Predicting Player Engagement Levels in Online Gaming Platforms Using Machine Learning and Behavioral Interaction Data Analysis

Gokulnath Anandakumar[1], R. Karthick Manoj[2,*]

[1,2]Department of Electrical and Electronics Engineering AMET University Kanathur, Chennai, India

ABSTRACT

Online gaming platforms have become an important component of digital society, generating large volumes of behavioral data that can be analyzed to understand player engagement patterns. This study aims to predict player engagement levels in online gaming environments using machine learning techniques. The dataset used in this research consists of 40,034 player records with 13 variables that include demographic characteristics and gameplay behavior indicators such as playtime hours, number of weekly sessions, average session duration, player level, and achievements unlocked. The target variable classifies engagement into three categories: Low, Medium, and High engagement. Several machine learning models were evaluated, including Gradient Boosting, Neural Network, and XGBoost, and their performance was assessed using accuracy, precision, recall, and F1 score. The experimental results show that XGBoost achieved the best predictive performance with an accuracy of 91.4 percent and an F1 score of 0.9137. Cross validation results further confirm the robustness of the model with a mean F1 score of 0.9163. Feature importance analysis reveals that SessionsPerWeek and AvgSessionDurationMinutes are the most influential factors in predicting player engagement, followed by achievements unlocked and player level. These findings indicate that engagement in online gaming platforms is primarily influenced by behavioral interaction patterns rather than demographic characteristics. The results highlight the potential of machine learning methods for analyzing human engagement behavior in digital environments and provide insights that may support the development of more engaging online gaming platforms.

Keywords Machine Learning, Online Gaming, Player Engagement, Digital Behavior, Predictive Modeling

Submitted 21 June 2026
Accepted 3 August 2026
Published 1 September 2026

Corresponding author
R. Karthick Manoj,
karthickmanoj.r@gmail.com

Additional Information and Declarations can be found on page 208

INTRODUCTION

The rapid growth of digital technologies has transformed how individuals interact, communicate, and participate in entertainment activities. One of the most prominent developments in this digital environment is the rise of online gaming platforms, which attract millions of players worldwide [1]. Online games are no longer limited to entertainment but have become important social and interactive spaces where users engage with digital systems and other players. These platforms generate large volumes of behavioral data that reflect how users interact with digital environments. Understanding these behavioral patterns is increasingly important for researchers and developers who seek to improve user experience and maintain player engagement within gaming platforms.

Player engagement has become a central concept in online gaming research because it reflects the intensity and frequency of interaction between users and

How to cite this article: G. Anandakumar and R. K. Manoj, "Predicting Player Engagement Levels in Online Gaming Platforms Using Machine Learning and Behavioral Interaction Data Analysis," *J. Digit. Soc.*, vol. 2, no. 3, pp. 192-210, 2026.

digital systems [2]. Engagement can be influenced by various factors such as gameplay duration, session frequency, player progression, and in game achievements. High engagement levels often indicate strong user involvement and long-term platform participation, while low engagement may suggest declining interest or reduced interaction. As a result, analyzing engagement patterns provides valuable insights into how users behave in digital environments and how gaming platforms can design experiences that sustain player interest.

Recent studies have applied machine learning techniques to analyze player behavior and predict engagement outcomes based on gameplay data [3]. These approaches allow researchers to identify complex patterns within large datasets that are difficult to detect using traditional statistical methods. Advanced machine learning models such as gradient boosting, neural networks, and ensemble learning methods have shown promising performance in classification and prediction tasks related to player behavior. These models are capable of processing multiple behavioral variables simultaneously and learning non linear relationships between gameplay activity and engagement levels.

Despite these advancements, several research gaps remain in the current literature. First, many previous studies focus primarily on specific aspects of gaming behavior such as player retention, addiction risk, or in game purchases rather than providing a comprehensive prediction of engagement levels based on multiple behavioral indicators [4]. Second, some existing research relies on relatively small datasets, which may limit the generalizability of the findings. Third, there is still limited research that integrates both demographic characteristics and gameplay behavior variables to predict engagement levels within large scale online gaming datasets. As a result, further investigation is needed to better understand how different types of player data contribute to engagement prediction.

The state of the art in this area increasingly emphasizes the use of ensemble machine learning models for behavioral prediction tasks. Techniques such as XGBoost and other gradient boosting methods have demonstrated strong performance in classification problems involving complex datasets because they can capture interactions between multiple features while reducing prediction errors through iterative learning [5]. These models have been widely applied in domains such as recommendation systems, user behavior analysis, and digital platform analytics. However, their application in predicting engagement levels using comprehensive gaming behavior datasets remains relatively underexplored.

Therefore, this study aims to predict player engagement levels in online gaming environments using machine learning techniques by analyzing both demographic attributes and gameplay behavior variables. Using a dataset consisting of more than forty thousand player records, this research evaluates several machine learning models to determine which approach provides the best predictive performance. In addition, the study identifies the most influential behavioral factors that contribute to engagement prediction. The findings of this research are expected to provide deeper insights into digital user behavior and contribute to the development of more engaging online gaming platforms.

Literature Review and Related Works

Online gaming has become one of the most rapidly expanding sectors within the digital entertainment industry, attracting millions of users who interact within virtual environments on a daily basis. These platforms generate extensive behavioral data that can be analyzed to understand patterns of user interaction and engagement in digital environments. Player engagement is widely considered a critical factor for the sustainability and success of online gaming platforms because it reflects the level of interaction between users and the gaming system. High engagement is often associated with increased player retention, longer gameplay sessions, and stronger participation in gaming communities [6]. As a result, understanding engagement patterns has become an important research focus in both digital behavior studies and gaming analytics.

Previous studies have explored various aspects of player engagement and behavioral analysis in online gaming environments. Several works have shown that player behavior data such as gameplay duration, session frequency, and in game achievements can provide meaningful insights into user engagement and gaming activity patterns [7]. Behavioral data generated from online games allows researchers to examine how players interact with digital systems and how their actions reflect different levels of engagement and participation. These behavioral indicators are often used to identify player preferences, predict user activity, and analyze gameplay dynamics within digital platforms [8].

With the increasing availability of large scale gaming datasets, machine learning techniques have become widely used for analyzing player behavior and predicting engagement outcomes. Machine learning methods are capable of identifying complex relationships within data and can model nonlinear patterns that traditional statistical approaches may fail to capture. Several studies have applied classification algorithms to predict player engagement levels based on gameplay behavior, demonstrating that machine learning models can effectively distinguish between different types of player activity patterns [9]. These predictive models enable developers and researchers to better understand player behavior and design systems that maintain user interest.

Ensemble learning methods have recently gained attention in gaming analytics due to their strong predictive performance in classification tasks. Models based on gradient boosting and decision tree ensembles are particularly effective in handling structured datasets that contain both numerical and categorical variables. These algorithms can combine multiple weak learners into a single strong predictive model, which improves accuracy and reduces prediction errors during the training process [10]. In addition, ensemble learning techniques are capable of handling complex interactions between variables, making them suitable for analyzing gaming behavior data that contains multiple behavioral indicators.

Deep learning approaches have also been explored for modeling player behavior and engagement patterns. Neural network architectures, including multilayer perceptrons, have been applied to analyze gaming data and predict various outcomes such as player retention, engagement intensity, and gameplay progression [11]. These models are capable of learning hierarchical representations of data, allowing them to capture complex relationships between gameplay features. However, deep learning models often require large datasets and extensive computational resources, which may limit their applicability in certain scenarios.

Another important area of research focuses on identifying key behavioral factors that influence player engagement. Previous studies have shown that variables such as gameplay frequency, session duration, and player progression levels can significantly influence engagement outcomes [12]. Frequent gameplay sessions and longer session durations often indicate stronger player involvement and higher engagement levels. Similarly, achievements and progression systems in games can motivate players to continue participating in gaming activities, thereby increasing engagement [13].

Recent research has also highlighted the importance of analyzing digital user behavior within broader digital society contexts. Online gaming platforms serve not only as entertainment systems but also as social and interactive environments where users engage in collaborative and competitive activities. The analysis of gaming behavior can therefore provide insights into how individuals interact with digital technologies and how engagement patterns evolve within virtual environments [14]. Understanding these patterns can contribute to the development of more effective digital platforms that promote sustained user participation.

Despite the growing body of research in gaming analytics and engagement prediction, several challenges remain in this field. Many previous studies focus on specific behavioral indicators or narrow aspects of gaming behavior rather than providing a comprehensive analysis of engagement prediction using multiple features simultaneously [15]. Additionally, some existing studies rely on limited datasets that may not fully represent the diversity of player behaviors observed in large scale gaming platforms [16]. These limitations highlight the need for further research that utilizes larger datasets and integrates multiple behavioral variables for engagement prediction.

Machine learning models such as gradient boosting and extreme gradient boosting have recently emerged as state of the art techniques for classification tasks involving structured data. These models are designed to improve predictive accuracy by iteratively minimizing prediction errors through gradient optimization. Their ability to capture nonlinear relationships and interactions between variables makes them particularly suitable for analyzing behavioral data generated by online gaming platforms [17]. Several studies have demonstrated that gradient boosting models outperform traditional machine learning algorithms in classification problems involving user behavior prediction [18].

In addition, feature importance analysis has become an important tool for interpreting machine learning models and identifying the variables that contribute most to prediction outcomes. By analyzing feature importance, researchers can gain insights into which gameplay behaviors have the strongest influence on engagement levels. This approach not only improves model interpretability but also provides valuable information about the behavioral patterns that drive user interaction in digital environments [19].

Based on these developments, the application of machine learning techniques to predict engagement levels using comprehensive gameplay behavior datasets represents a promising research direction. By combining behavioral variables, demographic characteristics, and advanced machine learning models, it is possible to develop predictive systems that better understand player interaction patterns within online gaming platforms [20]. This study contributes to the existing literature by applying multiple machine learning models to a large scale gaming dataset and identifying the key behavioral factors that influence player engagement levels in digital gaming environments.

Methodology

This study applies a machine learning based approach to predict player engagement levels in online gaming environments. The methodology consists of several stages, including data collection, data preprocessing, exploratory data analysis, model development, model evaluation, and feature importance analysis. These steps are designed to identify behavioral factors that influence player engagement and determine the most effective machine learning model for engagement prediction.

The overall research workflow is illustrated in figure 1, which shows the main stages of the research process starting from data collection to model evaluation.

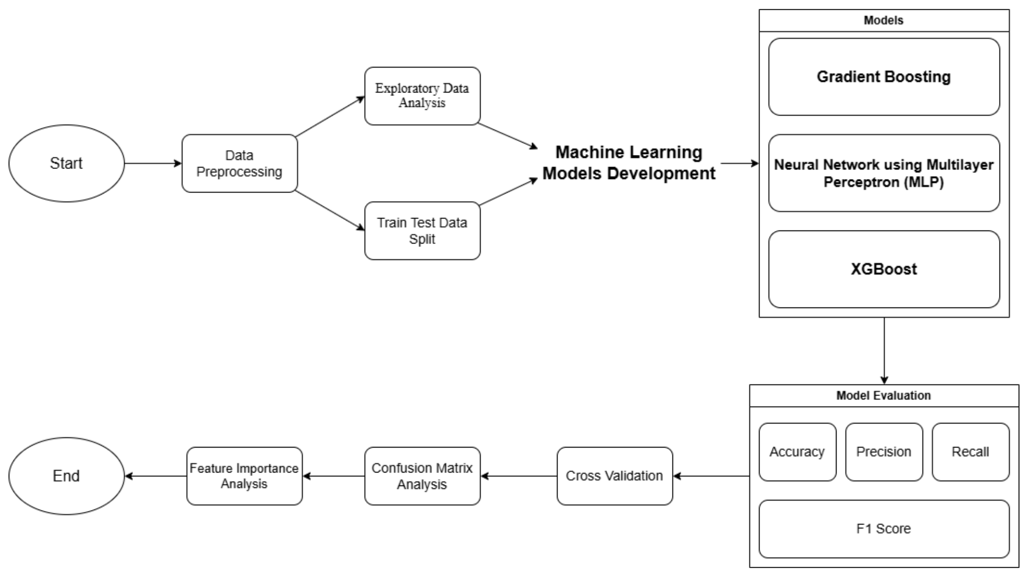


Figure 1 Research methodology workflow.

Dataset Description

The dataset used in this study consists of 40,034 player records with 13 attributes that represent demographic information, gameplay behavior, and engagement classification within online gaming environments. The dataset includes both numerical variables, such as age, playtime hours, sessions per week, average session duration, player level, and achievements unlocked, as well as categorical variables, including gender, location, game genre, and game difficulty. These variables collectively capture various aspects of player interaction patterns and behavioral activities during gameplay. The target variable in this dataset is EngagementLevel, which is classified into three categories, namely Low, Medium, and High engagement, representing different levels of player involvement in the gaming platform. All remaining variables are treated as input features for the machine learning modeling process. A detailed description of these variables and their corresponding data types is presented in table 1.

Table 1 Dataset variables description

Variable	Type	Description
Age	Numerical	Player age
Gender	Categorical	Player gender
Location	Categorical	Player region
GameGenre	Categorical	Type of game played
PlayTimeHours	Numerical	Total hours spent playing
InGamePurchases	Numerical	Purchase activity indicator
GameDifficulty	Categorical	Game difficulty level
SessionsPerWeek	Numerical	Number of weekly sessions
AvgSessionDurationMinutes	Numerical	Average session duration
PlayerLevel	Numerical	Player progression level
AchievementsUnlocked	Numerical	Number of achievements
EngagementLevel	Categorical	Player engagement classification

Before the modeling process, the dataset was inspected to ensure data quality. The analysis showed that the dataset contained no missing values, which allowed the modeling process to proceed without additional data imputation.

Data Preprocessing

Data preprocessing was performed to prepare the dataset for machine learning analysis by transforming and organizing the variables into a suitable format for modeling. The PlayerID variable was first removed from the dataset because it functions only as an identifier and does not contribute to the predictive process. After removing this variable, the dataset was separated into input features (X) and the target variable (y). Categorical variables, including gender, location, game genre, and game difficulty, were converted into numerical representations using one hot encoding to allow machine learning algorithms to process categorical information effectively. In addition, numerical variables such as age, playtime hours, sessions per week, average session duration, player level, and achievements unlocked were normalized using feature scaling techniques to ensure that variables with larger value ranges do not dominate the learning process and to improve the stability and performance of the machine learning models.

Data Splitting

To evaluate the predictive performance of the machine learning models, the dataset was divided into training and testing datasets.

The dataset was split using the following ratio

Training data = 80%

Testing data = 20%

This approach allows the models to learn patterns from the training data and then evaluate their predictive capability using unseen testing data. Stratified sampling was applied to maintain the distribution of engagement classes across both datasets.

Machine Learning Models

Three machine learning algorithms were implemented in this study, namely Gradient Boosting, Neural Network using Multilayer Perceptron (MLP), and Extreme Gradient Boosting (XGBoost). These algorithms were selected because they are widely used for classification tasks and have demonstrated strong performance in handling structured datasets. Gradient Boosting and XGBoost are ensemble learning methods that combine multiple decision trees to improve predictive accuracy. These models work by sequentially building trees where each new tree attempts to reduce the prediction errors produced by the previous trees through gradient optimization. This iterative learning process allows the models to capture complex patterns in the data and improve classification performance. In contrast, the Neural Network model uses a multilayer perceptron architecture that consists of interconnected layers of neurons capable of learning nonlinear relationships between input features and the target variable. By processing multiple features simultaneously, the neural network can identify hidden patterns within player behavioral data and contribute to the prediction of engagement levels in online gaming environments.

Model Evaluation

The predictive performance of the models was evaluated using several classification metrics including accuracy, precision, recall, and F1 score.

Accuracy measures the proportion of correct predictions among the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

Precision measures how many predicted positive observations are actually correct.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall measures the ability of the model to correctly identify all positive observations.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The F1 score combines precision and recall into a single metric.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

Cross Validation

To ensure the robustness of the models, 5-fold cross validation was conducted. In this approach, the dataset is divided into five subsets. During each iteration, four subsets are used for training and one subset is used for validation.

This process is repeated five times so that each subset is used as validation data once. Cross validation helps evaluate the stability and generalization capability of the models.

Confusion Matrix Analysis

A confusion matrix was used to analyze the classification performance of the best performing model by comparing the predicted engagement categories with the actual engagement labels in the dataset. The confusion matrix provides a detailed summary of how many observations were correctly classified and how many were misclassified for each engagement category. This analysis allows a clearer understanding of the model's predictive behavior by showing the distribution of predictions across the Low, Medium, and High engagement levels. Through this evaluation, it is possible to assess how effectively the model distinguishes between different engagement classes and identify whether certain categories are more difficult to predict than others. Overall, the confusion matrix helps provide deeper insight into the classification accuracy and reliability of the predictive model.

Feature Importance Analysis

Feature importance analysis was conducted using the XGBoost model to determine which variables contribute the most to predicting player engagement levels. Feature importance represents the relative contribution of each variable to the model's predictive performance by measuring how much each feature improves the model during the training process. By examining these importance values, the study identifies the behavioral variables that have the strongest influence on engagement prediction. This analysis allows the identification of key gameplay factors such as session frequency, session duration, player progression, and achievements that are associated with higher or lower engagement levels. Overall, the feature importance results provide valuable insights into how player activity patterns and in game behavior influence engagement within online gaming environments.

Algorithm 1: Player Engagement Prediction

Input: dataset D, feature variables X, target variable Y

Output: predicted engagement level, best machine learning model

Process:

Start

Load dataset D

Remove identifier variable

$$D = D - PlayerID$$

Encode categorical variables using one hot encoding

$$X_{cat} = OneHotEncoding(X_{cat})$$

Normalize numerical variables

$$X_{norm} = \frac{X - \mu}{\sigma}$$

Split dataset into training and testing data

$$D_{train}, D_{test} = Split(D, 80\%, 20\%)$$

Initialize machine learning models

$$M = \{GradientBoosting, NeuralNetwork, XGBoost\}$$

For each model $m \in M$

Train model using training data

$$m = Train(D_{train})$$

Predict engagement level

$$\hat{Y} = Predict(m, D_{test})$$

Compute evaluation metrics

Accuracy

Precision

Recall

F1Score

End for

Select best model based on highest F1Score

$$m^* = argmax(F1Score)$$

Compute feature importance using best model

$$FI = FeatureImportance(m^*)$$

Return predicted engagement level and important features

End

Results

Dataset Overview

The dataset used in this study consists of 40,034 records of online game players and includes 13 attributes that represent demographic characteristics, gameplay behavior, and engagement classification. The variables comprise age, gender, location, game genre, playtime hours, in game purchases, game difficulty level, number of sessions per week, average session duration in minutes, player level, and achievements unlocked. These variables capture different aspects of player interaction within online gaming environments, ranging from demographic background to behavioral patterns during gameplay. The target variable in this dataset is EngagementLevel, which categorizes player engagement into three classes: Low, Medium, and High engagement. This classification allows the study to analyze varying levels of player involvement and predict engagement behavior using machine learning techniques.

The dataset contains no missing values, which indicates that all observations are complete and suitable for predictive modeling. Descriptive statistical analysis shows that players have an average age of 31.99 years with a standard deviation of 10.04 years, suggesting that the dataset represents a broad range of player ages. The average playtime hours is approximately 12.02 hours, indicating moderate gameplay activity among participants. In addition, the average session duration is 94.79 minutes, which suggests that players typically spend a substantial amount of time in each gaming session. These descriptive statistics provide an overview of the player population and highlight the diversity of gaming behavior captured in the dataset, which is essential for developing reliable machine learning models to predict engagement levels.

Distribution of Engagement Levels

The distribution of engagement levels among players is illustrated in figure 2. The results show that the medium engagement category represents the largest proportion of players, totaling 19,374 individuals, which accounts for nearly half of the dataset. In comparison, the High engagement category includes 10,336 players, while the Low engagement category consists of 10,324 players, both representing smaller but almost equal portions of the dataset. This distribution indicates that most players interact with online gaming platforms at a moderate level of activity, suggesting consistent but not excessive gameplay behavior. Meanwhile, players classified as highly engaged or minimally engaged appear in smaller groups, which reflects the presence of both highly active users who frequently participate in gaming sessions and less active users who interact with the platform more occasionally. Overall, the engagement distribution demonstrates that the majority of players fall within the moderate engagement category, providing a balanced dataset that is suitable for training machine learning models to distinguish different levels of player involvement.

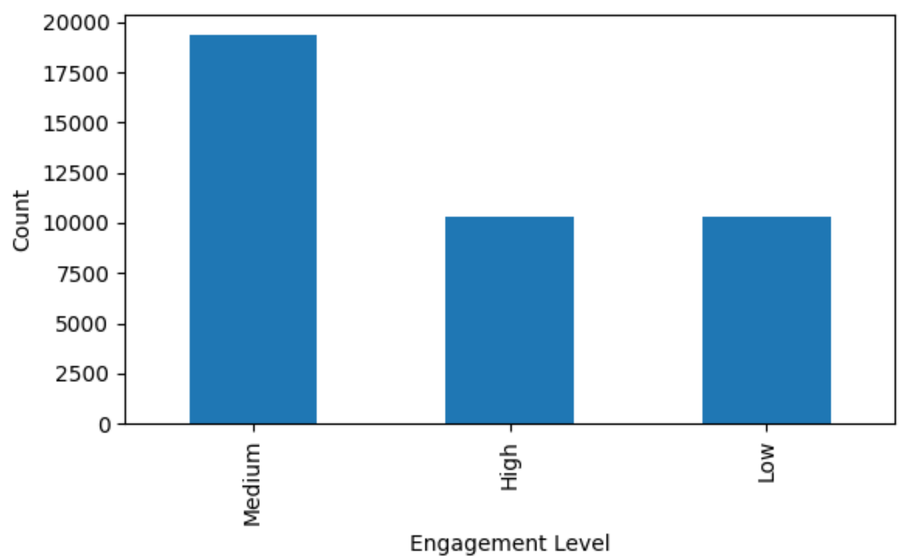
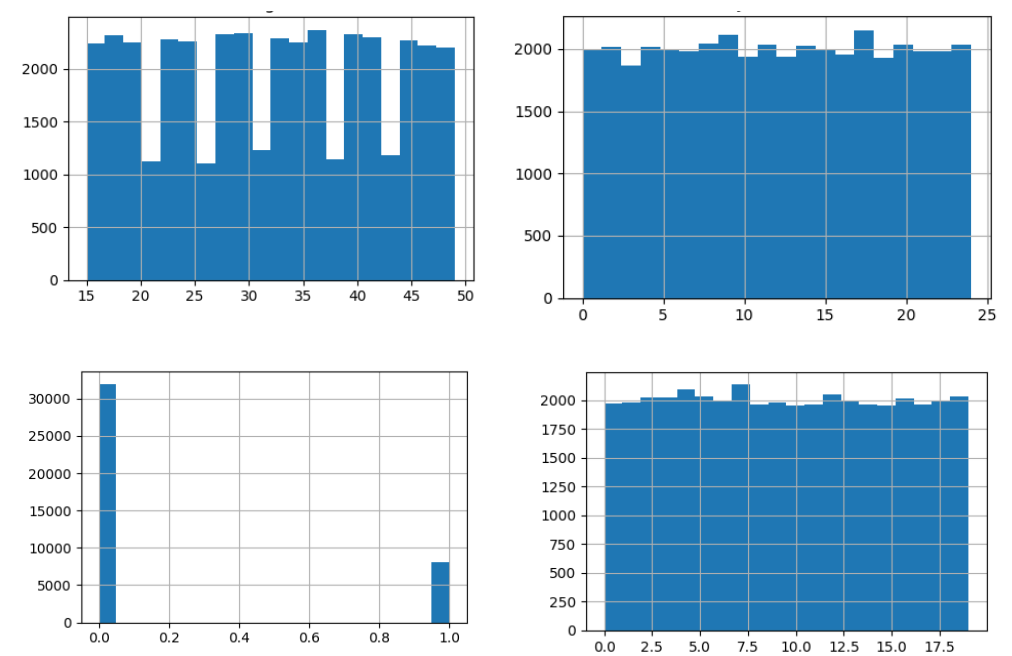


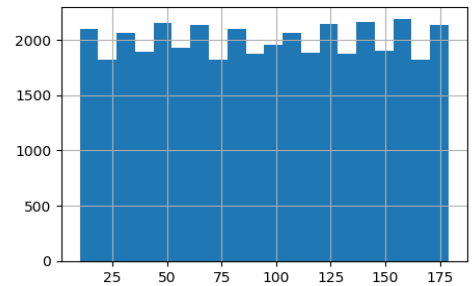
Figure 2 **Distribution of player engagement levels.**

Distribution of Numerical Features

Exploratory analysis was conducted to examine the distribution of several key numerical variables that represent player behavior within the gaming platform. These variables include age, playtime hours, sessions per week, average session duration, player level, and achievements unlocked. The distribution of player age, shown in figure 3, ranges from 15 to 49 years, indicating that online gaming platforms attract individuals from a wide range of age groups. Although the ages vary considerably, the distribution suggests that players are primarily concentrated within young adult and adult age groups. The distribution of PlayTimeHours in figure 3 demonstrates substantial variation in the total number of hours players spend in gaming activities. This variation suggests that players differ in their gaming intensity, with some individuals spending relatively limited time playing while others invest significantly more time in gaming sessions.

The distribution of SessionsPerWeek, presented in figure 3, shows that players participate in gaming sessions between 0 and 19 times per week, indicating substantial differences in gameplay frequency. Some players engage with the game only occasionally, while others interact with the platform almost daily. The AvgSessionDurationMinutes distribution in figure 3 indicates that the duration of each gaming session ranges between 10 and 179 minutes, reflecting a wide range of gameplay durations among players. Meanwhile, the PlayerLevel distribution in figure 3 shows progression levels between 1 and 99, which represent different stages of player advancement and experience within the game environment. Finally, the distribution of AchievementsUnlocked in figure 3 ranges from 0 to 49 achievements, indicating varying levels of accomplishment and in game success among players. Overall, these distributions highlight the diversity of behavioral patterns among players and provide important insights into how individuals interact with online gaming platforms.





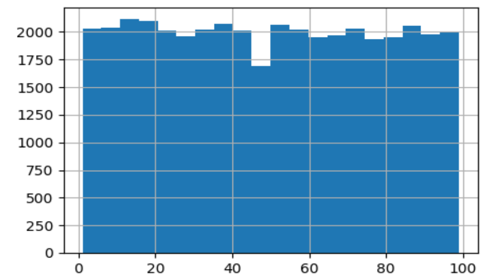


Figure 3 Distribution of numerical variables: age distribution of players, playtime hours, sessions per week, average session duration, player level, and achievements unlocked.

Machine Learning Model Performance

Three machine learning models were evaluated in this study to predict player engagement levels based on behavioral and demographic features within the dataset. The selected models include Gradient Boosting, Neural Network using a Multilayer Perceptron (MLP) architecture, and XGBoost, which are widely used algorithms for classification tasks due to their ability to capture complex patterns in large datasets. Each model was trained using the prepared dataset and then evaluated using several commonly used performance metrics, including accuracy, precision, recall, and F1 score. These metrics provide a comprehensive assessment of the model's predictive capability by measuring overall correctness, the ability to correctly identify engagement categories, and the balance between precision and recall across the classes. The comparative performance results of the evaluated models are presented in table 2, which summarizes the effectiveness of each algorithm in predicting player engagement levels within the online gaming environment.

Table 2 Performance comparison of machine learning models

Model	Accuracy	Precision	Recall	F1-score
XGBoost	0.914	0.914	0.914	0.9137
Gradient Boosting	0.9037	0.9040	0.9037	0.9033
Neural Network (MLP)	0.8905	0.8904	0.8905	0.8902

The results indicate that XGBoost achieved the highest predictive performance among the evaluated models, obtaining an accuracy of 91.4 percent and an F1 score of 0.9137. These results demonstrate that XGBoost is particularly effective in capturing complex patterns within player behavioral data, which contributes to more accurate classification of engagement levels. The superior performance of XGBoost can be attributed to its ensemble learning approach, which combines multiple decision trees and applies gradient boosting to iteratively improve prediction accuracy. By integrating several weak learners into a stronger predictive model, XGBoost is able to reduce prediction errors and enhance overall model performance. The comparison of model performance based on the F1 score, which provides a balanced evaluation between precision and recall, is illustrated in figure 4. The visualization further confirms that XGBoost consistently outperforms both Gradient Boosting and the Neural Network model, highlighting its effectiveness for predicting player engagement behavior in online gaming environments.

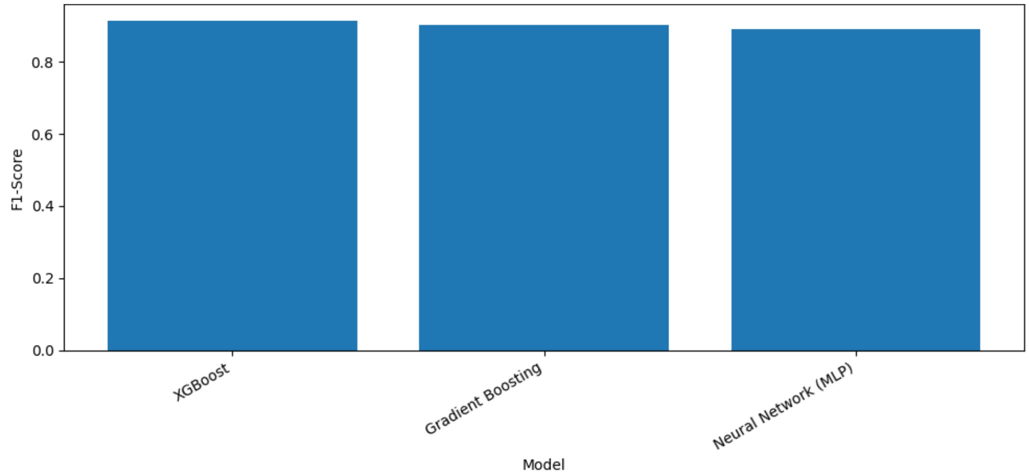


Figure 4 **Model comparison based on F1-score.**

Cross-Validation Results

To evaluate the robustness and reliability of the machine learning models, 5-fold cross-validation was performed. In this procedure, the dataset was divided into five subsets of approximately equal size. During each iteration, four subsets were used to train the model while the remaining subset was used for validation. This process was repeated five times so that each subset served as the validation data once. Cross-validation helps ensure that the evaluation results are not dependent on a single train-test split and provides a more comprehensive assessment of model performance. The results show that XGBoost achieved the highest average performance with a mean F1 score of 0.9163 and a standard deviation of 0.0013, indicating that the model consistently performs well across different data partitions.

The low standard deviation observed in the cross-validation results suggests that the model's predictive performance remains stable when trained and evaluated on different subsets of the dataset. This stability indicates that the model is able to capture generalizable patterns in player behavior rather than memorizing specific training samples. As a result, the XGBoost model demonstrates strong generalization capability and maintains consistent accuracy when applied to unseen data. These findings further support the suitability of XGBoost for predicting player engagement levels in online gaming environments, as it provides reliable performance and reduces the risk of overfitting during the training process.

Confusion Matrix Analysis

The confusion matrix of the best performing model, XGBoost, is presented in figure 5 and provides a detailed overview of the model's classification performance across the three engagement categories. The matrix shows how many instances from each actual engagement class were correctly predicted and how many were misclassified into other categories. The results indicate that the model correctly classified the majority of observations in each engagement category, demonstrating strong predictive capability across all classes. Based on the classification report, the model achieved an F1 score of 0.90 for High engagement, 0.90 for Low engagement, and 0.93 for medium engagement, indicating balanced predictive performance among the three categories. The slightly higher F1 score for the medium engagement class suggests that the model is particularly effective in identifying moderately engaged players, which also represent the largest portion of the dataset. Overall, the confusion matrix results confirm that the XGBoost model performs consistently across different engagement levels and is capable of accurately distinguishing between low, medium, and high player engagement in online gaming environments.

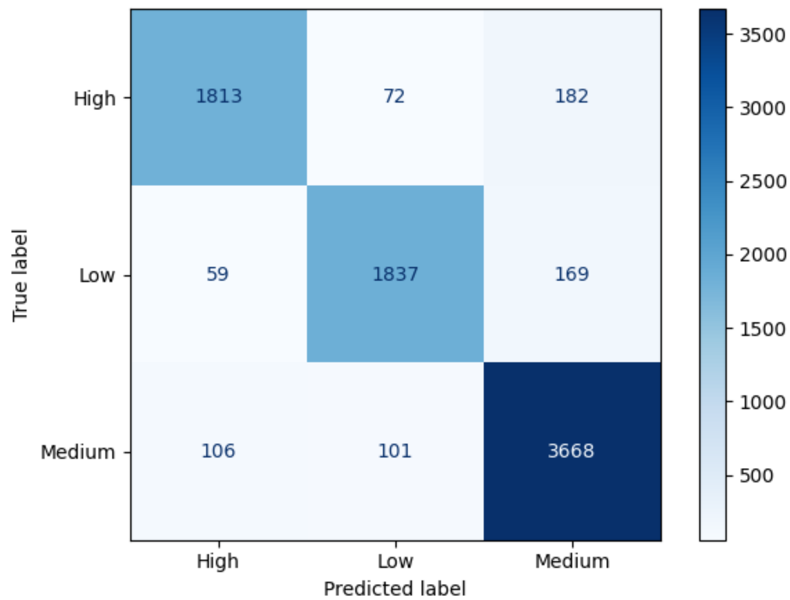


Figure 5 **Confusion matrix of the XGBoost model.**

Feature Importance Analysis

Feature importance analysis was conducted using the XGBoost model to determine which variables contribute the most to predicting player engagement levels. The results, illustrated in figure 6, indicate that behavioral variables related to gameplay activity play a dominant role in determining engagement outcomes. Among all predictors, SessionsPerWeek emerged as the most influential feature with an importance score of 0.3885, indicating that the frequency with which players access the game each week is the strongest indicator of engagement level. The second most important variable is AvgSessionDurationMinutes with an importance value of 0.2268, suggesting that the amount of time players spend during each gameplay session significantly contributes to engagement prediction. Other relevant features include AchievementsUnlocked (0.0357) and PlayerLevel (0.0334), which represent player progression and accomplishments within the game environment, reflecting long term interaction with the platform. Additionally, InGamePurchases (0.0198) also contributes to the prediction, although with a smaller influence compared to gameplay activity variables. Overall, these findings indicate that player behavioral patterns, particularly the frequency and duration of gameplay sessions, serve as the primary determinants of engagement levels in online gaming platforms, while progression related variables and in game transactions provide additional supporting signals for predicting player involvement.

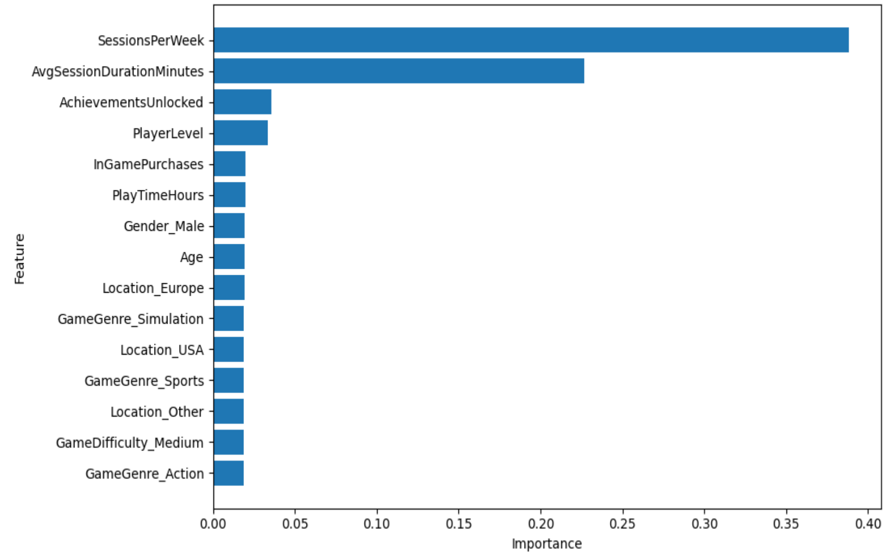


Figure 6 **Top 15 feature importance values from the XGBoost model.**

Discussion

The findings of this study demonstrate that machine learning techniques can effectively predict player engagement levels in online gaming environments using behavioral and demographic data. Among the evaluated models, XGBoost achieved the highest predictive performance, outperforming Gradient Boosting and the Neural Network model in terms of accuracy, precision, recall, and F1 score. This result indicates that ensemble-based tree models are particularly effective for capturing complex relationships within player behavioral data. The strong performance of XGBoost can be attributed to its ability to iteratively optimize prediction errors through gradient boosting while combining multiple decision trees to form a more robust model. These results suggest that machine learning approaches are capable of identifying patterns in player behavior that are associated with different levels of engagement, which can provide valuable insights for understanding user interaction within digital gaming platforms.

The feature importance analysis further reveals that behavioral variables play a more significant role than demographic characteristics in predicting engagement levels. In particular, the frequency of gameplay sessions and the duration of each session emerged as the most influential predictors of engagement. This indicates that engagement is primarily driven by how frequently players interact with the game and how long they remain active during each session. Variables related to player progression, such as achievements unlocked and player level, also contribute to engagement prediction by reflecting long term commitment and accomplishment within the game environment. In contrast, demographic variables such as age, gender, and location show relatively smaller influence on the prediction results. These findings suggest that player engagement in online gaming platforms is largely determined by behavioral interaction patterns rather than demographic background, highlighting the importance of analyzing gameplay activity to better understand user engagement in digital environments.

Conclusion

This study demonstrates that machine learning methods can effectively predict player engagement levels in online gaming environments by utilizing behavioral and demographic data derived from gameplay activity. Using a dataset consisting of 40,034 player records, three machine learning models were evaluated, including Gradient Boosting, Neural Network, and XGBoost. The experimental results indicate that XGBoost achieved the best predictive performance with an accuracy of 91.4 percent and an F1 score of 0.9137, confirming its effectiveness in modeling engagement behavior. Cross validation further confirmed the robustness and stability of the model across different data partitions. Feature importance analysis revealed that behavioral variables such as sessions per week and average session duration are the most influential predictors of engagement, followed by progression related variables such as achievements unlocked and player level. These findings suggest that player engagement in online gaming platforms is primarily driven by gameplay interaction patterns rather than demographic characteristics. Overall, this study highlights the potential of machine learning approaches for understanding and predicting human engagement behavior in digital gaming environments, which can provide valuable insights for game developers, platform designers, and researchers studying user interaction in digital societies.

Declarations

Author Contributions

Conceptualization: G.A. and R.K.M.; Methodology: R.K.M.; Software: G.A.; Validation: G.A. and R.K.M.; Formal Analysis: G.A. and R.K.M.; Investigation: G.A.; Resources: R.K.M.; Data Curation: R.K.M.; Writing Original Draft Preparation: G.A. and R.K.M.; Writing Review and Editing: R.K.M. and G.A.; Visualization: G.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] D. Zendle and P. Cairns, "Correction: Video game loot boxes are linked to problem gambling: Results of a large-scale survey," *PLOS ONE*, vol. 14, no. 3, p. e0214167, pp. 1-12, Mar. 2019, doi: 10.1371/journal.pone.0214167.

[2] J. Hamari and L. Keronen, "Why do people play games? A meta-analysis," *International Journal of Information Management*, vol. 37, no. 3, pp. 125–141, Jun. 2017, doi: 10.1016/j.ijinfomgt.2017.01.006.

[3] M. Seif El-Nasr, A. Drachen, and A. Canossa, Eds., *Game Analytics: Maximizing the Value of Player Data*. London, U.K.: Springer-Verlag, 2013, doi: 10.1007/978-1-4471-4769-5.

[4] A. Drachen, C. Thurau, J. Sifa, and C. Bauckhage, "A comparison of methods for player clustering via behavioral telemetry," *Foundations of Digital Games*, pp. 1–8, 2012, doi: 10.1145/2282338.2282370.

[5] T. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, vol. 2016, no. Aug., pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[6] N. Yee, "Motivations for play in online games," *CyberPsychology & Behavior*, vol. 9, no. 6, pp. 772–775, Jan. 2007, doi: 10.1089/cpb.2006.9.772.

[7] A. Drachen, M. Seif El-Nasr, and A. Canossa, "Game analytics – The basics," in *Game Analytics: Maximizing the Value of Player Data*, M. Seif El-Nasr, A. Drachen, and A. Canossa, Eds. London, U.K.: Springer, vol. 2013, no. January, pp. 13-40, 2013, doi: 10.1007/978-1-4471-4769-5_2.

[8] R. Sifa, C. Bauckhage, and A. Drachen, "The playtime principle: Large-scale cross-games interest modeling," in *Proc. IEEE Conf. Computational Intelligence and Games (CIG)*, vol. 2014, no. October, pp. 1-8, Aug. 2014, doi: 10.1109/CIG.2014.6932906.

[9] J. Runge, P. Gao, F. Garcin, and B. Faltings, "Churn prediction for high-value players in casual social games," in *Proc. IEEE Conf. Computational Intelligence and Games (CIG)*, Dortmund, Germany, 2014, vol. 2014, no. October, pp. 1–8, doi: 10.1109/CIG.2014.6932875.

[10] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Nov. 2000, doi: 10.1214/aos/1013203451.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[12] M. Claypool and K. T. Claypool, "Latency and player actions in online games," *Communications of the ACM*, vol. 49, no. 11, pp. 40–45, Nov. 2006, doi: 10.1145/1167860.

[13] R. Hunicke, M. LeBlanc, and R. Zubek, "MDA: A formal approach to game design and game research," in *Proc. AAAI Workshop on Challenges in Game AI*, 2004.

[14] J. Hamari, J. Koivisto, and H. Sarsa, "Does gamification work? — A literature review of empirical studies on gamification," in *Proc. 47th Hawaii Int. Conf. System*

Sciences (HICSS), Waikoloa, HI, USA, 2014, vol. 2014, no. March, pp. 3025–3034, doi: 10.1109/HICSS.2014.377.

[15] R. Sifa, A. Drachen, and C. Bauckhage, “Large-scale cross-game player behavior analysis on Steam,” in *Proc. AAAI Conf. Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, 2015.

[16] A. Drachen, M. Yancey, J. Maguire, and D. Chu, “Skill-based differences in spatio-temporal team behaviour in Defence of the Ancients 2 (DotA 2),” in *Proc. IEEE Games, Entertainment and Media Conf. (GEM)*, vol. 2015, no. February, pp. 1-8, Oct. 2014, doi: 10.1109/GEM.2014.7048109.

[17] T. Chen et al., *xgboost: Extreme Gradient Boosting*, version 3.2.1.1, Mar. 18, 2026, doi: 10.32614/CRAN.package.xgboost.

[18] J. H. Friedman, “Stochastic gradient boosting,” *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, Feb. 2002, doi: 10.1016/S0167-9473(01)00065-2.

[19] H. A. Salman, A. Kalakech, and A. Steiti, “Random forest algorithm overview,” *Babylonian Journal of Machine Learning*, vol. 2024, no. June, pp. 69–79, Jun. 2024, doi: 10.58496/BJML/2024/007.

[20] A. Drachen, A. Canossa, and G. N. Yannakakis, “Player modeling using self-organization in Tomb Raider: Underworld,” in *Proc. IEEE Symp. Computational Intelligence and Games (CIG)*, Milan, Italy, vol. 2009, no. October, pp. 1–8, 2009, doi: 10.1109/CIG.2009.5286500.