MIT | Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

Online Gaming Behavior Analysis Using Machine Learning

TY B.Tech
Computational Intelligence
Project Report

SUBMITTED BY

Vaishnav Lajulkar - 202301070140

Mukund Chavan -   202301070141

Aditya Ganeshe –   202301070142

Prathamesh Shinde- 202301070143

GUIDED BY
Dr. Abhilasha Joshi
Prof. Department of E&TC

MIT ACADEMY OF ENGINEERING, ALANDI (D), PUNE-412105

MAHARASHTRA (INDIA)

Nov, 20265

ACKNOWLEDGEMENT

It is with immense pleasure that we express our heartfelt gratitude to all those who have supported and guided us throughout the successful completion of this project.

First and foremost, we extend our deepest thanks to our guide, whose constant encouragement, expert guidance, and constructive feedback have played an instrumental role in shaping the direction and quality of this project. Their insights and mentorship motivated us to continuously refine our ideas and explore innovative approaches.

We also express our sincere appreciation to the Dean, Department of E&TC, for providing necessary resources, technical support, and a conducive environment for conducting this study. Their commitment to academic excellence has greatly benefited our work.

Furthermore, we are thankful to our peers, friends, and family who supported us with their encouragement and confidence throughout this endeavor. Their belief in our abilities helped us remain focused and dedicated.

Finally, we extend gratitude to everyone—directly or indirectly involved—who contributed to the completion of this project. This work stands as a reflection of collective effort, determination, and the pursuit of knowledge.

Thank you.

ABSTRACT

This project explores the use of machine learning techniques to analyze and predict online gaming behavior, focusing primarily on player engagement, gameplay patterns, and purchase tendencies. The dataset—containing 40,034 gamer records—underwent extensive preprocessing, including handling missing values, encoding categorical variables, and standardizing numerical features to ensure proper model compatibility.

A variety of machine learning algorithms were implemented, including Logistic Regression, Random Forest, Support Vector Machines (SVM), XGBoost, and LightGBM. These models were evaluated across performance metrics such as accuracy, precision, recall, and F1-score to determine the most effective approach for predicting player engagement levels.

Hyperparameter tuning was employed to optimize performance, resulting in improved accuracy and model generalization. Among all models tested, LightGBM delivered the highest accuracy of 88%, outperforming the rest.

This study highlights the importance of structured preprocessing, robust evaluation, and comparative analysis when using machine learning to understand digital behavior trends within the online gaming ecosystem.

1. INTRODUCTION

Online gaming has grown into one of the largest digital entertainment industries globally, with millions of users engaging across various platforms. Understanding user behavior—such as gameplay patterns, engagement levels, and spending habits—has become essential for game developers, publishers, and analytics companies.

Traditionally, player behavior analysis relied on basic statistics and manual insights. However, with the increasing volume and complexity of gaming data, these traditional methods are insufficient. Machine learning provides powerful tools to uncover deeper behavioral patterns, enabling personalized recommendations, improved user retention strategies, and better in-game experience design.

This project aims to develop a machine learning-based player engagement prediction system using a comprehensive online gaming dataset. The methodology includes data preprocessing, feature engineering, model development, and evaluation. Multiple machine learning models were tested and compared using standard performance metrics.

The resulting insights offer a deeper understanding of key factors driving player engagement, helping gaming companies make informed decisions regarding monetization, content design, and gameplay optimization.

1.1 OBJCETIVE
The primary objective of this project is to analyze online gamer behavior and develop a machine-learning model capable of accurately predicting player engagement levels. The objectives include:

1. Data Analysis and Preprocessing
   - Understanding key factors that influence online gaming behavior.
   - Handling missing values, inconsistent entries, and scaling numerical features.
   - Encoding categorical features such as Gender, Game Genre, and Difficulty Level.

2. Model Development and Evaluation
   - Implementing and comparing various ML algorithms including:
   - Logistic Regression
   - Random Forest

- Support Vector Machines (SVM)
- XGBoost
- LightGBM

Evaluating models based on accuracy, precision, recall, and F1-score.

3. Optimization
- Performing hyperparameter tuning using Randomized Search with cross-validation.
- Improving model performance and preventing overfitting.

4. Insights and Recommendations
- Identifying factors that significantly impact player engagement, playtime, and in-game purchases.
- Offering insights that can assist gaming companies in enhancing user experience and retention.
- The ultimate goal is to build a reliable and efficient prediction system that can help gaming platforms understand and respond to player behavior more effectively.

1.2 MOTIVATIONS

The project is motivated by the increasing need within the gaming industry to understand and predict user behavior. As online games continue to evolve, user engagement and play patterns have become crucial indicators of game success.

Traditional methods of analyzing gaming behavior often rely on basic descriptive analytics, which may overlook complex relationships. Machine learning offers advanced capabilities to analyze vast amounts of player data, uncover hidden patterns, and make accurate predictions.

Understanding player engagement is essential for:

- improving in-game mechanics

- reducing user churn

- personalizing content

- optimizing marketing strategies

- enhancing monetization models

This project aims to leverage machine learning techniques to develop predictive models that can significantly support the gaming industry by providing deeper behavioral insights and data-driven decision-making.

2. METHODOLOGY

The methodology follows a structured approach, integrating machine learning with CI/MLOps concepts for automation and performance.

1. Data Collection and Preprocessing
The dataset used includes demographics (age, gender, region), gameplay statistics (difficulty level, sessions per week, playtime hours), and behavioral features (engagement level, purchases).
Preprocessing involved:
- Removing missing or inconsistent entries
- Label encoding categorical attributes
- Standardizing continuous variables
- Performing Exploratory Data Analysis (EDA)

2. Feature Selection and Engineering
- Identifying relevant features influencing gamer engagement.
- Using model-based importance methods (Random Forest, LightGBM) to rank features.
- Creating new derived features such as:
- Playtime Ratio
- Weekly Activity Score

3. Model Development
Multiple algorithms were trained and evaluated:
- Logistic Regression
- k-Nearest Neighbors (kNN)
- Support Vector Machines (SVM)
- Random Forest
- XGBoost
- LightGBM (best performer)

- Train-test split: 80:20
- Cross-validation used to prevent overfitting.

4. Performance Evaluation
- Models were evaluated on:
- Accuracy
- Precision
- Recall
- F1-Score

Model Performance Summary

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.814	0.815	0.814	0.812
Random Forest	0.877	0.878	0.877	0.877
XGBoost	0.879	0.879	0.879	0.879
LightGBM	0.881	0.882	0.881	0.881

5. Continuous Integration (CI)
- CI pipeline automates:
- Code integration
- Model training
- Testing
- Deployment
-

6. MLOps Integration Includes:
- Model versioning
- Monitoring drift
- Automated retraining
- Performance logging

7. Deployment
The final model was deployed using Flask and cloud-based support, enabling real-time prediction and continuous monitoring.

3.   RESULTS

1. Model Performance
LightGBM delivered best performance with 88% accuracy, strong precision, and balanced recall.
Key observations:
- Random Forest and XGBoost also performed well but required longer training time.
- Logistic Regression lagged due to linear decision boundaries.
- Engagement level strongly predicted by playtime hours, difficulty level, and purchase behavior.

2. Key Insights
- Medium engagement players unlock the highest achievements.
- Players with easy difficulty settings tend to make more purchases.
- Average session duration = $\approx$95 minutes/day.
- Region-wise participation: USA > Europe > Asia.

3. Model Interpretability
SHAP values indicated top influential features:
- Playtime Hours
- Sessions Per Week
- Game Difficulty
- In-Game Purchases
- Age & Region

4. CI/MLOps Integration
Automated workflows ensured:
- Faster testing
- Continuous model updates
- Error-free deployment
- Model monitoring for performance consistency

4. CHALLENGES

1. Overfitting

Managed through:

- Cross-validation
- Regularization
- Early stopping

2. Data Quality Issues

Handled using:

- Imputation
- Feature scaling
- Removing outliers

3. Interpretability

Addressed via:

- SHAP
- LIME
- Visualization dashboards

5. FUTURE IMPROVEMENTS

- Integrating deep learning models (LSTM/ANN)
- Real-time behavioral tracking via IoT gaming devices
- Personalized recommendation engine
- Advanced MLOps pipeline with drift alerts
- Gamification-based analytics dashboards
- Larger and more diverse datasets

CONCLUSION

The project successfully applied machine learning techniques to analyze online gaming behavior and predict engagement levels. Using structured preprocessing, model comparison, and tuning strategies, LightGBM emerged as the best-performing model with high accuracy and generalization capability.

This work demonstrates the potential of ML-driven behavioral analytics in enhancing gaming experiences, improving user retention, and supporting data-driven decision-making for gaming platforms.