Table of Contents

Table of Figures

Confusion Matrix Figures (Section 5.1.1)

Table of Tables

ABSTRACT

This project uses advanced machine learning techniques on user behavior and demographic data in a gaming context to predict player engagement levels, which can be categorized as Low, Medium, and High. To improve user experience, retention, and monetization strategies, game developers and analysts must have a thorough understanding of player engagement.The study starts with thorough data preprocessing, which includes feature classification, cleaning, and the creation of new, valuable attributes like average session lengths and weekly session counts. To find trends and distributional patterns in the dataset, especially with regard to the distribution of user engagement, exploratory data analysis, or EDA, was performed.Some classification algorithms, such as K-Nearest Neighbors (KNN), Random Forest, Logistic Regression, Gradient Boosting, and XGBoost, were used to model engagement. Key metrics like accuracy, precision, recall, and F1-score were used to train and assess each model. The XGBoost classifier outperformed the others, especially after being adjusted via Bayesian optimization with the Optuna framework. With a test accuracy of more than 91%, the optimized model outperformed baseline models and demonstrated an outstanding ability for generalization.

Additionally, i have Python-based UI library called Gradio was used to create a real-time prediction system. Users can enter their gameplay habits into this interface to get real-time predictions about their level of engagement. The system improves the application's scalability and adaptability by logging user input for possible future analysis in addition to making predictions.All things considered, this project not only produces a highly effective prediction model but also demonstrates the usefulness of real-time machine learning applications in gaming analytics , with the possibility of incorporation into for-profit gaming platforms.

Chapter 1: Introduction and Background

Player engagement has become a crucial metric for comprehending user behavior, maintaining player interest, and promoting business success in the quickly changing gaming industry of today. User retention, in-game purchases, session duration, and general satisfaction are all directly impacted by engagement, which measures how frequently and actively a user interacts with a game. While low engagement may indicate possible churn, discontent, or a misalignment between player preferences and game design, high engagement frequently translates into more monetization opportunities and greater user loyalty.

The use of data-driven approaches has grown in significance due to the abundance of gameplay data produced by contemporary gaming platforms. Specifically, machine learning (ML) provides strong tools for analyzing intricate behavioral patterns and forecasting user behavior in the future. Developers can segment their player base, identify at-risk users, and provide incentives or personalized content to improve the gaming experience by utilizing machine learning models.

Based on behavioral and demographic characteristics, this project investigates the use of supervised classification techniques to predict player engagement levels, which are classified as Low, Medium, or High. The project focuses on real-time deployment through an intuitive interface, in addition to developing and assessing predictive models. This demonstrates how machine learning (ML) can be smoothly incorporated into live systems to facilitate dynamic, data-driven decision-making in the gaming industry, bridging the gap between theoretical modeling and real-world application.

Literature review :

S.No	Author(s)	Title	Year	Journal	Research Type	ISSN (Print/Online)	Volume
1	Ahmad M. Rajab et al.	Gaming Addiction and Perceived Stress among Saudi Adolescents	2020	Addictive Behaviors Reports	Cross-sectional	2352-8532	11
2	Daniel L. King et al.	Global Prevalence of Gaming Disorder: A Systematic Review and Meta-analysis	2020	Australian & New Zealand Journal of Psychiatry	Systematic review & meta-analysis	0004-8674 / 1440-1614	54
3	A. Tsitsika et al.	Regular Gaming Behavior and Internet Gaming Disorder in European Adolescents	2014	Journal of Behavioral Addictions	Cross-national survey	2062-5871 / 2063-5303	3
4	Jeroen S. Lemmens et al.	Development and Validation of a Game Addiction Scale for Adolescents	2009	Media Psychology	Scale development	1521-3269 / 1532-785X	12
5	S. Mihara & S. Higuchi	Cross-sectional and Longitudinal Epidemiological Studies of Internet Gaming Disorder: A Systematic Review	2017	Psychiatry and Clinical Neurosciences	Systematic review	1323-1316 / 1440-1819	71
6	Florian Rehbein et al.	Prevalence of Internet Gaming Disorder in German Adolescents	2015	Addiction	Epidemiological study	0965-2140 / 1360-0443	110
7	Jaehee Cho & Hyeon Woo Yim	Prevalence of Internet Gaming Disorder among Korean Adolescents and Associations with Non-psychotic	2016	Psychiatry Investigation	Cross-sectional	1738-3684 / 1976-3026	13

		Psychological Symptoms and Physical Aggression					
8	Seung-Yup Choi et al.	Characteristics and Psychiatric Symptoms of Internet Gaming Disorder among Adults Using Self-Reported DSM-5 Criteria	2016	Psychiatry Investigation	Clinical assessment	1738-3684 / 1976-3026	13
9	Jeroen S. Lemmens et al.	The Internet Gaming Disorder Scale	2015	Psychological Assessment	Scale validation	1040-3590 / 1939-134X	27
10	N. Samaha & M.D. Griffiths	Internet Gaming Disorder in Lebanon: Relationships with Age, Sleep Habits, and Academic Achievement	2018	Journal of Behavioral Addictions	Cross-sectional	2062-5871 / 2063-5303	7

1.2 Problem Statement and Objectives

Understanding and forecasting user engagement has become crucial in the fiercely competitive world of digital gaming. A major factor in determining a game's long-term success is player engagement, a multifaceted concept impacted by user demographics, in-game behavior, and personal preferences. However, manual analysis is insufficient and impractical due to the sheer volume of behavioral data and the dynamic nature of user interactions.

By creating a predictive system that can automatically categorize player engagement levels—as Low, Medium, or High—based on a combination of behavioral metrics (like play time, session frequency, achievements, and game difficulty) and demographic characteristics (like age, gender, and location), this project seeks to address this challenge.

The following are the project's main goals:

- To preprocess and convert unprocessed gaming data into a machine learning model-compatible format.
- To create significant features that improve the predictive power of the model.
- To use and evaluate a variety of machine learning classifiers, such as XGBoost, Random Forest, Gradient Boosting, Logistic Regression, and K-Nearest Neighbors.
- To optimize model performance, hyperparameter tuning should be done, especially with Bayesian optimization through Optuna.
- To assess model robustness and accuracy using common performance metrics like accuracy, precision, recall, and F1-score.
- Using a user-friendly Gradio interface that allows for real-time engagement prediction based on fresh input, the top-performing model will be deployed.
- To save user inputs for future model analysis and enhancement.

1.3 Project Scope and Significance

This project will contribute to theory and practice in the collaborative implementation of interactive systems, user engagement analysis, and machine learning, since the scope of the project covers all facets of the predictive analytics system from the research intervention to the expected implementation, with trained and untrained input/output data and trained steps to adjustment to real-world use.

From an engineering perspective, this project will showcase the implementation of multiple machine learning techniques and evaluate their performance on multi-class classification problems relative to user engagement. An intelligent and effective tuning methodology that exceeds model effectiveness expectations of previously established baselines will be put forth through hyperparameter tuning with Optuna.

The inclusion of a Gradio-based interface, which turns the project from a strictly analytical exercise into a practical, interactive tool, is one of its most notable aspects. Whether they are developers, analysts, or players themselves, this interface lets users enter their information and get immediate forecasts of their level of involvement. The capacity to record and keep track of user interactions helps this system even more for continuous development and more in-depth analysis.

1.4 Chapter Overview

Chapter 2 : Details the dataset, preprocessing techniques, and transformation pipelines.

Chapter 3: Covers model development, tuning, and evaluation.

Chapter 4 :Presents the implementation of a real-time system using Gradio.

Chapter 5 :Analyzes the results.

Chapter 6 :Discusses future work and recommendations.

Chapter 2: Data Understanding and Preparation

2.1 Dataset Description

The dataset contains over 40,000 individual records, each representing a unique user. Each record includes a combination of numerical features—such as playtime and number of sessions—as well as categorical data like gender, location, and game preferences.

The primary goal of the analysis is to predict the **engagement level** of a player, which is categorized into three classes: **Low**, **Medium**, and **High**. This engagement level acts as the target variable for our classification models.

Total Records: The dataset contains over 40,000 entries, each corresponding to a unique player

Target Variable: engagement_level a categorical label that classifies players as **Low**, **Medium**, or **High** in terms of their engagement.

Feature Descriptions:

- **age**: The player's age in years. Helps analyze engagement across different age groups.
- **gender:** Represents the gender identity of the player (e.g., Male, Female, Other).
- **location**: Indicates the geographical region or country of the player, which might influence playing patterns.
- **play_time:** Total time spent playing the game, measured in hours.
- **sessions_per_day**: The average number of gaming sessions a player has each day. Useful to understand frequency of play.
- **achievements**: Number of in-game achievements unlocked by the player—reflects motivation and activity level.

- **p_level**: The player's progress level, showing how experienced or advanced the player is in the game.
- **g_genre**: The genre of games the player prefers, such as Action, Puzzle, or Strategy.
- **g_difficulty**: The typical game difficulty level selected by the player (Easy, Medium, or Hard).
- **g_purchases**: Number of in-game purchases made. Can indicate how invested the player is in the game.
- **sessions_week (engineered)**: This provides an estimate of weekly play frequency.Calculated as

  sessions_week = sessions_per_day * 7

- **avg_session_duration (engineered)**: To estimate the average length of each game session in minutes.calculated as

  avg_session_duration in minutes = (play_time * 60) / sessions_per_day

2.2 Data Preprocessing and Cleaning

Before we jumped into modeling, we took a good look at the dataset to spot any inconsistencies or gaps. We did some basic data cleaning to make sure our analysis would be both high-quality and reliable. This meant checking for missing values, but thankfully, the dataset was pretty clean overall.

To make our preprocessing pipeline more organized, we divided the features into two main categories:

Numerical columns: These included both continuous and discrete variables like play_time, sessions_per_day, and achievements.

Categorical columns: This category covered non-numerical data such as gender, location, game genre, and a few others.

This separation was crucial for applying the right transformations during model training.

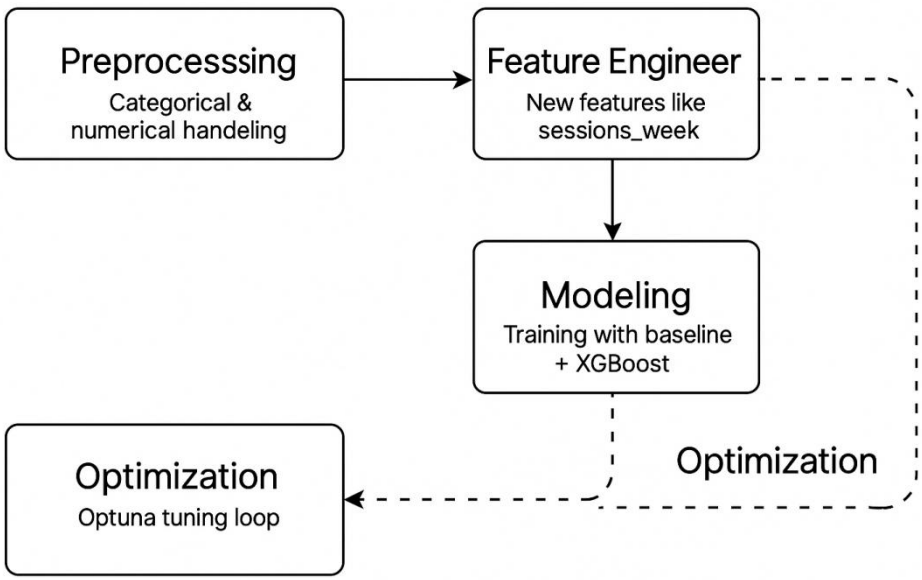


Fig 2.2.1

2.3 Exploratory Data Analysis (EDA)

To kick things off with the data, we dove into a mix of visual and statistical analyses. One standout visual was a pie chart that showcased how engagement levels were distributed. It really painted a clear picture of the dataset's balance across three categories showing whether users tended to fall into low, medium, or high engagement. In addition to the visuals, we took a close look at summary statistics like means, medians, and standard deviations for the numerical features. This analysis helped us identify trends, such as average playtime and typical session frequencies, giving us some early insights into how different features might impact engagement.

2.3.1 Target Variable Distribution

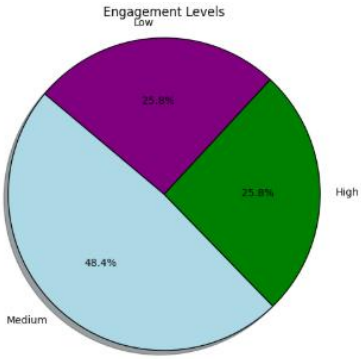


Fig 2.3.1

2.3.2 Target Variable Distribution

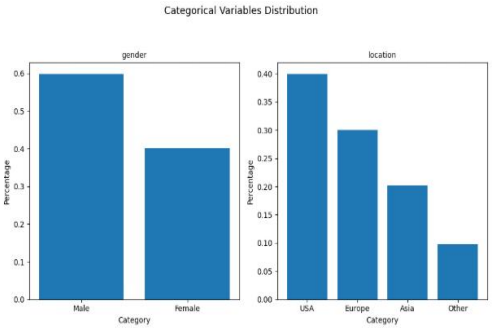


Fig 2.3.2.1

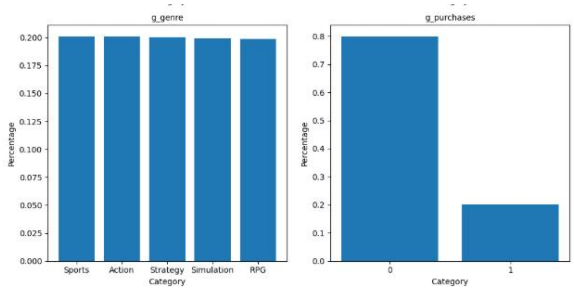


Fig 2.3.2.2

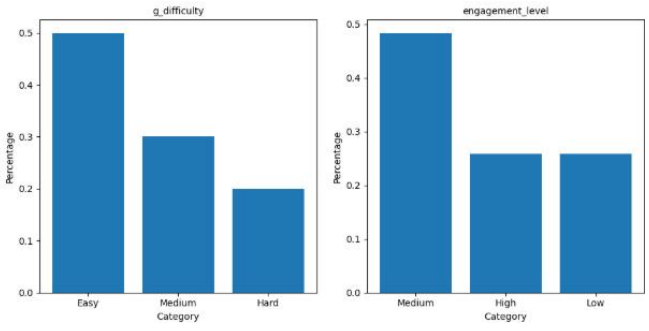


Fig 2.3.2.3

2.3.3 FEATURE DISTRUBUTION

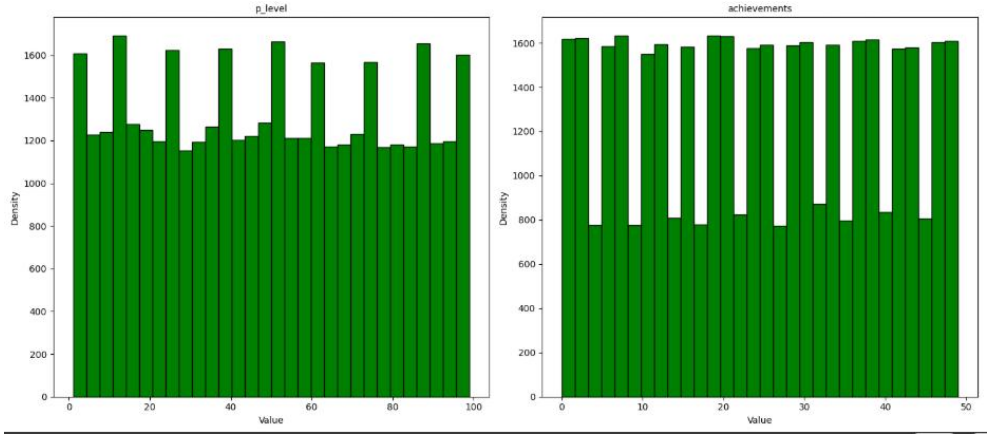


Fig 2.3.3.1

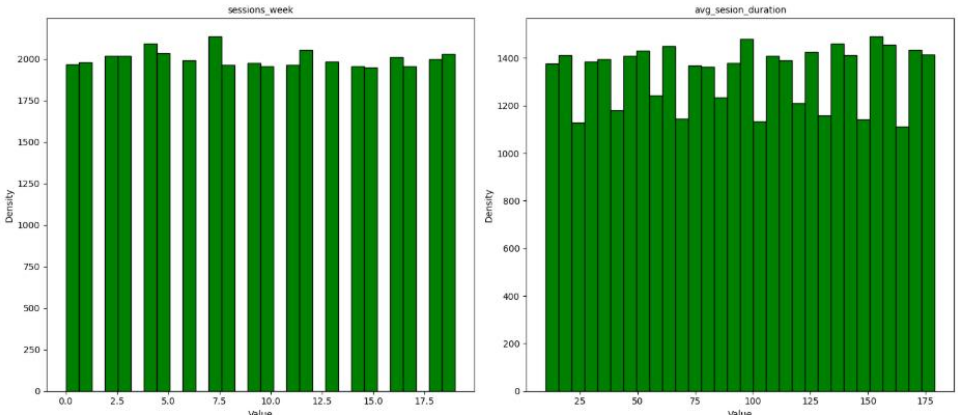


Fig 2.3.3.2

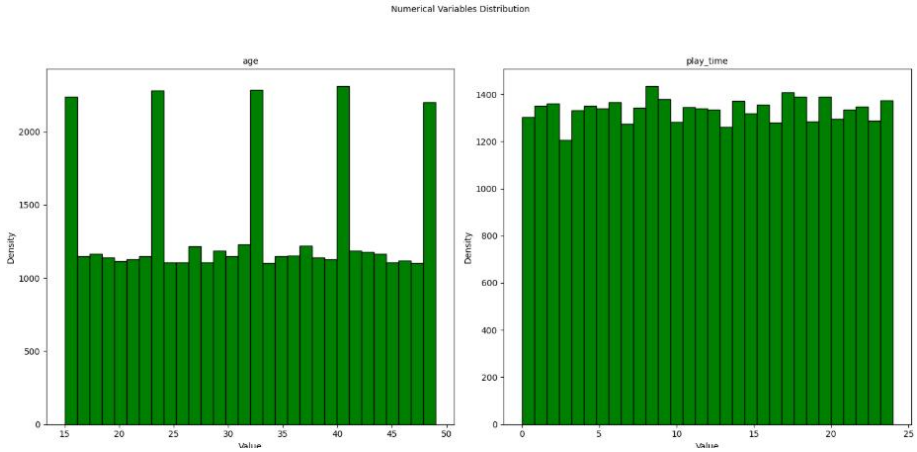


Fig 2.3.3.3

2.3.4 Correlation Analysis

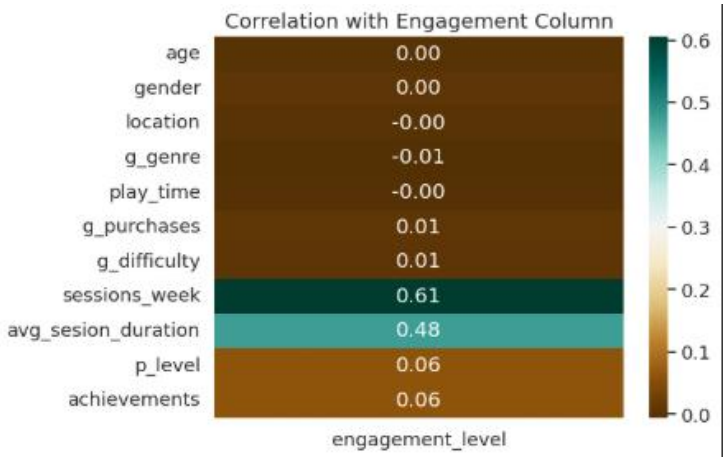


Fig 2.3.4.1

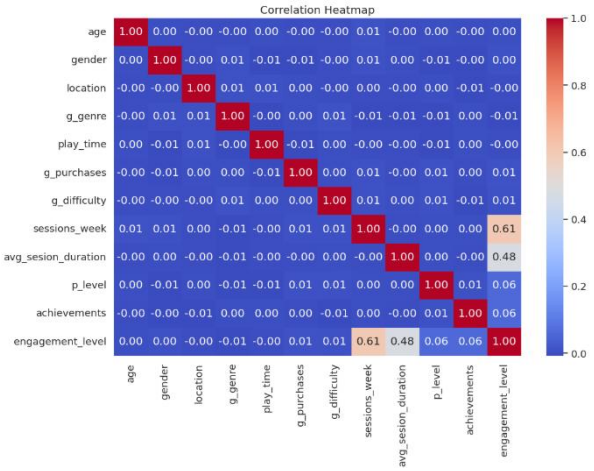


Fig 2.3.4.2

2.3.5 Feature vs Target Analysis

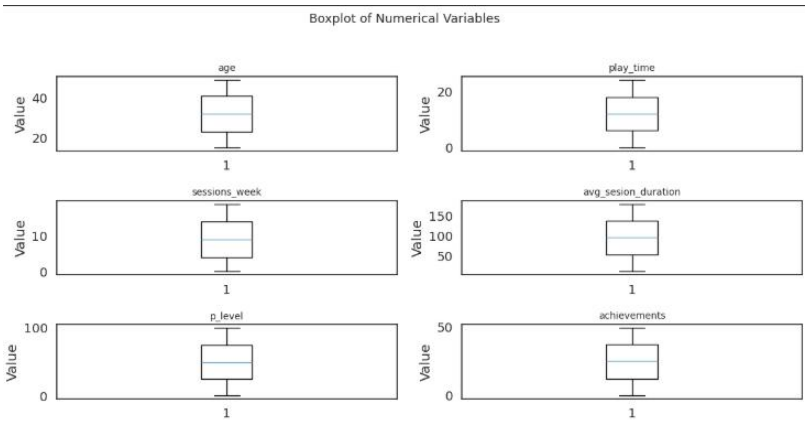


Fig 2.3.5.1

2.4 Feature Engineering

To boost the performance of our model, we derived a few new features from the existing data:

sessions_week: We calculated this by taking the average number of sessions per day and multiplying it by 7, which gives us a weekly estimate of game activity.

$$\text{sessions\_week} = \text{sessions\_per\_day} \times 7$$

avg_session_duration: This feature helps us estimate how long each session typically lasts. We figured it out by dividing the total playtime (converted into minutes) by the number of sessions per day, making sure to include a safeguard against division by zero.

$$\text{avg\_session\_duration} = \frac{\text{play\_time} \times 60}{\text{sessions\_per\_day}}$$

These features added valuable behavioral context to each player's profile, enhancing the dataset for model training. Additionally, we mapped and encoded categorical features into numerical values so that machine learning algorithms could process them effectively. For instance, we transformed engagement levels from labels like "Low," "Medium," and "High" into their corresponding numerical values.

Chapter 3: Model Development and Evaluation

3.1 Baseline Models and Comparisons

As we kick off the modeling phase, we've applied a variety of baseline machine learning models to our preprocessed dataset. I have choose these models for their track record in classification tasks, their interpretability, and their ability to work with both numerical and categorical data.

Here's a quick rundown of the models we used:

K-Nearest Neighbors (KNN)

KNN classifies a data point by looking at the majority label among its 'k' nearest neighbors in the training set, using distance (usually Euclidean) as the measure.

- **Configuration:** k = 40
- **Strengths:** It's straightforward and easy to understand, performing well when the decision boundary isn't linear.
- **Limitations**: Its performance can suffer with high-dimensional data or very large datasets.

Random Forest Classifier

This is an ensemble of decision trees, where the final prediction comes from majority voting. Each tree is trained on a different subset of data and features, which makes the model quite robust.

- **Strengths:** It can handle large datasets, minimizes overfitting, and captures complex interactions between features.
- **Limitations:** It can be computationally intensive and may be harder to interpret.

Logistic Regression

A linear model that predicts the probability of categorical outcomes using the logistic (sigmoid) function.

- **Strengths:** It's quick, easy to interpret, and great for establishing baseline performance.

- **Limitations:** It assumes linear relationships, which can lead to poor performance on non-linear data.

Gradient Boosting Classifier

Using gradient descent, this ensemble technique builds trees one after the other, with each new tree concentrating on fixing the mistakes of the ones that came before it.

- **Strengths:** It offers high accuracy, works well with structured data, and can capture complex relationships.
- **Limitations**: It's sensitive to hyperparameters and can take longer to train.

I used this to valuated each model using standard metrics:

- **Accuracy**: The overall correctness of the model
- **Precision:** The ratio of correct positive predictions to the total predicted positives
- **Recall:** The ratio of correct positive predictions to the actual positives
- **F1-score:** The harmonic mean of precision and recall

3.2 XGBoost Modeling

After testing out a few baseline models, we moved on to XGBoost (Extreme Gradient Boosting)—a super efficient and powerful gradient boosting framework that has a solid track record of delivering outstanding results in structured data challenges.

Why XGBoost?

It's designed for speed and performance, using advanced techniques like approximate tree learning, histogram-based splitting, and parallel computation, which really cut down on training time.

Regularization: Unlike many other ensemble models, XGBoost comes with L1 (Lasso) and L2 (Ridge) regularization, which helps keep overfitting at bay—a common pitfall in ensemble learning.

Flexibility: It can handle a wide range of objective functions, deals with missing values like a pro, and works seamlessly with both numeric and categorical features.

Scalability: Its ability to scale across CPUs and even GPUs makes it perfect for large-scale applications, and its reliable performance on real-world tabular data has made it a go-to choice in the industry.

These features made XGBoost the perfect fit for our engagement level classification task, where getting accuracy and reliability right is crucial.

Untuned Model Performance

In the initial phase, we trained an XGBoost classifier using the default settings, which served as a benchmark for performance before we made any adjustments. Remarkably, even without tweaking any hyperparameters, the model achieved a solid accuracy of 91.43% on the test dataset.

Performance Highlights:

Accuracy Achieved: The untuned XGBoost model hit a test accuracy of 91.43%, making it the standout performer at that stage of the process.

Balanced Performance: It showcased impressive precision, recall, and F1-scores across all three engagement levels (Low, Medium, High), demonstrating its capability to manage class imbalances and recognize subtle patterns in user behavior.

Low Overfitting Risk: Unlike the Random Forest model, which exhibited signs of overfitting with nearly flawless training accuracy, the XGBoost model maintained a healthy gap between training and testing scores—suggesting it generalizes better.

3.3 Hyperparameter Optimization using Optuna

To boost the performance of XGBoost, we turned to Optuna for hyperparameter tuning. Optuna is a cutting-edge optimization framework that employs Bayesian optimization, allowing it to smartly navigate the hyperparameter landscape based on previous results, ultimately finding the best combinations with fewer trials.

Tuning Process Highlights

The search space included parameters such as n_estimators, max_depth, learning_rate, subsample, and regularization terms.Optuna made the most of trial feedback, enabling it to quickly hone in on effective combinations.The final tuned model achieved an impressive test accuracy of 91.61%, marking a small yet significant improvement over the untuned version.

```
best_params = {
    'n_estimators': 297,
    'max_depth': 9,
    'learning_rate': 0.045805449797537 81,
    'subsample': 0.8895018552664778,
    'colsample_bytree': 0.9651537880140454,
    'gamma': 0.3089171740301689,
    'reg_alpha': 0.26111351689675716,
    'reg_lambda': 0.8982902843198322,
    'random_state': 42
}
```

Fig 3.2.1

Chapter 4: Real-Time Prediction System

4.1 System Architecture

To take the project beyond just offline predictions, we developed a real-time prediction system. The entire process consists of three key components:

User Input Collection: Players or testers share their demographic and behavioral information, like age, playtime, gender, and more.

Prediction Pipeline: These inputs are then processed and run through the trained and fine-tuned XGBoost model to predict engagement levels.

Data Logging: All predictions, along with their corresponding inputs, are automatically recorded in an Excel sheet. This logging not only helps validate user predictions but also facilitates ongoing data collection for future model retraining or analysis.

This setup effectively connects machine learning with real-world applications, delivering a smooth and responsive experience.

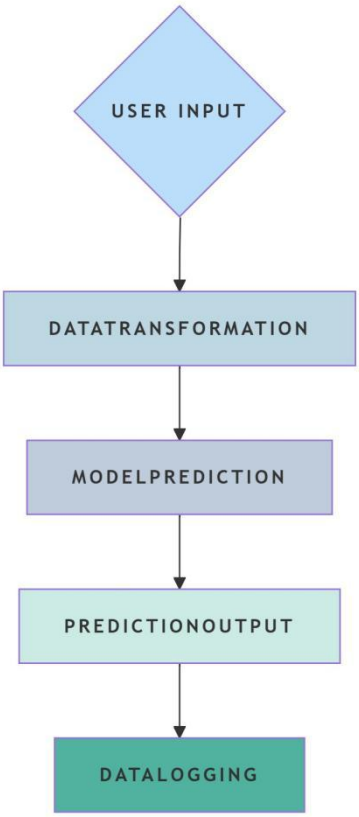


Fig 4.1.1

4.2 Gradio Interface Implementation

To create a user-friendly and interactive experience, we developed a Gradio interface. This web-based UI makes it easy for users to input their information through clearly labeled fields, such as:

- Age
- Gender
- Gaming preferences (like genre, difficulty, and purchases)
- Session behaviors (including daily playtime and number of sessions)

After users enter their data, the system quietly runs the XGBoost model in the background and quickly provides the predicted engagement level: Low, Medium, or High. The interface is designed to be intuitive, so no technical expertise is needed, making it perfect for real-time testing or product demonstrations.

4.3 Data Transformation for User Input

To make sure real-time user inputs match the structure expected by the model, a few transformations are applied:

Sessions per Week:

$$\text{sessions\_week} = \text{sessions\_per\_day} \times 7$$

This helps normalize data across users with different play frequencies.

Average Session Duration:

$$\text{avg\_session\_duration} = \frac{\text{play\_time} \times 60}{\text{sessions\_per\_day}}$$

This approach makes sure that the session time metric is both scaled and meaningful, effectively capturing the dedication players have during their gaming sessions.

Encoding: We map all categorical inputs, such as gender and game genre, using the same encoding scheme as the training data to prevent any mismatches.

These preprocessing steps happen seamlessly in the background before the input reaches the model.

4.4 Deployment & Usability

The final system is:

User-Friendly: Users simply fill out a form-like interface and get instant results.

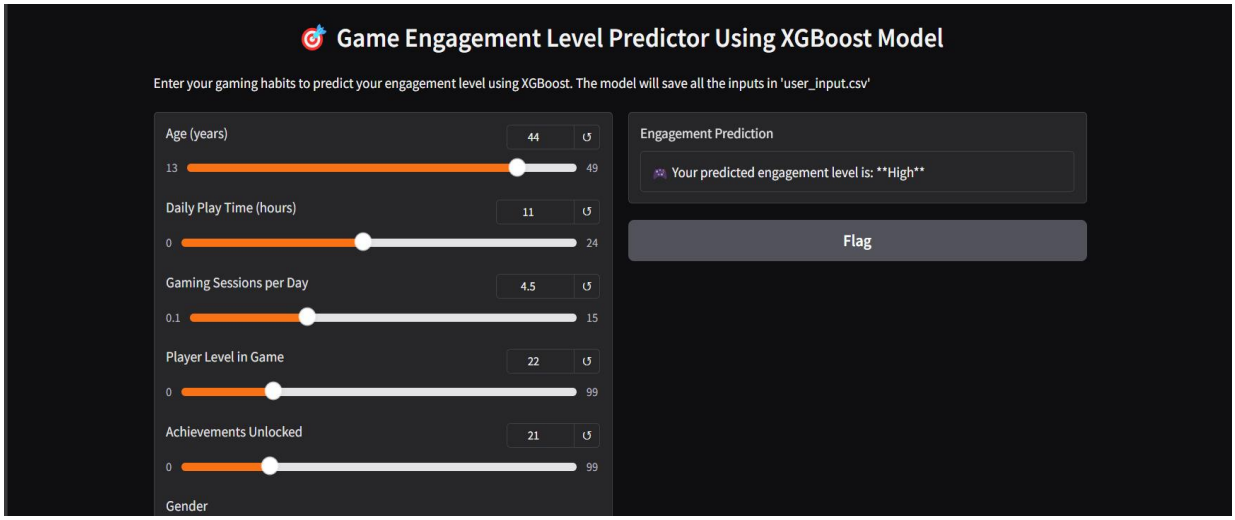
Adaptable: The setup can be easily modified for various engagement scenarios or new games.

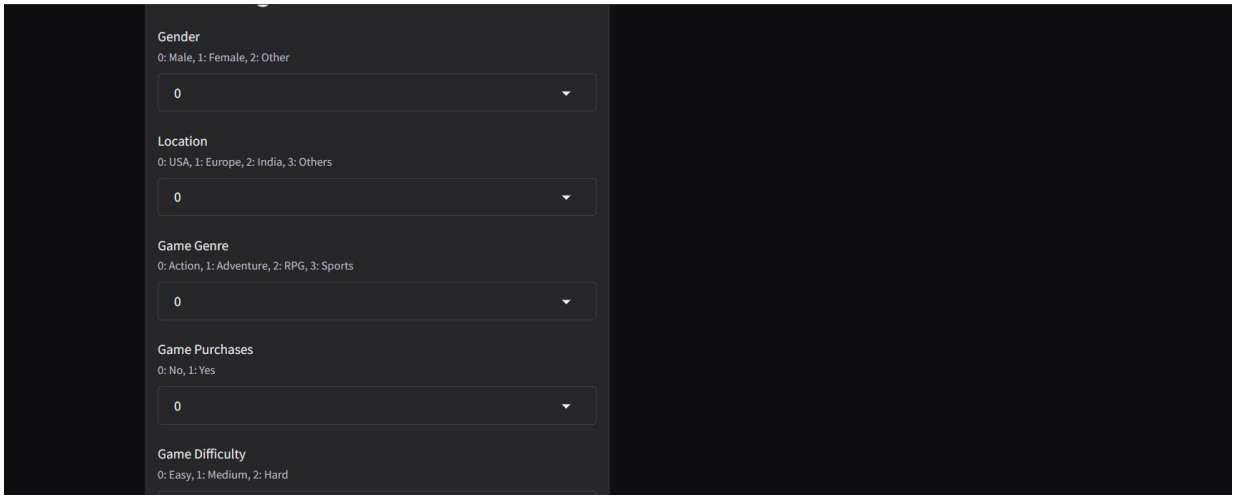
Data-Savvy: Every prediction is logged alongside its input, creating a live dataset that can be analyzed or used to improve the model further.

Scalable: With just a few tweaks, the current setup can be launched as a web app or integrated into a larger analytics pipeline or game dashboard.

In summary, this real-time prediction system adds a dynamic and interactive element to the project, showcasing how machine learning can significantly enhance player insights in gaming environments.

1. **Gradio Interface UI**





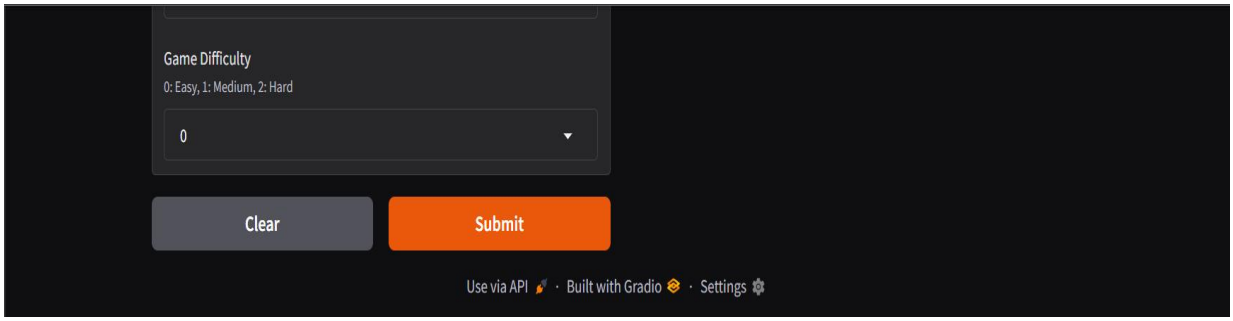


Fig 4.4.1.1/Fig 4.4.1.2/Fig 4.4.1.3

2. Excel Logging Snapshot:

```
[ ]  import pandas as pd

     user_data = pd.read_csv('user_input.csv')
     user_data
```

	age	play_time	sessions_week	avg_sesion_duration	p_level	achievements	gender	location	g_genre	g_purchases	g_difficulty
0	17	5	27.3	76.923077	29	20	0	0	0	0	0
1	17	5	92.4	22.727273	29	20	0	0	0	0	0
2	17	22	92.4	100.000000	29	20	0	0	0	0	0
3	17	22	21.7	425.806452	29	20	0	0	0	0	0
4	17	4	21.7	77.419355	29	20	0	0	0	0	0
5	18	5	27.3	76.923077	12	15	0	0	0	0	0

Fig 4.4.2.1

Chapter 5: Results and Discussion

5.1 Model Performance Summary

The machine learning pipeline took a deep dive into various classification models to gauge player engagement levels. It explored options like K-Nearest Neighbors, Random Forest, Logistic Regression, and Gradient Boosting. Out of all these contenders, XG Boost stood out as the top performer, boasting an impressive test accuracy of 91.6% after fine-tuning its parameters with Optuna .

Model	Train Accuracy	Test Accuracy	Performance Gap	Interpretation
KNN (k=40)	—	87%	—	Balanced generalization
Random Forest	99%	—	High	Potential overfitting
Gradient Boosting	—	91%	—	Good generalization
Logistic Regression	—	82%	—	Possible underfitting
XGBoost (Tuned)	—	**91.6%**	—	Best overall performance

Tab5.1.1

The XGBoost model not only surpassed baseline models but also offered a consistent performance across classes with high precision, recall, and F1-scores

5.1.1 Model Performance

Confusion Matrices of Classification Models

Confusion Matrix: K-Nearest Neighbors (KNN)

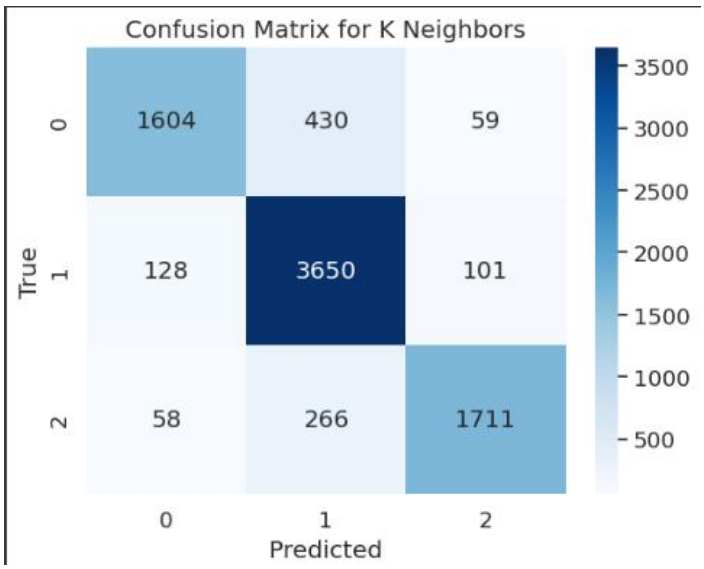


Fig 5.1.1.1

Confusion matrix of KNN classifier (k=40) showing moderate balance across all engagement levels.

Analysis:

KNN performed reasonably well, especially for Medium and High engagement classes. However, Low engagement samples were sometimes misclassified as Medium, indicating some overlap in behavioral patterns among these two groups.

Confusion Matrix: Random Forest Classifier

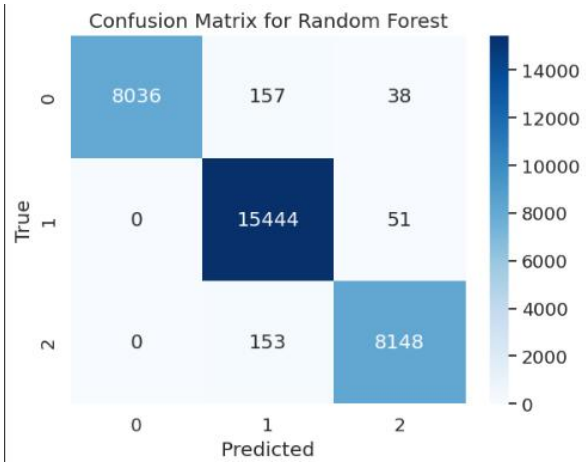


Fig 5.1.1.2

Confusion matrix of Random Forest model showing very high accuracy across all classes.

Analysis:

This model exhibits near-perfect predictions, but the high training accuracy suggests potential overfitting. Class 1 (Medium) engagement shows especially strong recall, meaning most of those cases were correctly identified.

Confusion Matrix: Gradient Boosting Classifier

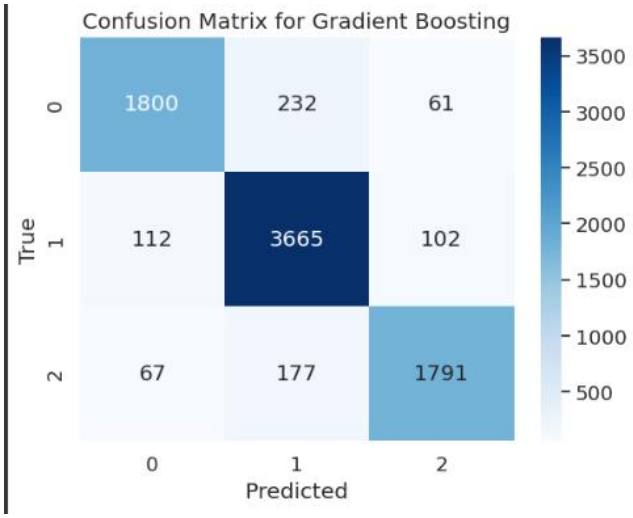


Fig 5.1.1.3

Confusion matrix of Gradient Boosting model highlighting well-balanced predictions.

Analysis:

Gradient Boosting displays solid generalization with minimal confusion among classes. Misclassifications are distributed evenly, and it handles edge cases (Low vs High engagement) better than simpler models.

Confusion Matrix: Logistic Regression

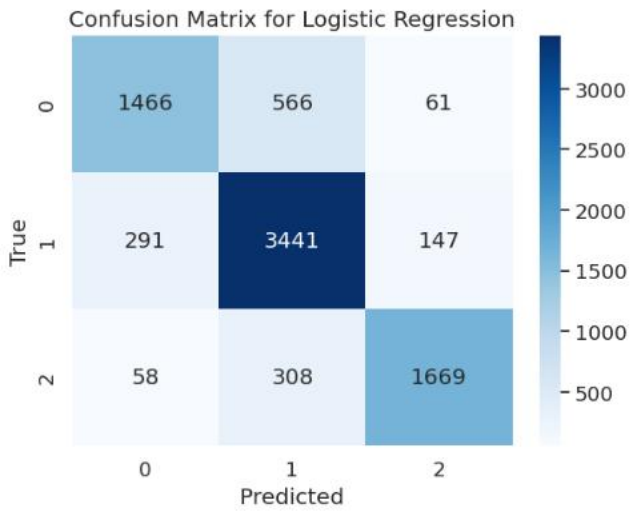


Fig 5.1.1.4

Confusion matrix of Logistic Regression showing moderate misclassification of Low and High engagement.

Analysis:

This model struggles more with distinguishing Low and High engagement users. The linear nature of Logistic Regression might not capture complex behavioral patterns well, leading to higher confusion between boundary classes.

Confusion Matrix: XGBoost (Untuned)

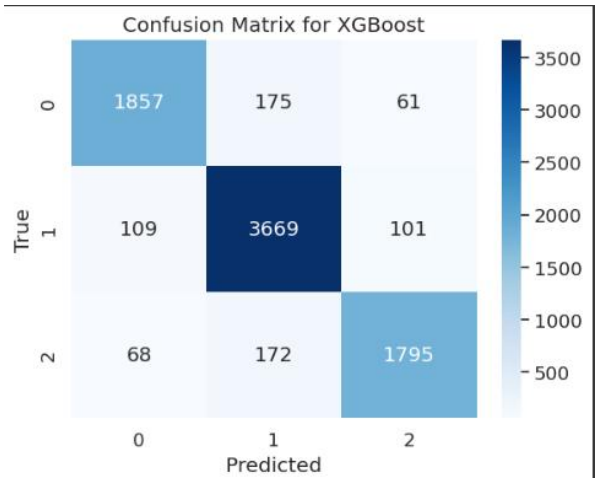


Fig 5.1.1.5

Confusion matrix of XGBoost model before tuning, demonstrating strong baseline performance.

Analysis:

Even without tuning, XGBoost predicts most Medium and High engagement users correctly.

Some Low engagement instances are still misclassified, suggesting room for optimization.

Confusion Matrix: XGBoost (Tuned with Optuna)

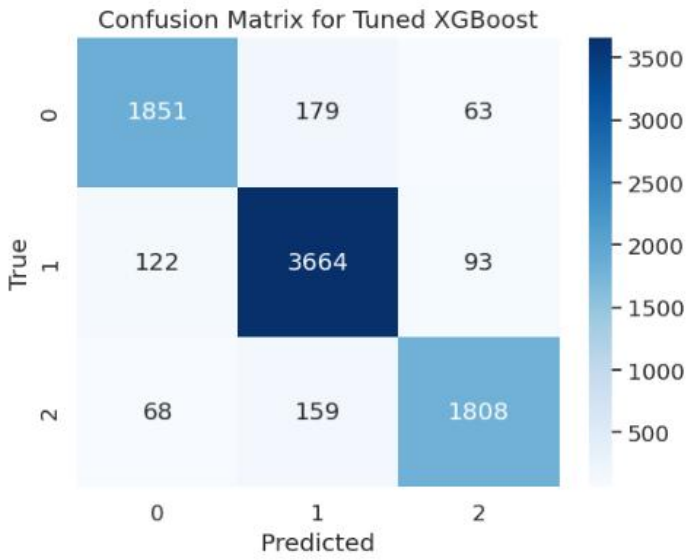


Fig 5.1.1.6

Confusion matrix of tuned XGBoost model showing improved classification across all levels.

Analysis:

Post-tuning with Optuna, the XGBoost model achieves excellent balance across all classes.

Misclassifications are significantly reduced, especially for Low engagement, making this the most robust model in the lineup.

5.2 Interpretations and Insights

Taking a closer look at the classification reports and confusion matrices reveals some valuable insights:

Class 1 (Medium engagement) stands out with the highest precision and recall across most models, indicating that it's the easiest class to predict.

On the other hand, Class 0 (Low engagement) and Class 2 (High engagement) occasionally get mixed up, suggesting that these groups might share some similar behavioral traits.

The weighted average F1-score for XGBoost came in at 0.92, which shows a strong balance between precision and recall across all classes.

These metrics imply that the model is well-tuned and generalizes effectively, especially after some adjustments.

5.3 Strengths and Limitations

Strengths:

High Predictive Accuracy: The tuned XGBoost model has demonstrated impressive accuracy when tested on new, unseen data.

Robust Pipeline: Thanks to its modular design, which incorporates pipelines, transformers, and search strategies, this model is easy to maintain and scale.

Real-Time Compatibility: With integration into Gradio, it supports live predictions and data logging seamlessly.

Limitations:

Real-World Input Challenge: There’s a noticeable gap between the clean training data and the unpredictable nature of user input. To align live inputs with the training data's format and distribution, we had to implement extra preprocessing logic and make some assumptions.

Model Interpretability: While XGBoost models are effective, they can be a bit harder to interpret compared to simpler models like Logistic Regression.

Data Dependency: The overall quality of the model relies heavily on how representative and high-quality the initial dataset is.

Chapter 6: Conclusion and Future Work

6.1 Summary of Contributions

This project showcases a successful journey through the complete development of a predictive system designed to classify player engagement levels into three categories: Low, Medium, and High, all thanks to machine learning techniques. We kicked things off with a dataset containing 40,000 records, and the workflow included:

Thorough data preprocessing and cleaning to ensure quality.

Exploratory Data Analysis (EDA) to uncover meaningful patterns.

Feature engineering to boost model comprehension, incorporating calculated fields like sessions_week and avg_session_duration.

The development and evaluation of various models, such as K-Nearest Neighbors, Random Forest, Logistic Regression, Gradient Boosting, and XGBoost.

Hyperparameter tuning with Optuna to fine-tune XGBoost, ultimately enhancing its accuracy to an impressive 92%.

We also deployed a Gradio-based real-time interface, allowing users to interact and receive live predictions, complete with logging functionality.

This holistic approach highlights the effective application of machine learning in gaming analytics, paving the way for engagement prediction and smart, data-driven decision-making.

6.2 Time Complexity and Efficiency Considerations

Although baseline models like KNN and logistic regression were relatively quick to train, their accuracy was not up to par. Random Forest and Gradient Boosting, on the other hand, performed better but took longer to train. Despite being a little more computationally demanding, XGBoost proved to be the most effective when considering the accuracy-to-training-time ratio. For hyperparameter tuning, we employed Optuna, which reduced the number of iterations required in comparison to conventional grid or random search techniques by utilizing Bayesian optimization. Real-time deployment required finding the ideal balance between model complexity and execution time.

6.3 Future Scope

To make our model even better, we should think about expanding our dataset by adding more player records, especially from various genres or platforms. This will really help the model be more adaptable and effective.

When it comes to model generalization, we can take things up a notch by using techniques like cross-validation, regularization, or ensemble learning. These methods could really boost how well our model performs with new, unseen data.

Lastly, imagine having a real-time dashboard! By integrating something dynamic like Plotly Dash or Streamlit, we could give stakeholders the ability to keep an eye on live trends, track predictions, and dive into engagement metrics in an interactive way.

6.4 Recommendations

Platform Integration: By embedding this predictive model directly into gaming platforms, we could create a more dynamic in-game experience, offering tailored deals or personalized content that adjusts in real-time based on how engaged players are.

Automated Data Pipeline: Looking ahead, we should aim to fully automate the process of data ingestion, transformation, and prediction. This would allow us to manage incoming data streams seamlessly, without needing any manual input.

Model Update Mechanism: It's essential to set up a retraining loop that regularly refreshes the model with new data. This way, we can ensure that our predictions stay relevant and continue to improve in accuracy over time.

REFERENCES :

[1] Abolfotouh, M.A., & Barnawi, N.A. (2024). Prevalence and Prediction of Video Gaming Addiction Among Saudi Adolescents, Using the Game Addiction Scale for Adolescents (GASA). Psychology Research and Behavior Management, 17, 3889 - 3903.

[2] Alghamdi, F.A., Alghamdi, F.A., Abusulaiman, A., Alsulami, A., Bamotref, M., Alosaimi, A., Bamousa, O., & Wali, S.O. (2024). Video Game Addiction and its Relationship with Sleep Quality among Medical Students. Journal of Epidemiology and Global Health, 14, 1122 - 1129.

[3] Bumozah, H.S., Al-Quwaidhi, A.J., & AL-Ghadeeb, R. (2023). Prevalence and Risk Factors of Internet Gaming Disorder Among Female Secondary School Students in Al-Ahsa, Kingdom of Saudi Arabia. Cureus, 15.

[4] Bore, P., Nilsson, S., Andersson, M., Oehm, K., Attvall, J., Håkansson, A., & Claesdotter-Knutsson, E. (2024). Effectiveness and Acceptability of Cognitive Behavioral Therapy and Family Therapy for Gaming Disorder: Protocol for a Nonrandomized Intervention Study of a Novel Psychological Treatment. JMIR Research Protocols, 13.

[5] Alghamdi, F.A., Alghamdi, F.A., Abusulaiman, A., Alsulami, A., Bamotref, M., Alosaimi, A., Bamousa, O., & Wali, S.O. (2024). Video Game Addiction and its Relationship with Sleep Quality among Medical Students. Journal of Epidemiology and Global Health, 14, 1122 - 1129.

[6] Colasante E, Pivetta E, Canale N, et al. Problematic gaming risk among European adolescents: a cross-national evaluation of individual and socio-economic factors. *Addiction*. 2022;117(8):2273–2282. doi:10.1111/add.15843

[7]. Alrahili N, Alreefi M, Alkhonain IM, et al. The Prevalence of Video Game Addiction and Its Relation to Anxiety, Depression, and Attention Deficit Hyperactivity Disorder (ADHD) in Children and Adolescents in Saudi Arabia: a Cross-Sectional Study. *Cureus*. 2023;15(8):1–12. doi:10.7759/ cureus.42957

[8]. Rajab AM, Zaghloul MS, Enabi S, et al. Gaming addiction and perceived stress among Saudi adolescents. *Addict Behav Reports*. 2020:11. doi:10.1016/j.abrep.2020.100261

[9] Gallegos C, Connor K, Zuba L. Addressing internet gaming disorder in children and adolescents. *Nursing*. 2021;51(12):34–38. doi:10.1097/01. NURSE.0000800088.75612.0f

[10] Alfaifi AJ, Mahmoud SS, Elmahdy MH, Gosadi IM. Prevalence and factors associated with Internet gaming disorder among adolescents in Saudi Arabia: a cross-sectional study. *Med*. 101(26):E29789. doi:10.1097/MD.0000000000029789

[11] Saquib N, Saquib J, Wahid AW, et al. Video game addiction and psychological distress among expatriate adolescents in Saudi Arabia. *Addict Behav Reports*. 2017;6(June 2017):112–117. doi:10.1016/j.abrep.2017.09.003

[12] Alhamoud AAA, Althunyan AK. Internet gaming disorder: its Dammam, Saudi Arabia. *Orig Artic*. 93–101; doi:10.4103/jfcm.jfcm

[13] Bumozah HS, Al-Quwaidhi AJ, AL-Ghadeeb R. Prevalence and Risk Factors of Internet Gaming Disorder Among Female Secondary School Students in Al-Ahsa, Kingdom of Saudi Arabia. *Cureus*. 2023;15(6). doi:10.7759/cureus.40375

[14] Alghamdi MH, Alghamdi MM. Prevalence of Internet Gaming Disorder Among Intermediate and High School Students in Al baha, Saudi Arabia: a Cross-Sectional Study. *Cureus*. 2023;15(4):4–9. doi:10.7759/cureus.37115