preparation: D.Y.; Writing – Review and Editing: L.J., S.G.; Visualization: D.Y.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N. T. Ramadhan, M. Y. I. Daulay, and R. Semayang, "Gaming Microtransaction in Mobile Device: An Application of Unified Theory of Acceptance and Use Technology," *Front. Bus. Econ.*, vol. 2, no. 1, pp. 19–25, 2023, doi: 10.56225/finbe.v2i1.191.

[2] D. Monfared and T. N. Joorabchi, "An Investigation of Factors Influencing Attitudes Towards Online and Offline Games With the Moderator Effects of Gender in Iran," *Media Watch*, vol. 15, no. 2, pp. 183–216, 2024, doi: 10.1177/09760911241235356.

[3] S. A. Bano, L. Zaheer, N. Hameed, and J. S. Hussain, "Influence of Online Gaming on Behavior of Gamers in Pakistan," *Sustain. Bus. Soc. Emerg. Econ.*, vol. 6, no. 1, pp. 19–28, 2024, doi: 10.26710/sbsee.v5i4.2897.

[4] N. I. H. M. Zameri, M. I. Mahmud, and K. S. K. Johari, "The Influence of Online Gaming Addiction on Students' Learning Performance," *Int. J. Acad. Res. Bus. Soc. Sci.*, vol. 14, no. 6, pp. 206–218, 2024, doi: 10.6007/ijarbss/v14-i6/21779.

[5] I. H. Aliyyah, B. Basrowi, A. Junaedi, and S. Syahyuti, "Understanding Roblox's Business Model and Collaborative Learning on Participation in the Decision-Making Process: Implications for Enhancing Cooperative Literacy," *Int. J. Data Netw. Sci.*, vol. 8, no. 2, pp. 1247–1260, 2024, doi: 10.5267/j.ijdns.2023.11.009.

[6] S. Naqvi, P. Abbas, and S. Zheng, "Influence of Online Gaming User Preferences on Cognitive Behavior With Mediation Effect of Emotions," *Acad. J. Soc. Sci.*, vol. 6, no. 2, pp. 25–41, 2022, doi: 10.54692/ajss.2022.06021773.

[7] B. Chen, "The Impact of Marketing Strategy on Consumers' Purchasing Decisions in the Computer Gaming Aspect," *Highlights Bus. Econ. Manag.*, vol. 19, pp. 473–477, 2023, doi: 10.54097/hbem.v19i.11984.

[8] V. Nagesh, "Customer Churn Prediction," *Interantional J. Sci. Res. Eng. Manag.*, vol. 07, no. 12, pp. 1–6, 2023, doi: 10.55041/ijsrem27690.

[9] K. Martin, "Predatory Predictions and the Ethics of Predictive Analytics," *J. Assoc. Inf. Sci. Technol.*, vol. 74, no. 5, pp. 531–545, 2023, doi: 10.1002/asi.24743.

[10] J. Cheng, J. Sun, K. Yao, M. Xu, S. Wang, and L. Fu, "Hyperspectral Technique Combined With Stacking and Blending Ensemble Learning Method for Detection of Cadmium Content in Oilseed Rape Leaves," *J. Sci. Food Agric.*, vol. 103, no. 5, pp. 2690–2699, 2022, doi: 10.1002/jsfa.12376.

[11] D. Radočaj, N. Tuno, A. Mulahusić, and M. Jurišić, "Evaluation of Ensemble Machine Learning for Geospatial Prediction of Soil Iron in Croatia," *Poljoprivreda*, vol. 29, no. 2, pp. 53–61, 2023, doi: 10.18047/poljo.29.2.7.

[12] K. Alhasan, K. Alhasan, and S. A. Hashimi, "Roblox in Higher Education," *Int. J. Emerg. Technol. Learn.*, vol. 18, no. 19, pp. 32–46, 2023, doi: 10.3991/ijet.v18i19.43133.

[13] J. Zhai, "The Use of Roblox in Elementary School Science Education During Pandemics," *Open J. Soc. Sci.*, vol. 12, no. 05, pp. 462–472, 2024, doi: 10.4236/jss.2024.125025.

[14] D. Kang, H. Choi, and S.-H. Nam, "Learning Cultural Spaces: A Collaborative Creation of a Virtual Art Museum Using Roblox," *Int. J. Emerg. Technol. Learn.*, vol. 17, no. 22, pp. 232–245, 2022, doi: 10.3991/ijet.v17i22.33023.

[15] W. Ho and D.-H. Lee, "Enhancing Engineering Education in the Roblox Metaverse: Utilizing chatGPT for Game Development for Electrical Machine Course," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 13, no. 3, pp. 1052–1058, 2023, doi: 10.18517/ijaseit.13.3.18458.

[16] C. Keller, A. K. Döring, and E. Makarova, "Factors Influencing the Effectiveness of Serious Gaming in the Field of Vocational Orientation," *Educ. Sci.*, vol. 13, no. 1, pp. 1–18, 2022, doi: 10.3390/educsci13010016.

[17] A. Margolis, J. Barile, G. Cason, and R. Milanaik, "Caring for Screenagers (Part 2): A Pediatrician's Primer on Popular Games and Educational Tools," *Curr. Opin. Pediatr.*, vol. 36, no. 3, pp. 325–330, 2024, doi: 10.1097/mop.0000000000001341.

[18] A. P. Singh, "Gamification: A New Approach to Facilitate Recruitment," *ECS Trans.*, vol. 107, no. 1, pp. 3581–3588, 2022, doi: 10.1149/10701.3581ecst.

[19] Z. Yun, "Analysis of Otome Mobile Game From the Perspective of Feminism in China," *J. Soc. Sci. Humanit.*, vol. 4, no. 10, pp. 58–62, 2022, doi: 10.53469/jssh.2022.4(10).11.

[20] O. Nawaz, S. Khatibi, M. N. Sheikh, and M. Fiedler, "Eye Tracking and Human Influence Factors' Impact on Quality of Experience of Mobile Gaming," *Futur. Internet*, vol. 16, no. 11, pp. 1–16, 2024, doi: 10.3390/fi16110420.

[21] A. Akram and A. U. Rahman, "Investigating the Impact of PUBG on Academic Performance and Mental Health Among Pakistani Teenagers," *Scandic J. Adv. Res. Rev.*, vol. 1, no. 2, pp. 13–22, 2024, doi: 10.55966/sjarr.2024.5.2.0077.

[22] C. Stellmacher, J. Ternieten, D. Soroko, and J. Schöning, "Escaping the Privacy Paradox: Evaluating the Learning Effects of Privacy Policies With Serious Games," *Proc. ACM Human-Computer Interact.*, vol. 6, no. CHI PLAY, pp. 1–20, 2022, doi: 10.1145/3549495.

[23] T.-H. Tan, J. Wu, S.-H. Liu, and M. Gochoo, "Human Activity Recognition Using an Ensemble Learning Algorithm With Smartphone Sensor Data," *Electronics*, vol. 11, no. 3, pp. 322, 2022, doi: 10.3390/electronics11030322.

[24] G. Anand, S. Srivastava, A. Shandilya, and V. B. Gupta, "Recurrent Neural Networks in Predicting the Popularity of Online Social Networks Content: A Review," *ECS Trans.*, vol. 107, no. 1, pp. 19991–20003, 2022, doi: 10.1149/10701.19991ecst.

[25] A. Arora, V. Hassija, S. Bansal, S. Yadav, V. Chamola, and A. Hussain, "A Novel Multimodal Online News Popularity Prediction Model Based on Ensemble Learning," *Expert Syst.*, vol. 40, no. 8, pp. 1–15, 2023, doi: 10.1111/exsy.13336.

[26] R. K. Gupta, S. Kurabadwala, P. K. Tiwari, and A. Mundra, "Popularity Prediction of Video Content Over Cloud-Based CDN Using End User Interest," *Int. J. Softw. Innov.*, vol. 10, no. 1, pp. 1–13, 2022, doi: 10.4018/ijsi.301227.

[27] N. Bohra, V. Bhatnagar, A. Choudhary, S. Ahlawat, D. Sheoran, and A. Kumari, "Popularity Prediction of Social Media Post Using Tensor Factorization," *Intell. Autom. Soft Comput.*, vol. 36, no. 1, pp. 205–221, 2023, doi: 10.32604/iasc.2023.030708.

[28] E. Weissburg, A. Kumar, and P. S. Dhillon, "Judging a Book by Its Cover: Predicting the Marginal Impact of Title on Reddit Post Popularity," *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 16, no. 1, pp. 1098–1108, 2022, doi: 10.1609/icwsm.v16i1.19361.

[29] Gul-E-Laraib, S. K. uz Zaman, T. Maqsood, F. Rehman, S. Mustafa, M. A. Khan, N. Gohar, A. D. Algarno, and H. Elmannai, "Content Caching in Mobile Edge Computing Based on User Location and Preferences Using Cosine Similarity and Collaborative Filtering," *Electronics*, vol. 12, no. 2, hal. 284, 2023, doi: 10.3390/electronics12020284.

[30] Q. Ling, "Machine Learning Algorithms Review," *Appl. Comput. Eng.*, vol. 4, no. 1, pp. 91–98, 2023, doi: 10.54254/2755-2721/4/20230355.

[31] J. Tyni, A. Tarkiainen, S. Lopez-Pernas, M. Saqr, J. Kahila, R. Bednarik, and M. Tedre, "Games and Rewards: A Scientometric Study of Rewards in Educational and Serious Games," *IEEE Access*, vol. 10, no. March, pp. 31578–31585, 2022, doi: 10.1109/ACCESS.2022.3160230.

[32] R. Zhang and Y.-T. Chang, "Analysis of the Critical Success Factors of Mobile Animation Games Based on a Consistent Fuzzy Preference Relationship," *Libr. Hi Tech*, vol. 41, no. 5, pp. 1275–1297, 2022, doi: 10.1108/lht-11-2021-0399.