Online Gaming Behavior Analysis - Project Report

Title: “Online Gaming Behavior Analysis using Machine Learning”

Name: Sharan Thakur - GH1031360

University Name: GISMA University of Applied Sciences, Potsdam

Module Name: M606 - Machine Learning

Professor: Prof. Dr. Mahmoudreza Babaei

Introduction

The gaming industry is a huge market with exponential growth every year. With fierce competition amongst game developers/publishers, understanding player engagement has become crucial for success.

This project applies machine learning techniques to analyze player engagement patterns in online games, providing some business insights that can help game publishers and developers alike.

By examining a dataset of 40,000+ players, I want to identify the key factors that influence engagement levels and build Machine Learning model that can classify players based on their engagement potential. These insights will help gaming publishers prioritize features that maximize player retention while keeping them engaged.

Problem Statement

Business Problem

Game development studios and publishers invest millions in creating engaging experiences, yet player retention remains a significant challenge. Understanding what drives engagement can drastically improve:

- Player retention rates
- Monetization opportunities
- Effective marketing

Importance

High player engagement directly correlates with:

- Increased player loyalty
- Higher in-game purchasing
- Better word of mouth marketing
- Reduced player acquisition costing

Data Collection Methodology

Normally, we will collect data from: the game itself (tracking how people play), surveys (asking them what they think), and the game servers (to see how long sessions last).

For this project, though, I used an anonymized dataset from Kaggle. It contains player info, how they behaved in the game, and how engaged they seemed to be.

Machine Learning Formulation

Formulating this as a **classification problem**:

- **Target Variable**: Player engagement level (Low, Medium, High)
- **Features**: AchievementUnlocked, PlayerLevel, Age, InGamePurchasing, PlayTimeBehavior, SessionsPerWeek, AverageSessionDuration

- **Goal**: Build a model that predicts engagement levels based on player characteristics and behaviors

Project Resources

Please check out my code on GitHub and the deployed dashboard with all interactive charts on GitHub Pages

- **GitHub Repository**: https://github.com/c2p-cmd/online-gaming-behavior-analysis
- **Interactive Dashboard**: https://c2p-cmd.github.io/online-gaming-behavior-analysis

Technologies Used

This project was developed using:

- **Python 3.13** as the primary programming language
  - Python has many great libraries for Machine Learning tasks
- **Jupyter Notebook** for interactive development and documentation
  - For running code on cell basis and seeing the outputs immediately to avoid re-compiling the entire code
- **Pandas** for data manipulation
  - Reading the dataset and performing statical analysis
- **Plotly.js** for data visualization
  - Plotly.js has rich and visual charts which are interactive in nature which can better help understand out data
  - Plotly.js also has python (Plotly) package which is supported in Jupyter
- **Scikit Learn** for Machine Learning model development and evaluation
  - For easy implementing Machine Learning algorithms and getting important metrics about the model.

Conclusion

Strengths

1. **High Predictive Power**: Our fine-tuned Random Forest model achieved 91% accuracy after hyperparameter tuning in predicting engagement levels
2. **Clear Insights**: Clear identification of behavioral metrics (especially session frequency) as primary engagement predictors
3. **Balanced Performance**: The model performs well across all three Player Engagement classes

Limitations

1. **Static Dataset**: The analysis uses historical data that has been anonymized and may not capture newer player preferences
2. **Limited Demographic Diversity**: The dataset may not represent all kinds of players
3. **Limited Context Variables**: Due to the anonymized nature of the data, we missed out on information regarding game design elements as well as UI/UX factors.

Business Recommendations

Our findings indicate that Player Engagement is primarily driven by:

1. **Session Frequency** (43% importance) - more important than session duration
2. **Session Duration** (31% importance) - players must be playing for substantial hours
3. **Progression Systems** (Player Level 6.6%, Achievements 5.9%) - moderately important
4. **Monetization** (<1% importance) - surprisingly minimal impact on engagement so, more money in the game doesn’t always mean more engagement

These findings suggest that game publishers should prioritize on the game’s progression system as well as the a good reward system as opposed to selling goods inside the games.

Data-Driven Recommendations

1. **Implement Engagement Systems**: Design mechanics that encourage players to return frequently (daily quests, login rewards)
2. **Optimize Session Design**: Structure game content for satisfying 90-95 minute sessions

3. **Enhance Progression Systems**: Make achievements and level progression more meaningful
4. **Rethink Monetization Approach**: Focus on player satisfaction rather than in game purchase

Model Explainability

Our **Random Forest** model offers good explainability through feature importance analysis:

- It clearly points to SessionsPerWeek and AvgSessionDuration as the big factors.
- The importance values align with gaming industry research on engagement factors
- We can trace the decision-making process to see why the model makes certain predictions.
- The model's predictions make sense in context of gaming industry