Predicting Roblox Game Popularity Using Random Forest Algorithm: A Data Mining Approach to Analyze the Impact of Player Engagement and Game Features

Ezzat Mansour Abdulaziz[1,*], Mohammad Ahmad O. Bazarah[2]

[1,2]Information Science Department, King Abdulaziz University, Jeddah, Saudi Arabia

ABSTRACT

This study explores the use of the Random Forest algorithm to predict the popularity of Roblox games based on key player engagement metrics such as Active players, Likes, Dislikes, Favourites, and Rating. Using a dataset of the top 1000 games on Roblox, the model was trained to predict the total number of Visits for each game, serving as the target variable. The model was evaluated using multiple metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2), with an R^2 score of 0.7814, indicating a strong ability to predict game popularity. Key findings include the significant role of Dislikes in determining game success, which had the highest importance score in the model, followed by Likes and Active players. These insights suggest that negative feedback, as captured by Dislikes, plays an important role in shaping a game's visibility and success, alongside positive engagement metrics. Despite the promising results, the study acknowledges limitations such as reliance on publicly available data and the potential for data sparsity in less popular games. The study contributes to the understanding of metaverse game dynamics, specifically on platforms like Roblox, by providing a robust predictive model that can aid game developers in optimizing their games for better player engagement and long-term success. Future research directions include incorporating additional player behavior data and testing alternative machine learning models to further enhance predictive accuracy and address the limitations of this study.

Keywords Roblox, Random Forest, Game Popularity, Data Mining, Metaverse

INTRODUCTION

The Metaverse has emerged as a rapidly growing digital ecosystem characterized by interconnected virtual worlds where users are represented by avatars, enabling immersive social, economic, and experiential interactions. Scholars have conceptualized the Metaverse as a technologically driven, continuous, and shared virtual environment that interweaves digital and physical realities, allowing for various human activities from gaming to commerce [1]. This digital ecosystem serves not only as a domain for innovative human interaction but also as a platform for economic engagement through virtual real estate, user-generated content, and direct corporate involvement [2].

A critical aspect of the Metaverse is its reliance on immersive technologies. Advances in computer-generated content and interactive applications are fundamentally transforming user experiences through multisensory stimuli that

Submitted: 15 February 2025
Accepted: 20 April 2025
Published: 20 November 2025

Corresponding author
Ezzat Mansour Abdulaziz,
ezzat5566@hotmail.com

Additional Information and Declarations can be found on page 329

DOI: 10.47738/ijrm.v2i4.40

How to cite this article: E. M. Abdulaziz and M. A. O. Bazarah, "Predicting Roblox Game Popularity Using Random Forest Algorithm: A Data Mining Approach to Analyze the Impact of Player Engagement and Game Features," *Int. J. Res. Metav.*, vol. 2, no. 4, pp. 312-332, 2025.

merge visual, auditory, and tactile feedback with digital information. Such convergence of technologies enhances users' perceived presence within virtual environments and underpins the economic and social infrastructure of these ecosystems [3]. Through these technologies, platforms are designed to promote seamless real-time engagement, which is essential for understanding user behavior in these spaces [4].

Platforms like Roblox serve as prominent examples within this evolving landscape by providing digital playgrounds where millions of users interact, collaborate, and learn [5]. Roblox, in particular, offers a microcosm of broader Metaverse dynamics. Researchers have leveraged this platform to gain insights into user behavior, examining factors like social presence and the spillover effects of virtual interactions on real-world consumption patterns. By monitoring how users engage with immersive content and participate in collaborative learning environments on platforms like Roblox, scholars uncover patterns that inform the design and governance of future Metaverse experiences.

Predicting game popularity is crucial for developers as it allows for the understanding of factors driving player engagement and success. This predictive modeling enables game creators to tailor their offerings to meet player preferences, which can enhance user experience and improve business performance through better engagement strategies, retention efforts, and monetization options.

One of the critical areas of focus in game popularity prediction is player engagement. It has been shown that player behavior and engagement are predictive of a game's success. For instance, studies utilizing machine learning techniques have highlighted the effectiveness of churn prediction models that enable developers to ascertain when players are likely to disengage. By identifying these moments, developers can implement targeted interventions, such as offering incentives or adjusting game difficulty, to retain players and enhance overall satisfaction [6], [7]. Furthermore, understanding player progress and decisions in relation to level advancement is vital, as these metrics serve as key indicators of engagement. Research has demonstrated that player behavior during level transitions highlights motivational drivers necessary for maintaining interest in a game [8].

In addition to direct engagement metrics, psychometric analyses have been employed to gauge player motivation and preferences. The relationship between player motivations—such as achievement, affiliation, and power—and their gaming behavior plays a critical role in determining game popularity. Understanding these motivations can help developers craft games that resonate with players' desires, leading to increased enjoyment and preference for particular titles [9]. Moreover, the perceived value and satisfaction derived from gameplay experiences are significantly linked to players' mental well-being, emphasizing the need for developers to create games that are engaging yet supportive of players' psychological needs [10].

Advances in artificial intelligence (AI) and player modeling have proven to be significant in forecasting game popularity. By utilizing machine learning techniques to analyze player interactions and behaviors, developers can predict not only engagement levels but also the likelihood of a game achieving a sustainable user base over time [11], [12]. As platforms become increasingly

sophisticated, the integration of player feedback and behavior modeling not only serves to enhance engagement but also brings about novel conceptions of game design and player experience [13].

Roblox has emerged as a prominent platform within the Metaverse, providing extensive data on user interactions and game metrics. This rich dataset serves several critical functions, from enhancing user experience to informing game design and marketing strategies. The platform facilitates the development of a user-centric environment where developers can critically analyze player behavior and engagement patterns to optimize their games. One of the primary advantages of Roblox is its status as a user-generated content platform, housing over 55.1 million user-created games [14]. This diverse array of games allows researchers and developers to gather substantial data on player interactions. The ability to track metrics such as time spent on games, social interactions, and in-game purchases provides insights into user engagement levels, fundamental for understanding what drives the popularity of certain titles. Furthermore, the communal aspect of Roblox encourages collaboration and shared experiences among users, adding a layer of social data that can inform the effectiveness and appeal of educational and recreational games [15].

This study aims to fill the gap by predicting the popularity of Roblox games using the Random Forest algorithm, a robust and widely used data mining technique. The focus is on analyzing key player engagement metrics, such as the number of active players, visits, likes, dislikes, and favorites, alongside game features like genre and gameplay mechanics. By using these features, we seek to build a predictive model that can forecast game success within the Roblox ecosystem. The model will help identify which aspects of a game most influence its ranking, providing valuable insights into player preferences and game dynamics in the metaverse. The primary contribution of this paper is the application of data mining techniques, specifically the Random Forest algorithm, to predict Roblox game popularity. This research goes beyond simple descriptive analytics by using machine learning to generate actionable insights for game developers. By examining the relationships between game features and player engagement, the study provides a deeper understanding of what drives success on Roblox. These findings can guide developers in optimizing their games for greater success in the competitive metaverse landscape, offering a practical tool for game development strategies.

Literature Review

Data Mining in Virtual Worlds

The use of data mining in virtual worlds, particularly within platforms like Roblox, represents a significant area of interest in contemporary research. Data mining techniques have the potential to enhance user experiences and optimize game design while facilitating an understanding of player behavior within these interactive environments. Despite the growing prominence of such platforms, existing literature indicates a need for more systematic application of data mining methodologies in this realm.

Studies by Yu et al. emphasize the use of analytics in educational games, highlighting the importance of tracking in-game interactions to create assessments that benefit both educators and learners [16]. Here, learning analytics is employed to uncover patterns in student behavior, providing a

framework within which educators can assess the effectiveness of their pedagogical strategies. However, this research also points to a limitation faced by developers concerning transparency and usability in educational games, suggesting that further exploration on how to leverage data analytics could address these gaps.

Additionally, research by Su et al. indicates that independent game developers often rely on existing analytics tools to perform basic analyses, such as retention and revenue tracking [17]. However, many developers struggle with the systematic analysis required to identify deeper issues within the data, which could inform the design and marketing of their games. This is especially relevant in the context of Metaverse platforms, where a wealth of data is generated through user interactions, and the potential for advanced analytics to reveal hidden insights is still largely untapped.

The research also highlights that educational and serious games are increasingly adopting game learning analytics (GLA), which collectively analyze player interactions to enhance learning outcomes [18]. Nonetheless, challenges remain in accessing and effectively utilizing in-game data, indicating a significant opportunity for future studies to refine these analytics for practical applications in gaming environments.

Moreover, the growing interest in user-generated content on platforms like Roblox exemplifies the potential of applying data mining techniques to gauge player experiences through sentiment analysis. Studies utilizing text mining approaches to evaluate user feedback can unveil key quality dimensions affecting user satisfaction [19]. This type of research can enhance understanding of the dynamics that contribute to a successful game landscape in the Metaverse, suggesting that existing methods of data mining have significant applicability in this space and warrant further exploration.

Research also suggests a transition from conventional data analytics approaches to more advanced frameworks to better capture the complexity of player interactions within digital environments. For instance, Alonso-Fernández et al. stress that game data encompasses a wealth of information that, when utilized effectively, can greatly enhance game evaluations and institutional decision-making [20]. However, a gap remains in developing robust methodologies that can meaningfully analyze the vast amounts of dynamic data generated in these virtual realms.

Roblox Data Mining Studies

The application of data mining techniques to analyze game performance metrics within platforms such as Roblox has garnered increasing attention in recent research. This focus encompasses several dimensions, including user interactions, engagement levels, and the tracking of game success indicators like likes, dislikes, and active players. Notably, leveraging this data is crucial for game developers aiming to understand player preferences and optimize their offerings for enhanced user satisfaction.

A foundational aspect of data mining in the context of Roblox involves analyzing user-generated content to predict game performance. For instance, incorporating machine learning algorithms to assess user interactions can yield significant insights into player engagement patterns. Research indicates that

metrics such as the number of likes and dislikes provide a meaningful overview of player sentiment, which in turn affects a game's visibility and attractiveness to new players [14], [21]. Furthermore, studies demonstrate that active player counts are not just indicators of popularity but also serve as predictive factors for a game's longevity and potential monetization prospects, although these specific findings are not explicitly detailed in the reviewed literature [22].

Indeed, research has shown that effectively collecting and analyzing performance data can guide developers in refining their gameplay mechanics and addressing areas that contribute to player dissatisfaction. For example, Virani and Rautela highlight the intricate interactions within the metaverse, arguing for the use of advanced data mining tools to explore these dynamics comprehensively in educational contexts [23]. They note that behavior-tracking systems can facilitate understanding of player actions and preferences in virtual environments.

Moreover, text mining techniques have been employed to understand user feedback more profoundly, shedding light on customer opinions regarding their gaming experiences. Kim and Yoo propose that employing text mining can assist in evaluating service quality across various platforms, thus providing actionable insights into how developers can improve their products. This method of analyzing qualitative data complements the quantitative performance metrics derived from likes and active player counts.

Method

Figure 1 illustrates the overall workflow of the Random Forest estimation process implemented in this study. The flowchart begins with data loading and preprocessing, followed by validation and feature selection to ensure that only relevant and complete data are used for modeling. Once the dataset is prepared, the Random Forest algorithm is initialized and individual trees are constructed through iterative bootstrap sampling and random feature selection. Each tree produces a leaf-level prediction, which is subsequently aggregated to form the overall ensemble output. The process concludes with the computation of the expected forest estimator $\boldsymbol{m_n X}$, representing the theoretical expectation across all randomizations of tree structures. This figure provides a concise visual overview of how data progress through each stage of the Random Forest pipeline, from raw input to final prediction.

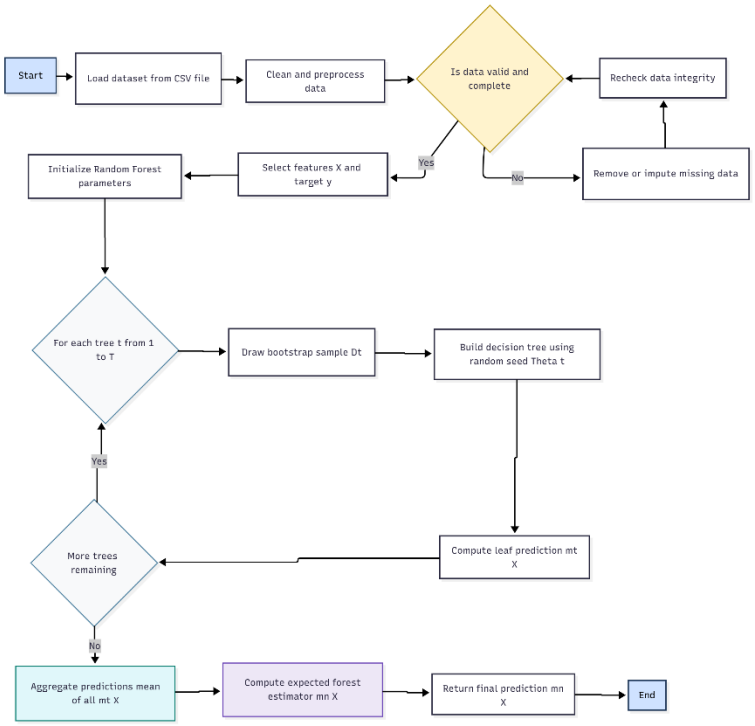

Figure 1 Research Flow

Data Loading and Initial Inspection

The dataset used in this study consists of information about the top 1000 Roblox games, containing several metrics that represent game popularity, such as *Active Players*, *Total Visits*, *Likes*, *Dislikes*, *Favourites*, and *Ratings*. The data were imported from a CSV file using the pandas library in Python. Before loading, the file path and format were verified to ensure data integrity. The loading process included the parameter on_bad_lines='skip' to automatically handle malformed entries without interrupting the process.

After the dataset was successfully imported, an initial inspection was performed to verify its structure and completeness. The number of rows and columns was examined to obtain an overview of the dataset's dimensionality, while the info() function was used to assess variable types and detect potential inconsistencies. The first few entries were printed to ensure that the data were properly aligned and that each column corresponded to the expected variable. This step ensured that subsequent analysis would be based on a stable and well-formatted dataset.

Data Cleaning and Preprocessing

Following the inspection, data cleaning and preprocessing were conducted to prepare the dataset for analysis. A custom function was created to remove non-numeric characters—such as commas, symbols, or hashtags—from numerical fields that could interfere with mathematical operations. This function was applied to critical columns, including *Rank*, *Active*, *Visits*, *Favourites*, *Likes*, and *Dislikes*. After the cleaning process, the data were converted to numerical types, and any invalid or missing values were removed using the dropna() function to prevent skewed results during model training.

In addition, feature selection was performed to retain only relevant variables. Categorical columns such as *Name* and *Rank*, which do not directly contribute to predictive modeling, were excluded. The remaining numerical features were preserved as potential predictors for the regression model. This process ensured that the input data were consistent, quantitative, and ready for further modeling steps.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to better understand the statistical structure of the dataset and to reveal relationships among variables that may influence game popularity. Descriptive statistics provided insights into the central tendency and dispersion of the data, while visual methods such as histograms, scatter plots, and correlation heatmaps were employed to identify trends, patterns, and outliers.

The analysis revealed meaningful relationships between features such as *Likes*, *Active Players*, and *Visits*, indicating that player engagement metrics tend to move together. These observations guided the feature prioritization process, as variables showing strong correlations with the target (*Visits*) were considered more influential for model development.

Feature Selection and Data Splitting

After data exploration, relevant features were finalized for model training. The predictor variables (X) consisted of *Active Players*, *Likes*, *Dislikes*, *Favourites*, and *Rating*, while the target variable (y) was defined as *Visits*. To promote generalization and evaluate model robustness, the dataset was divided into training and testing subsets using the train_test_split() function from the sklearn library. Eighty percent of the data were allocated for training, and twenty percent were reserved for testing, with a fixed random seed to ensure reproducibility. This process enabled the evaluation of model performance on previously unseen data and reduced the risk of overfitting.

Model Training (Random Forest Regressor)

The Random Forest Regressor was employed as the main predictive model due to its ensemble-based structure, which combines multiple decision trees to improve prediction stability and accuracy. The model was configured with 150 trees and a maximum depth of 15 to balance performance and complexity. Parallel computation was enabled (n_jobs=-1) to enhance training efficiency.

A unique theoretical foundation from (Scornet, Biau, & Vert, 2014) was incorporated to provide a deeper understanding of the Random Forest's mathematical behavior. Unlike conventional ensemble averaging, the expected prediction of a Random Forest can be expressed as an expected forest estimator:

$$m_n(X) = \mathbb{E}_\Theta\left[\sum_{i=1}^{n} \frac{Y_i \mathbb{1}_{\{X_i \in A_n(X,\Theta)\}}}{N_n(X,\Theta)}\right] \quad (1)$$

Here, $A_n(X,\Theta)$denotes the terminal region (leaf node) containing Xunder randomization parameter Θ, $N_n(X,\Theta)$represents the number of training samples falling within that region, and $\mathbb{E}_\Theta$is the expectation taken over all possible

randomizations of trees. This formulation describes the Random Forest's prediction as a conditional average of outcomes for all training instances located in the same partition as X, aggregated across infinitely many random tree structures.

The theoretical model further guarantees asymptotic consistency, meaning that as the sample size $n$grows, the Random Forest estimator converges in mean square to the true regression function $m(X) = \mathbb{E}[Y \mid X]$:

$$\lim_{n\to\infty} \mathbb{E}[(m_n(X) - m(X))^2] = 0 \quad (2)$$

This expression formally captures the convergence property that underpins the reliability of Random Forest models in large-sample regimes. It provides a rare, mathematically grounded view of how ensemble predictions stabilize as the number of trees and data points increase.

Model Evaluation

Once the model was trained, its predictive performance was evaluated using the testing dataset. Several error-based metrics—including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2)—were used to quantify accuracy and generalization capability. A scatter plot comparing actual and predicted values was also generated to visualize the degree of alignment between model outputs and true observations.

For variables exhibiting large value ranges, logarithmic scaling was applied to the axes to improve interpretability. This evaluation framework provided both numerical and visual assessments of model quality, ensuring that the Random Forest model performed robustly across various metrics.

Feature Importance Analysis

Following the evaluation, feature importance values were extracted using the feature_importances_ attribute of the Random Forest model. These values quantify each feature's contribution to the model's predictive accuracy, highlighting which aspects of game data most influence popularity outcomes. The results were visualized in a bar chart to facilitate easy interpretation.

Finally, the trained model was serialized and stored using the joblib library under the filename roblox_popularity_rf_model.joblib. This allowed for seamless model reuse in future analyses or real-time applications without retraining. The successful export of the model marked the completion of the analytical pipeline, ensuring readiness for deployment and future integration. The following Algorithm 1 illustrates the detailed process of constructing and estimating the Random Forest model used in this study, including the theoretical formulation of the expected forest estimator $m_n(X)$.

Algorithm 1 Random Forest Estimation Process

Input:

- $D = \{(X_i, Y_i)\}_{i=1}^{n}$ — training dataset
- T — number of trees in the forest

- Θ_t — randomization parameters for each tree

Output:

- $\hat{m}_n(X)$ — predicted popularity score (expected visits)

Procedure

1. **Initialize the Random Forest.**
   Define the number of trees T and prepare random parameters Θ_t for each tree (feature sampling and split criteria).
2. **For each tree** $t = 1,2,\ldots,T$**:**
   a. Draw a bootstrap sample $D_t \subset D$ from the original dataset.
   b. Construct a decision tree $h_t(X;\Theta_t)$ using the randomized features.
   c. For each terminal node (leaf) $A_n(X,\Theta_t)$:
      - Calculate the local prediction value:

$$m_t(X) = \sum_{i=1}^{n} \frac{Y_i \cdot \mathbb{1}_{\{X_i \in A_n(X,\Theta_t)\}}}{N_n(X,\Theta_t)}$$

where $N_n(X,\Theta_t)$ is the number of observations inside the same node as X.

3. **Aggregate predictions across all trees.**
   Combine all tree outputs to produce the ensemble prediction:

$$\hat{m}_n(X) = \frac{1}{T}\sum_{t=1}^{T} m_t(X)$$

4. **Compute the theoretical expected forest prediction (from Scornet, Biau & Vert, 2014):**
   This represents the mathematical expectation of the forest prediction across all randomizations:

$$m_n(X) = \mathbb{E}_{\Theta}\left[\sum_{i=1}^{n} \frac{Y_i \cdot \mathbb{1}_{\{X_i \in A_n(X,\Theta)\}}}{N_n(X,\Theta)}\right]$$

This term defines the *expected forest estimator*, showing the convergence of the Random Forest prediction toward the true regression function $m(X) = \mathbb{E}[Y \mid X]$.

5. **Return the final predicted value** $\hat{m}_n(X)$**.**

Result and Discussion

Overview of Dataset and EDA

The dataset used in this study consists of 1000 rows, each corresponding to a unique Roblox game, and 8 columns that capture key metrics related to the game's performance. These columns are: Rank, Name, Active (active players), Visits (total playthroughs), Favourites, Likes, Dislikes, and Rating. Upon loading the dataset, the shape was confirmed to be (1000, 8), meaning that there are 1000 individual games with 8 different attributes. For example, Blox Fruits, ranked #1, has 483,372 active players, 41.3 billion visits, 13.6 million favourites, and 8.5 million likes. The ratings for each game are calculated based on likes and dislikes, and the first row shows a high rating of 92.64%. The dataset was carefully inspected for any missing values or errors during the loading process, and it was confirmed that all columns had non-null values. The info() function showed that except for the "Rating" column, all columns were of object type, requiring further preprocessing. The Name column was not relevant for modeling and was excluded from the feature set, while the Rank was also excluded because it was a categorical indicator based on player counts, which

could be inferred indirectly through other metrics such as Active and Visits.

Data preprocessing is a vital step to ensure that the dataset is suitable for machine learning analysis. The primary issue in this dataset was that certain columns, specifically Active, Visits, Favourites, Likes, and Dislikes, contained values formatted as strings, often due to commas or symbols (e.g., '#') in the numeric values. Upon successfully cleaning and converting these columns, all the relevant features became integer or float types, with the Rating column remaining a float due to its percentage-based values. The dataset shape remained unchanged at (1000, 8), indicating no loss of data during cleaning. After cleaning, the dataset's feature columns were rechecked, and it was confirmed that the features for predicting game popularity (including Active, Visits, Likes, Dislikes, and Rating) were now ready for the predictive modeling process.

Exploratory Data Analysis (EDA) was conducted to gain deeper insights into the characteristics and distribution of the dataset. Descriptive statistics were first calculated to summarize the central tendency, spread, and shape of the data. For the Active column, the mean value is 5,736 active players, with a standard deviation of 26,584. This large standard deviation suggests significant variation in the number of active players across games, with some games attracting millions of players while others have relatively small player bases. Similarly, the Visits column has an average of approximately 636 million total visits, but the spread is much larger, with values ranging from just 28,198 to a staggering 55.6 billion visits. The high standard deviation of over 2.9 billion visits highlights that some games, particularly the top-ranked ones, have an overwhelming majority of visits compared to others. The Likes and Dislikes columns follow a similar distribution, with an average of 323,743 likes and a standard deviation of 763,031, showing that while the majority of games have a moderate number of likes, some games are heavily favored by users, while others have significant dislike counts. The Rating column has a mean of 84.1, with a standard deviation of 11.12, indicating that most games enjoy positive reception, although there are outliers with significantly lower ratings (minimum of 25.55).

Figure 2 presents a series of histograms displaying the distributions of key numerical features in the dataset, which include Rank, Active, Visits, Favourites, Likes, Dislikes, and Rating. Each subplot represents one of these features and provides insights into their distribution across the dataset. The Rank variable, which represents the position of the games based on current players, appears relatively uniform across the first 1000 games, with each rank appearing once. This suggests that the ranks are not highly concentrated, but the distribution is sparse due to the large spread of player counts within these ranks. The histogram for Active players shows a right-skewed distribution, meaning that most games have fewer active players, while a small number of games (mostly in the top ranks) have a very large number of active players. This reflects the nature of Roblox, where a few games dominate player engagement. Similar to Active players, the Visits distribution is highly right-skewed, with most games having relatively few visits, while a few games (top-ranking ones) have extraordinarily large numbers of visits (up to several billion). This indicates the highly unequal distribution of engagement across Roblox games.

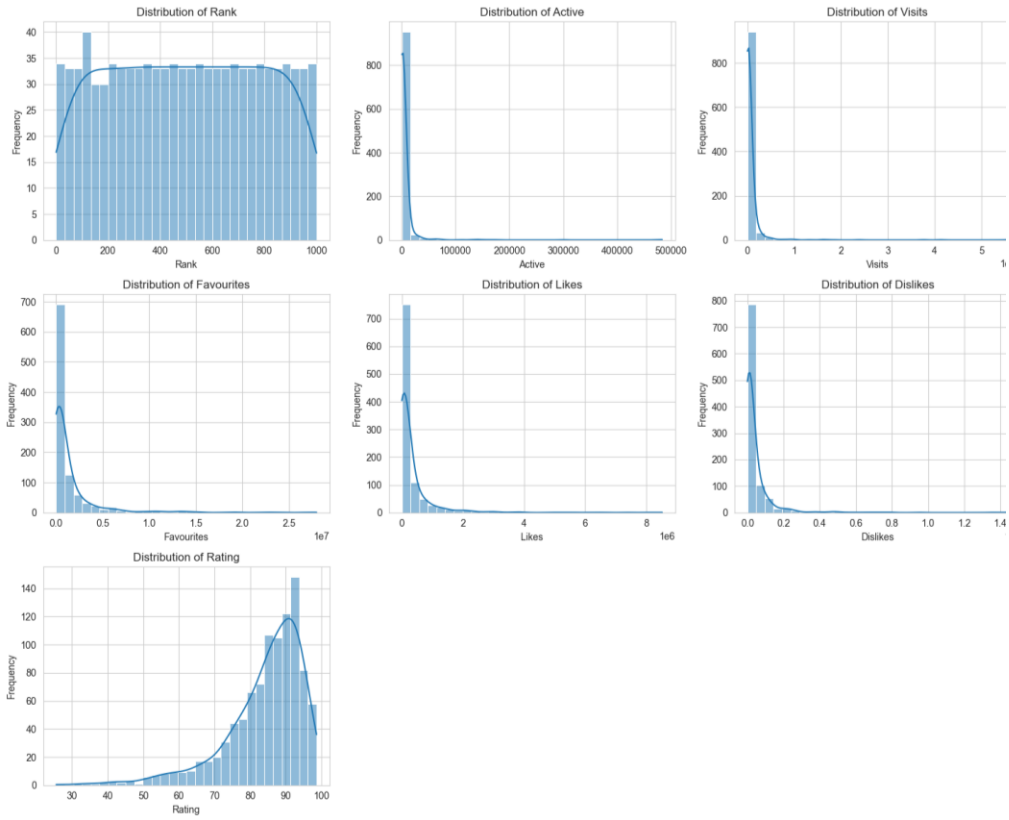

Figure 2 Histograms of Numerical Features

The Favourites feature also displays a right-skewed distribution, with most games having fewer favourites, while top games accumulate millions of favourites. The pattern indicates that while some games attract a large loyal following, the majority of games have a smaller, less engaged player base. Both the Likes and Dislikes distributions show a similar right-skewed pattern, indicating that a small number of games accumulate the majority of likes and dislikes. The spread of likes and dislikes further highlights that while most games receive a modest number of likes, some games garner millions, contributing to their visibility. The Rating distribution is more concentrated toward the higher end, with most games having ratings between 80% and 95%. This suggests that most games are well-received, with a relatively small number of outliers having lower ratings.

Figure 3 is a correlation matrix showing the relationships between the numerical features of the dataset. Each cell in the matrix represents the correlation between two features, with values ranging from -1 to +1. The strongest positive correlation is between Active players and Visits (correlation of 0.75). This indicates that games with more active players tend to accumulate more visits, which is intuitive since higher engagement likely leads to more players returning to play, increasing the total visits. There is also a strong positive correlation between Likes and Visits (correlation of 0.77), suggesting that games with higher user satisfaction, as measured by likes, tend to have more total visits. This relationship is crucial as it emphasizes the importance of positive player feedback in driving game success.

Figure 3 Correlation Matrix of Numerical Features

There is a high negative correlation between Likes and Dislikes (correlation of -0.87). This implies that games with more likes generally receive fewer dislikes, further supporting the idea that a game's overall sentiment, driven by likes, directly impacts its popularity. Rating shows weaker correlations with other features. For example, it has a minimal correlation with Likes (0.18), Dislikes (0.02), and Visits (0.04), suggesting that while ratings are important, they do not have as strong a direct relationship with game success as metrics like Active players and Likes. Favourites and Rating are also weakly correlated, indicating that while players may favourite a game, it does not necessarily reflect its quality in terms of ratings. Overall, the correlation matrix underscores the significant role of Active players, Likes, and Dislikes in predicting Visits, with Rating playing a comparatively minor role. This suggests that while player engagement is the most important factor in determining a game's success, the game's rating and user satisfaction also contribute, though to a lesser extent.

Figure 4 presents a scatter plot showing the relationship between Active players and Total visits for each Roblox game, using a logarithmic scale on both the x-axis and y-axis. This log scale is applied due to the wide range of values in these variables, particularly the Total visits, which can range from millions to billions. The plot reveals a clear positive correlation between the number of active players and the total visits a game receives. As expected, games with higher Active players tend to have higher Total visits, reflecting the increased engagement and retention that often accompanies more active users. However, there is significant variance in the data, especially at the higher end of the spectrum. Some games with high active player counts (ranging from tens of thousands to hundreds of thousands) still show substantial variance in visits, which suggests that factors beyond just the active player count contribute to a

game's popularity, such as game content, frequency of updates, and player retention.

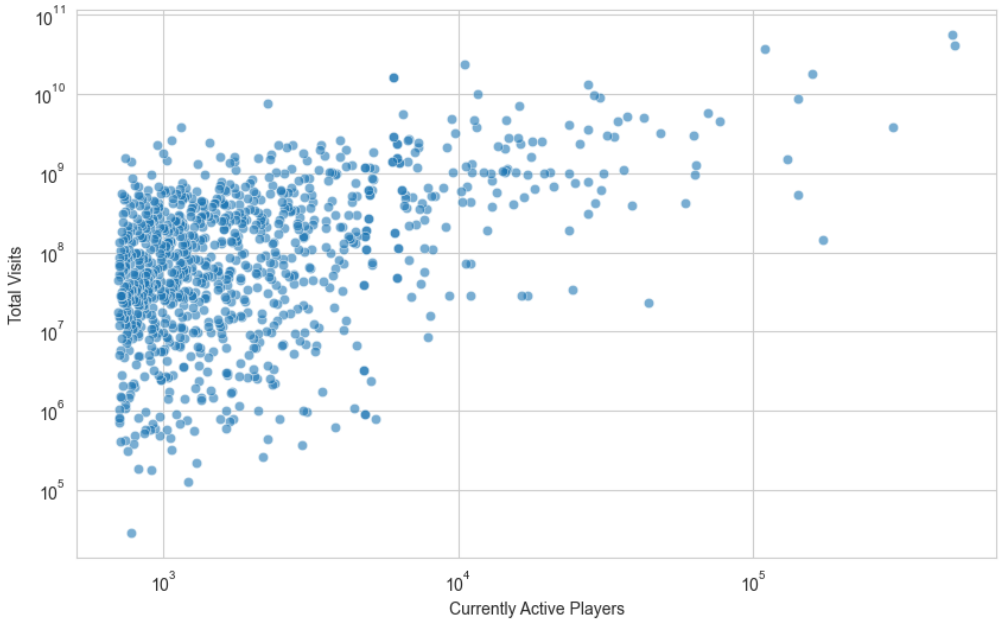

Figure 4 Active Players vs. Total Visits (Log Scale)

The outliers visible in the top-right corner of the plot represent games with millions of active players and billions of visits, highlighting the few dominant games that significantly outpace others in terms of player engagement and traffic. These findings reinforce the idea that active player count is a key indicator of a game's success but does not fully account for the variance seen in Roblox game popularity.

Figure 5 shows the relationship between Game Rating and Total visits, again using a logarithmic scale for both axes. The Game Rating, represented as a percentage, ranges from 30% to 100%. The scatter plot shows that there is a slight positive correlation between the Game Rating and Total visits, where games with higher ratings tend to have more visits. However, the correlation is weaker compared to Active players. This suggests that while player satisfaction, as measured by ratings, plays a role in the game's success, it is not the sole determinant of its popularity. Many games with ratings in the 80% to 90% range still receive vastly different numbers of visits, indicating that other factors, such as game mechanics, promotion, or genre, influence a game's ability to attract players. The plot also shows that low-rated games (below 50%) tend to have fewer visits, confirming that negative player sentiment (captured through lower ratings) typically leads to reduced popularity. However, there are some outliers where games with low ratings still manage to attract a substantial number of visits, which could be explained by factors like viral trends, the game's niche appeal, or its ability to engage players despite its negative feedback. The relationship between rating and visits is thus complex, suggesting that while ratings contribute to a game's overall success, they are not the only determinant, and high ratings alone do not guarantee a high number of visits.

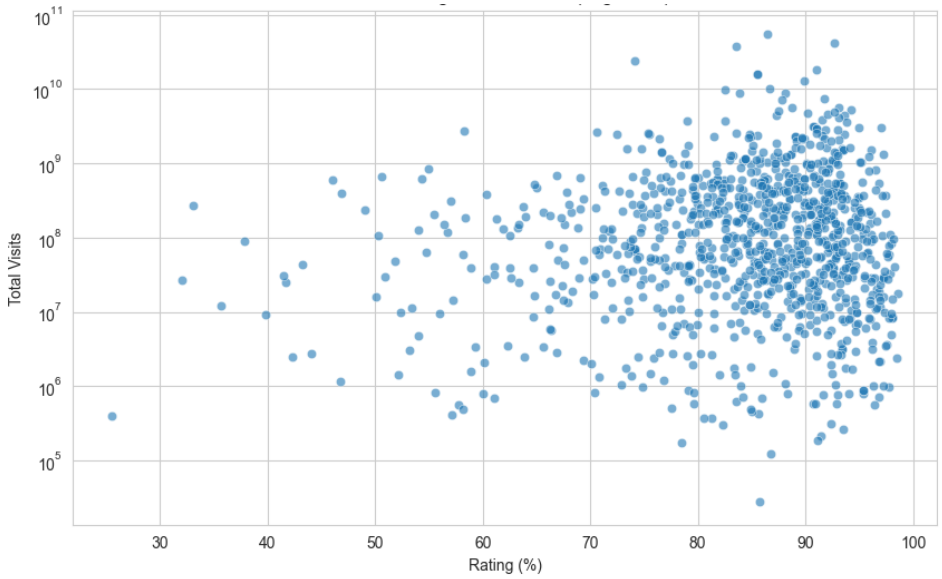

Figure 5 **Game Rating vs. Total Visits (Log Scale)**

Evaluation of Random Forest Regressor Performance

The training process was completed quickly, with the model taking just 0.33 seconds to train on the 800 training samples. During the training, we also calculated the Out-of-Bag (OOB) score, which is a built-in feature of Random Forest that estimates the model's generalization error without the need for a separate validation set. The OOB score of 0.5676 indicates that the model performs reasonably well, with a moderate level of predictive power. This score suggests that the model has learned useful patterns from the training data, but there is still room for improvement in terms of accuracy. After training and saving the model, the next step was to evaluate its performance on the test set. To assess how well the model predicted the Visits (target variable) for the test set, we used several evaluation metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2).

The Mean Squared Error (MSE) was found to be an extremely large number, 544,899,627,096,328,000.00. This large value is mainly due to the presence of extremely high values in the target variable, Visits, which can range from tens of thousands to billions. The MSE metric is sensitive to large errors, especially when the target variable has large values, so it needs to be interpreted carefully in such contexts. To better interpret the error, we calculated the Root Mean Squared Error (RMSE), which is the square root of MSE and provides an error metric that is on the same scale as the target variable. The RMSE was found to be 738,173,168.77, which means that, on average, the model's predictions are off by approximately 738 million visits. This still indicates significant error but gives a more interpretable measure of prediction accuracy. The Mean Absolute Error (MAE) was 208,532,159.38, which tells us that, on average, the model's predictions are off by about 208.5 million visits. This metric reflects the average magnitude of the errors, without considering their direction, and provides a more straightforward indication of prediction accuracy than MSE or RMSE. The R-squared (R^2) score was calculated to be 0.7814, which means that the model explains approximately 78.14% of the variance in the Visits data. This is a relatively high R^2 value, suggesting that the model has good predictive power

and captures the majority of the trends in the data. While not perfect, this R^2 value indicates that the model is highly useful for predicting game popularity in Roblox based on the available features. To visualize the model's performance, a prediction vs. actual plot was generated, which compares the predicted values of Visits to the actual values from the test set.

Figure 6 presents a scatter plot comparing the actual versus predicted Visits for the games in the test set. Both the x-axis (representing actual visits) and the y-axis (representing predicted visits) are in logarithmic scale to handle the wide range of values, especially considering the skewed nature of the data. The dashed red line represents the ideal scenario where the predicted values exactly match the actual values (i.e., y = x). The data points in the plot show how well the model performed: points that are close to the red dashed line indicate accurate predictions, while points that are further away represent larger prediction errors.

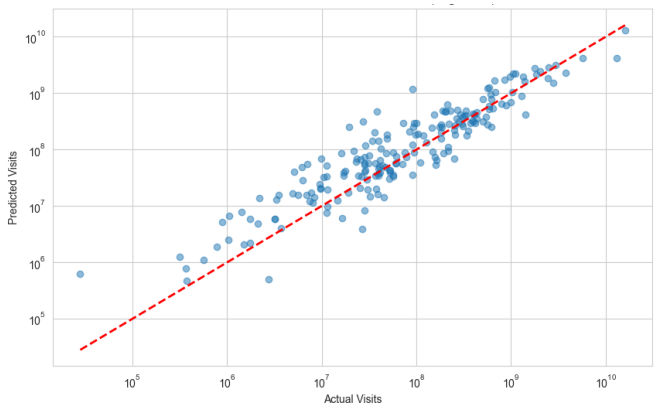

Figure 6 Actual vs. Predicted Visits on Test Set (Log Scale)

The plot demonstrates a strong positive correlation between the actual and predicted visits, indicating that the model is capable of predicting visits reasonably well, particularly for games that have moderate visit counts. However, there is still noticeable scatter in the data, especially for games with extreme values of visits, suggesting that the model struggles more with games that have very high or low visit counts. Overall, the plot confirms that the Random Forest model performs well for most games but may still face challenges when predicting extreme values in the test set.

Feature Importance Analysis

A key advantage of Random Forest models is their ability to assess the importance of each feature in the prediction process. The feature importance scores indicate how much each feature contributes to the model's ability to predict the target variable, Visits. The results showed that Dislikes was the most important feature, with an importance score of 0.48, meaning that it had the greatest influence on the model's predictions. This suggests that user dissatisfaction, as captured by the number of dislikes, plays a significant role in determining a game's popularity on Roblox.

Likes came in second, with an importance score of 0.25, indicating that games with more likes are more likely to attract higher visits. Active players also contributed to the model, with an importance score of 0.22, highlighting the importance of player engagement in predicting game success. The features

Favourites and Rating had lower importance scores of 0.03 and 0.02, respectively, indicating that while these features contribute to the model, they are less influential compared to user engagement metrics like Likes and Dislikes. To visualize these feature importances, a feature importance plot was created, showing the relative importance of each feature in predicting Visits.

Figure 7 displays a bar plot of the feature importance scores for the variables used in predicting Total Visits. The plot highlights the relative importance of each feature, with Dislikes having the highest importance score of 0.48. This result suggests that Dislikes play a significant role in determining game popularity on Roblox, which aligns with earlier findings in the study. Following Dislikes, Likes (with an importance score of 0.25) and Active players (0.22) are also important features for predicting visits, underscoring the role of player sentiment and engagement in driving game success. Favourites and Rating are the least influential features in the model, with scores of 0.03 and 0.02, respectively. This suggests that while these features contribute to the model's predictions, their impact on the overall success of a game is comparatively smaller than engagement metrics like Likes and Dislikes.

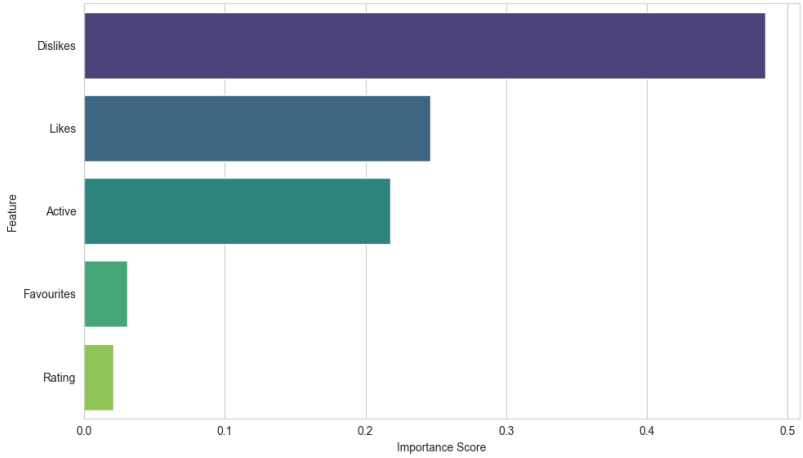

Figure 7 Feature Importance for Predicting Roblox Game Visits

The feature importance plot provides valuable insights into which variables matter most when predicting Roblox game popularity, guiding developers to focus on improving player engagement through positive feedback and active player engagement, rather than relying heavily on features like ratings or favourites. This analysis further emphasizes the utility of Random Forest in identifying the most important factors driving game popularity, allowing developers to prioritize the right features for optimization in the highly competitive Roblox environment.

Comparison with Previous Research

The application of data mining to analyze Roblox game performance has been widely explored in recent research, with a particular focus on user interactions, engagement levels, and key success indicators like likes, dislikes, and active players [14]. Previous studies have emphasized the importance of these metrics in predicting game success, with much of the focus being on analyzing player feedback (likes/dislikes) and the number of active players as primary indicators of a game's popularity. These metrics have been shown to be strongly correlated with a game's overall visibility, attractiveness, and potential for long-

term success. In line with this, our study also highlighted the significant role of player engagement metrics—Likes, Dislikes, and Active Players—in determining the success of Roblox games, confirming that these features are crucial predictors of game popularity, as evidenced by our Random Forest model's findings.

A key difference between our study and previous research lies in the modeling approach. While earlier studies often used simpler regression models or basic machine learning techniques to predict game success, our study leverages the Random Forest Regressor, an ensemble learning algorithm well-suited for capturing complex relationships between variables. The use of Random Forest aligns with findings from [24], who noted the algorithm's effectiveness in handling non-linear interactions and its resilience to overfitting. Previous studies found success with simpler models, but Random Forest excels at managing the interactions between multiple game features, which often involve non-linear relationships (e.g., the combined effects of Likes and Dislikes on game popularity). Our OOB score of 0.5676 reflects this improvement, showing that the Random Forest model was able to learn complex patterns in the dataset, despite some room for improvement in terms of prediction accuracy.

Moreover, the feature importance analysis in our study further sets our work apart from prior research. While earlier studies identified basic metrics such as Active Players and Likes as key predictors of game success, our model highlights the unexpected yet significant role of Dislikes, with an importance score of 0.48. This finding builds on the research by [25], who emphasized the importance of player sentiment in shaping game popularity. Our analysis demonstrates that negative feedback, reflected by Dislikes, plays a surprisingly large role in determining a game's success, suggesting that player dissatisfaction can be a significant factor influencing a game's visibility and retention on platforms like Roblox.

In terms of model evaluation, our study's R-squared value of 0.7814 is comparable to other Random Forest-based studies on game prediction, such as [24], where similar models achieved R^2 scores ranging from 0.7 to 0.8. While this suggests that our model does a good job of explaining the variance in Visits, the relatively large Mean Squared Error (MSE) value reflects the challenges of predicting games with extreme popularity. This issue of variance in game popularity has been acknowledged in previous research [22], where models struggled to predict the success of extremely popular games accurately due to their outlier status in the dataset. Despite these challenges, the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) metrics provide more interpretable measures, which show that, on average, the model's predictions are off by around 738 million visits, a relatively high error when considering the large range of values in the Visits feature.

This study builds on previous research by applying a more advanced machine learning technique, Random Forest, to predict Roblox game popularity. The combination of Likes, Dislikes, Active Players, and Rating as features for prediction aligns with existing literature but also introduces new findings, particularly regarding the importance of Dislikes. While there are still challenges, particularly with predicting extremely popular games, our study contributes to the growing body of work on game prediction in the metaverse, providing a more nuanced and powerful tool for game developers to optimize their offerings for

enhanced player engagement and success.

Conclusion

This study demonstrates the effectiveness of the Random Forest algorithm in predicting the popularity of Roblox games based on key player engagement metrics such as Active players, Likes, Dislikes, Favourites, and Rating. The model achieved a strong R-squared value of 0.7814, indicating its ability to explain a significant portion of the variance in game Visits, and provides valuable insights into how various features influence game success. Our analysis highlighted the crucial role of both positive and negative player feedback, particularly Dislikes, which was found to be the most important feature in predicting game popularity. These findings contribute to the growing body of research on metaverse game dynamics, offering a data-driven tool that can help developers optimize their games for greater engagement and success on platforms like Roblox.

Despite these promising results, the study has some limitations. The reliance on publicly available data means that certain player behavior patterns, which could provide deeper insights into game success, were not captured. Additionally, the dataset may have been subject to data sparsity, particularly for less popular games, which could affect the model's accuracy for games with lower player engagement. Future research could address these limitations by incorporating more comprehensive player behavior data, such as session duration and in-game interactions, or by experimenting with other machine learning algorithms, such as Gradient Boosting or Neural Networks, to further improve prediction accuracy and capture more complex relationships in the data.

Declarations

Author Contributions

Author Contributions: Conceptualization, E.M.A. and M.A.O.B.; Methodology, E.M.A. and M.A.O.B.; Software, E.M.A. and M.A.O.B.; Validation, E.M.A. and M.A.O.B.; Formal Analysis, E.M.A.; Investigation, E.M.A. and M.A.O.B.; Resources, M.A.O.B.; Data Curation, E.M.A.; Writing—Original Draft Preparation, E.M.A.; Writing—Review and Editing, M.A.O.B.; Visualization, E.M.A. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

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

Funding

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

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

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] M. Traub and M. Weinberger, "APIs in the Metaverse—A Systematic Evaluation," *Virtual Worlds*, vol. 3, no. 2, pp. 1–23, 2024, doi: 10.3390/virtualworlds3020008.

[2] M. Weinberger, "What Is Metaverse?—A Definition Based on Qualitative Meta-Synthesis," *Future Internet*, vol. 14, no. 11, pp. 1–14, 2022, doi: 10.3390/fi14110310.

[3] S. E. Bibri, "The Metaverse as a Virtual Model of Platform Urbanism: Its Converging AIoT, XReality, Neurotech, and Nanobiotech and Their Applications, Challenges, and Risks," *Smart Cities*, vol. 6, no. 3, pp. 1231–1260, 2023, doi: 10.3390/smartcities6030065.

[4] S. Mansoor, S. M. Rahman, and J. Bowden, "Purchase Spillovers From the Metaverse to the Real World: The Roles of Social Presence, Trialability, and Customer Experience," *Journal of Consumer Behaviour*, vol. 23, no. 1, pp. 37–52, 2024, doi: 10.1002/cb.2353.

[5] J. Han, G. Liu, and Y. Gao, "Learners in the Metaverse: A Systematic Review on the Use of Roblox in Learning," *Education Sciences*, vol. 13, no. 3, pp. 1–17, 2023, doi: 10.3390/educsci13030296.

[6] S. Roohi, C. Guckelsberger, A. Relas, H. Heiskanen, J. Takatalo, and P. Hämäläinen, "Predicting Game Difficulty and Engagement Using AI Players," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. 10, pp. 1–22, 2021, doi: 10.1145/3474658.

[7] K. Mustač, K. Bačić, L. Skorin-Kapov, and M. Sužnjević, "Predicting Player Churn of a Free-to-Play Mobile Video Game Using Supervised Machine Learning," *Applied Sciences*, vol. 12, no. 6, pp. 1–18, 2022, doi: 10.3390/app12062795.

[8] Y. Zhao, S. Yang, M. Shum, and S. Dutta, "A Dynamic Model of Player Level-Progression Decisions in Online Gaming," *Management Science*, vol. 68, no. 3, pp. 1491–1509, 2022, doi: 10.1287/mnsc.2021.4255.

[9] X. Liu, K. Kasmarik, and H. A. Abbass, "Assessing Player Profiles of Achievement, Affiliation, and Power Motivation Using Electroencephalography," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 52, no. 11, pp. 6910–6922, 2022, doi: 10.1109/TSMC.2021.3073216.

[10] N. Ballou, M. Vuorre, T. Hakman, K. Magnusson, and A. K. Przybylski, "Perceived Value of Video Games, but Not Hours Played, Predicts Mental Well-Being in Casual Adult Nintendo Players," *R. Soc. Open Sci.*, vol. 12, no. 1, pp. 1–12, 2025, doi: 10.1098/rsos.241174.

[11] A. Guitart, P. P. Chen, and Á. Periáñez, "The Winning Solution to the IEEE CIG 2017 Game Data Mining Competition," *Mach. Learn. Knowl. Extr.*, vol. 1, no. 1, pp. 103–118, 2018, doi: 10.3390/make1010016.

[12] S. Bunian, A. Canossa, R. Colvin, and M. S. El-Nasr, "Modeling Individual

Differences in Game Behavior Using HMM," *Proc. AAAI Conf. Artif. Intell. Interact. Digit. Entertain.*, vol. 2021, no. 9, pp. 1–12, 2021, doi: 10.1609/aiide.v13i1.12942.

[13] J. Zhou, "Studies Advanced in Artificial Intelligence Based Game," *Applied Computing and Engineering*, vol. 2023, no. 2, pp. 1–12, 2023, doi: 10.54254/2755-2721/8/20230255.

[14] K. Alhasan, K. Alhasan, and S. A. Hashimi, "Roblox in Higher Education," *Int. J. Emerg. Technol. Learn. (iJET)*, vol. 18, no. 19, pp. 109–123, 2023, doi: 10.3991/ijet.v18i19.43133.

[15] L. Molyneux, K. Vasudevan, and H. G. de Zúñiga, "Gaming Social Capital: Exploring Civic Value in Multiplayer Video Games," *J. Comput.-Mediat. Commun.*, vol. 20, no. 4, pp. 381–399, 2015, doi: 10.1111/jcc4.12123.

[16] J. Yu, W. Ma, J. Moon, and A. R. Denham, "Developing a Stealth Assessment System Using a Continuous Conjunctive Model," *J. Learn. Anal.*, vol. 9, no. 3, pp. 63–82, 2022, doi: 10.18608/jla.2022.7639.

[17] Y. Su, P. Backlund, and H. Engström, "Business Intelligence Challenges for Independent Game Publishing," *Int. J. Comput. Games Technol.*, vol. 2020, no. 6, pp. 1–12, 2020, doi: 10.1155/2020/5395187.

[18] C. Alonso-Fernández, A. Calvo-Morata, M. Freire, I. Martínez-Ortiz, and B. Fernández-Manjón, "Game Learning Analytics," *J. Learn. Anal.*, vol. 9, no. 3, pp. 33–55, 2022, doi: 10.18608/jla.2022.7633.

[19] Y. K. Oh, J. B. Yi, and J.-D. Kim, "What Enhances or Worsens the User-Generated Metaverse Experience? An Application of BERTopic to Roblox User eWOM," *Internet Research*, vol. 33, no. 2, pp. 637–658, 2023, doi: 10.1108/intr-03-2022-0178.

[20] C. Alonso-Fernández, A. Calvo-Morata, M. Freire, I. Martínez-Ortiz, and B. Fernández-Manjón, "Applications of Data Science to Game Learning Analytics Data: A Systematic Literature Review," *Computers & Education*, vol. 141, no. 11, pp. 1–15, 2019, doi: 10.1016/j.compedu.2019.103612.

[21] C. Meier, J. L. Saorín Pérez, A. B. de León, and A. G. Cobos, "Using the Roblox Video Game Engine for Creating Virtual Tours and Learning About the Sculptural Heritage," *Int. J. Emerg. Technol. Learn. (iJET)*, vol. 15, no. 20, pp. 86–100, 2020, doi: 10.3991/ijet.v15i20.16535.

[22] X. Zhang, Y. Chen, L. Hu, and Y. Wang, "The Metaverse in Education: Definition, Framework, Features, Potential Applications, Challenges, and Future Research Topics," *Front. Psychol.*, vol. 13, no. 10, pp. 1–17, 2022, doi: 10.3389/fpsyg.2022.1016300.

[23] S. Virani and S. Rautela, "Metaverse and Education: Identifying Key Themes and Future Research Trajectories," *Int. J. Inf. Learn. Technol.*, vol. 41, no. 5, pp. 659–674, 2024, doi: 10.1108/ijilt-05-2024-0083.

[24] J. Junaidi, A. Julianto, N. Anwar, S. Safrizal, H. L. H. Warnars, and K. Hashimoto, "Perfecting a Video Game With Game Metrics," *Telkomnika Telecommun. Comput. Electron. Control*, vol. 16, no. 3, pp. 1203–1211, 2018, doi: 10.12928/telkomnika.v16i3.7209.

[25] J. Hodge, J. Wilson, R. Cooper, P. Price, D. Reeves, and E. Johnson, "How the Business Model of Customizable Card Games Influences Player Engagement," *IEEE Trans. Games*, vol. 11, no. 3, pp. 272–283, 2019, doi: 10.1109/TG.2018.2803843.