Gaming Behavior Analysis and Prediction Report

Contents

Abstract

This report provides comprehensive analysis to predict players' gaming behavior. The machine learning algorithms used in this project are Logistic Regression, Random Forest, and Support Vector Machine (SVM). The study uses a good set of facts that includes player details, kinds of games played, and player's engagement with the game. The main goal is to see how well these algorithms can classify a player's engagement level as 'Low', 'Medium', or 'High'.

The analysis begins with data processing in which categorical features are encoded and numerical features are standardized. Exploratory data analysis is performed to understand the feature distribution. The machine learning models are then trained and evaluated using cross-validation and hyperparameter tuning to optimize performance.

The results indicate that the Random Forest program did the best out of the three, doing better than both Logistic Regression and SVM in predicting the right labels.

In summary, this report shows how machine learning algorithms can help predict player's engagement level from the player's details, providing actionable insights for game developers and marketers to tailor experiences and strategies based on player engagement.

Introduction

Gamers have been drastically increasing all over the world. The gaming industries have to keep up in order fulfil demands and deliver intriguing content to the players. In order to do so, predicting gaming behavior is one of crucial aspects of the gaming industries, as it offers valuable insights into player engagement, preferences, habits and patterns. Accurate predictions can greatly impact a player's experience, game design, and monetization strategies. The rapid advancement of technology has expanded access to gaming, eliminating restrictions based on time or location.

Variety of video games are popular among people as they not only serve as a good way to spend leisure time but also have positive impact. Researchers have indicated that engagement with video games correlates with the enhancement of technology-related competencies and cognitive functions, such as task-switching abilities (James W. Karle, 2010), response time (Bialystok, 2006), and decision-making skills (C. Shawn Green, 2010). These findings suggest that playing video games may contribute positively to the development of various cognitive and technological aptitudes.

Games have been part of humans' daily life since early age and this trend will only accelerate in the future. According to a report by Newzoo in 2022, the number of gamers is projected to reach 3.5 billion by 2025. This widespread availability of gaming has had a significant impact on businesses, with a contribution of $196.8 billion in 2022 expected to increase to $225.7 billion by 2025. (Laurence, 2023) The surge in online gaming has captured the interest of management scholars seeking to understand the remarkable performance of the industry.

Project Overview

With the increase in internet connectivity, the global gaming industry is experiencing increased demands. The predictive analysis using machine learning algorithms not only plays a vital role in targeted marketing but also in in-game customization and delivering tailored gaming experiences by analyzing the players' preferences and behavior in order to modify the in-game contents, rewards and challenges. This strategy is crucial as game developers can use the player behavior analytics to keep the player engaged by providing targeted rewards or incentives.

Dataset

The dataset used has been taken from Kaggle. It contains 40,034 records and 13 features capturing comprehensive metrics and demographics related to player behavior. Key features include:

- **PlayerID:** Unique identifier for each player.
- **Age:** Age of the player.
- **Gender:** Gender of the player.
- **Location:** Geographic location of the player.
- **GameGenre:** Genre of the game the player is engaged in.
- **PlayTimeHours:** Average hours spent playing per session.
- **InGamePurchases:** Indicator of in-game purchases (0 = No, 1 = Yes).
- **GameDifficulty:** Difficulty level of the game.
- **SessionsPerWeek:** Number of gaming sessions per week.
- **AvgSessionDurationMinutes:** Average duration of each gaming session in minutes.
- **PlayerLevel:** Current level of the player in the game.
- **AchievementsUnlocked:** Number of achievements unlocked by the player.
- **EngagementLevel:** Target variable indicating the level of player engagement categorized as 'High', 'Medium', or 'Low'.

Libraries Used

- **Pandas:** Offers data structures and data analysis tools for handling structured data.
- **Seaborn:** Simplifies the creation of informative and complex statistical graphics.
- **Matplotlib:** Enables the creation of static, animated, and interactive visualizations in Python.
- **Scikit-Learn Metrics**
    - classification_report: Generates a report showing the main classification metrics.
- **Scikit-Learn Model Selection**
    - train_test_split: Splits data into training and testing sets for model evaluation.
    - GridSearchCV: Performs hyperparameter tuning using cross-validation to find the best model parameters.

- **Scikit-Learn Preprocessing**
  - LabelEncoder: Converts categorical data into numerical format for machine learning models.
- **Scikit-Learn Models**
  - LogisticRegression
  - RandomForestClassifier
  - SVC
- **Scikit-Learn Utilities**
  - Pipeline: Facilitates the creation of machine learning pipelines to streamline model training and evaluation.

Goals

The primary goal of this project is to develop and evaluate models that can predict how engaged players are based on different gaming factors. The project focuses on:

- **Predicting Player Engagement:** Making machine learning models that can accurately predict if players are 'Low', 'Medium', or 'High' in engagement.
- **Comparing Algorithm Performance:** Testing and comparing how well different machine learning methods like Logistic Regression, Random Forest, and Support Vector Machines can classify player engagement levels.
- **Improving Model Performance:** Using hyperparameter tuning and cross validation to enhance the model's performance and its ability to predict accurate results for unseen data.

Such a predictive model with high accuracy will enable the developers and advertisers to make smart choices that improve the overall gaming experience and make players happier. This aids the game development and marketing more personalized, which leads to more engagement and success in the gaming world.

Methodology

Preprocessing of data

The data was first cleaned by exploring it. Any value that is null or with incorrect format was removed. Afterwards data was visualized to understand the complexity and pattern better. The dataset consists categorical data i.e., 'Gender', 'Location', 'GameGenre', 'GameDifficulty'. The target feature/label 'EngagementLevel' is also categorical data. This categorical data is encoded using Label Encoder. All the numerical data is then standardized using Standard Scaler.

Model Training and Evaluation

In this project, three machine learning algorithms were employed to classify player engagement levels: Logistic Regression, Random Forest, and Support Vector Machine (SVM). Each algorithm was chosen on the basis of its strengths and its ability to handle multiclass classification problems. The process included model training, hyperparameter tuning, and performance evaluation using Grid Search and cross-validation.

1. **Logistic Regression (Multinomial):**
   Logistic regression is one the widely used supervised machine learning algorithm. It is mostly used for binary classification tasks however it can also be used for multiclass classification. In the multinomial variant, it can handle multiple classes by using one-vs-rest or softmax techniques.

   Model Configuration

   - Model: LogisticRegression(multi_class='multinomial', solver='lbfgs')
   - Hyperparameters:
     - 'penalty': ['l2']
     - 'C': [0.1, 1, 10]

   It has been used in this project due to its simplicity and interpretability. Since the targeted label falls into more than two categories i.e., 'High', ' Medium' and 'Low' so 'multinomial' parameter is used. As the 'multinomial' option is supported only by the 'lbfgs', 'sag', 'saga' and 'newton-cg' solvers as stated in scikit-learn documentation. So, lbfgs was chosen as it can handle large datasets and complex models well. The penalty l2 can reduce the possibility of overfitting by penalizing the larger coefficients. The parameter 'C' controls the inverse of regularization which means smaller value offers greater regularization strength. Hence, the range was selected which helps in identifying the best trade-off between bias and variance. The hyperparameters were tuned using Grid Search which helps in exploring different values for regularization parameters to best fit the data.

2. **Random Forest**
   Random Forest is one of the popular ensemble learning method. It works by constructing multiple decision trees during training and merging their results to improve accuracy and control overfitting. It is used for both, the regression tasks as well as classification tasks, making it a versatile tool in machine learning toolkit.

   Model Configuration

   - Model: RandomForestClassifier()
   - Hyperparameters:
     - n_estimators: [50, 100, 200]
     - max_depth: [None, 10, 20]

It has been used in the project due to its robustness against overfitting and its ability to handle the high-dimensional data. The 'n_estimators' specify the number of trees. Larger number of trees offer better performance but leads to higher resources cost. So, the values were chosen with the goal of balancing the performance and computational efficiency. The parameter 'max_depth' controls maximum depth of trees. Deeper trees can capture more details however it is more prone to overfitting. Hence, the range was selected which can capture the most information without overfitting to the model. Grid Search was used to find the optimal combination of the parameters for best performance.

3. **SVM (Support Vector Machine)**

Support Vector Machine is supervised machine learning algorithm that handles classification and regression analysis tasks. It works by finding the optimal hyperplane that separates different classes in feature space. SVM can handle both linear and non-linear data by transforming input features into higher dimensional data by the use of kernel functions.

Model Configuration:

- **Model:** SVC()
- **Hyperparameters:**
  - kernel: ['linear', 'rbf']
  - C: [0.1, 1, 10]

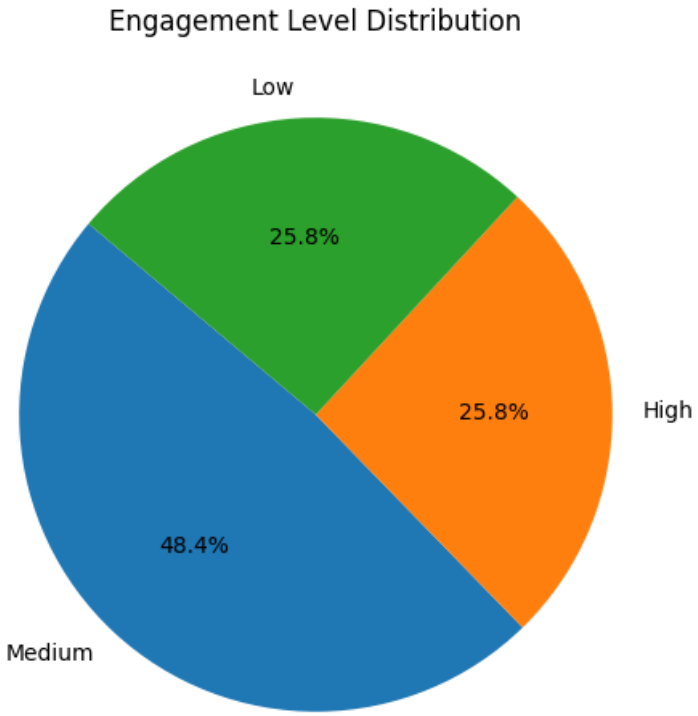
SVM was chosen due to its versatility in kernel functions and its effectiveness in high-dimensional data. The kernel function 'linear' is effective in linearly separable data. 'rbf' (Radial basis function) was also included due to its ability to handle non-linear data. Range of the parameter 'C' was selected to find the optimal balance between bias and variance as different values offer different regularization strengths. Grid Search was used for the optimal combination of parameter for best performance.

Results

Interpretations from Data

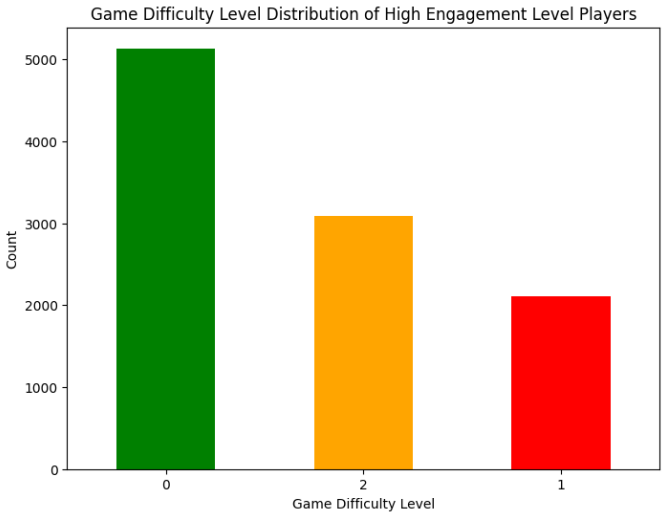
After the data exploration and analyzing the visualizations, interpretations can be made which can help the gaming industry's developers and advertisers to align the goals perfectly with the needs and interests of players.

Engagement Level Distribution



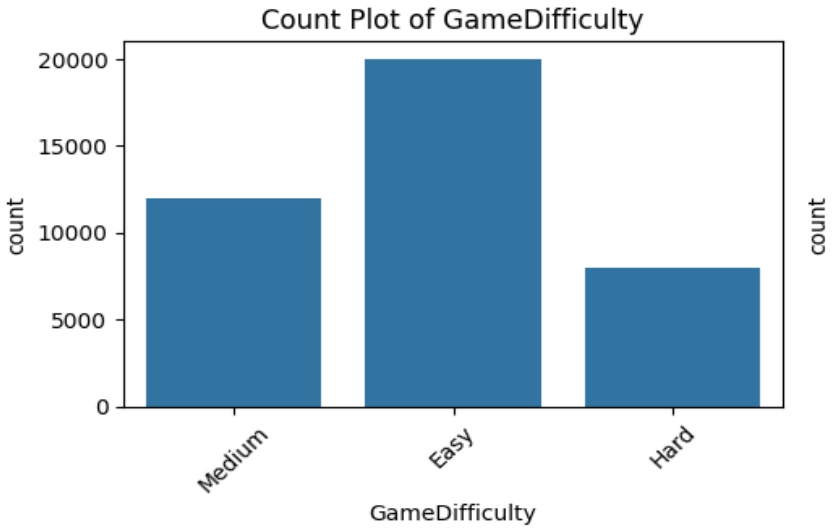
By analyzing the pie chart, the results of the captured data in dataset can be interpreted. Most of the players have the engagement level ‘Medium’ while the players with ‘High’ and ‘Low’ engagement levels are on par with one another. In order to have more in-depth analysis, other features can also be analyzed in the same way.

For instance, the difficulty level that high engagement players like to play.

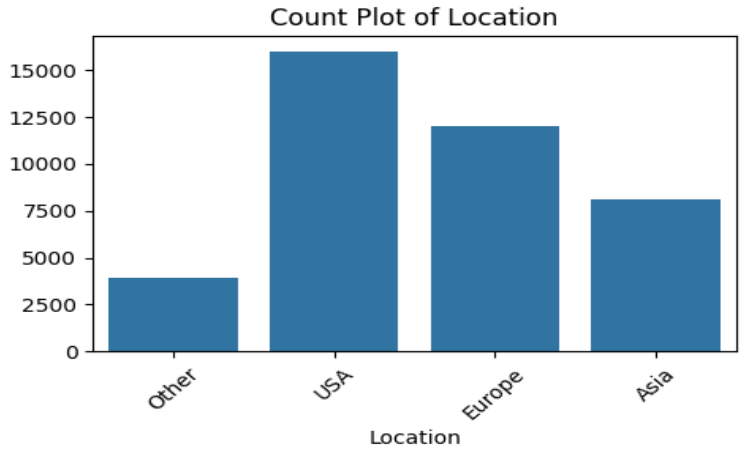


It can be seen that players with high engagement level like to play games with easier difficulty more rather than hard or medium difficulty.

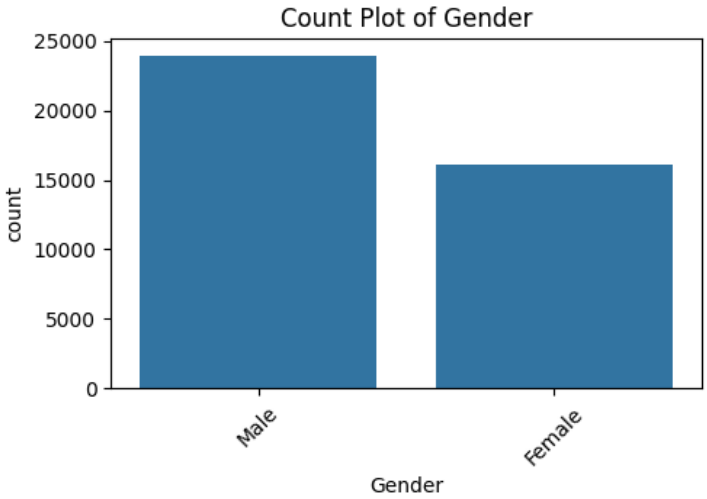
In general, the players tend to play games with easier and medium difficulty more rather than hard.



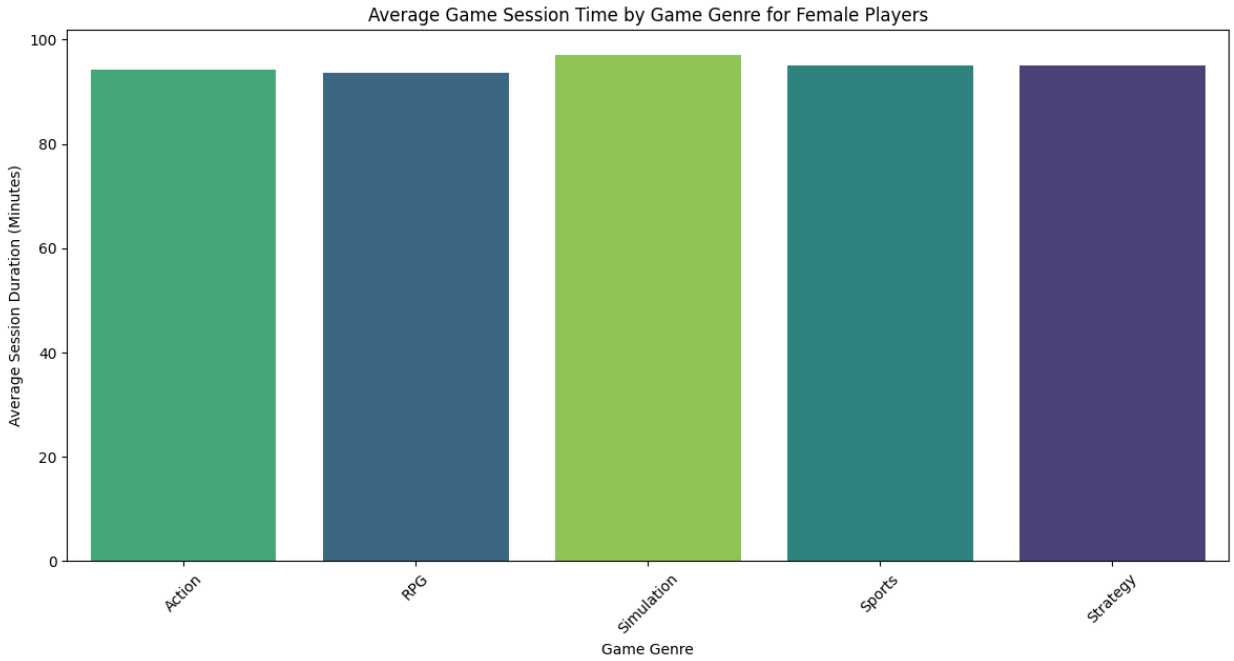
So, developers can change the game's difficulty to retain the player's interest and keep their satisfaction level up.



Location based analysis can provide insights that can aid the gaming industries to target specific location or region. As the above figure shows that according to dataset used in this project, there are more players from USA following Europe and Asia. It can help in making decisions such as releasing a new game in a specific country first. Further analysis can be made such as the most popular genre in specific location or region, the age range or gender. This insight can help in targeting and appealing more players.



Games are popular among both females and males. According to data collected and used in this project, there are more male players than female players. The strategies can be implemented to target more female players by looking at their stats such as the games that female players usually play or the genres that intrigue them. The gaming session analysis can be made for a more thorough targeting strategy.



Looking at the figure above, it can be seen that simulation games are slightly more popular among female player base than other genres. The survey conducted in 2017 also shows similar results. (Yee, 2017)

Similarly, an analysis can be made for male players base in the same way.

The dataset used in this project has many features that can be used for the analysis. One of them being in-game purchases. Players tend to enjoy more free content however an in-depth analysis can be made by looking at those who do spend money in the games by exploring the features too such as their age range, gender, location etc.

Performance of models

The models used i.e., Logistic Regression, Random Forest and Support Vector Machine all performed well on the given dataset however the Random Forest outperformed the other two.

Following are the classification reports of all three:

Classification Report for Logistic Regression:

	precision	recall	f1-score	support
0	0.81	0.70	0.75	2093
1	0.80	0.89	0.84	3879
2	0.89	0.82	0.85	2035
accuracy			0.82	8007
macro avg	0.83	0.80	0.81	8007
weighted avg	0.82	0.82	0.82	8007

Classification Report for Random Forest:

	precision	recall	f1-score	support
0	0.91	0.88	0.89	2093
1	0.90	0.95	0.92	3879
2	0.92	0.87	0.89	2035
accuracy			0.91	8007
macro avg	0.91	0.90	0.90	8007
weighted avg	0.91	0.91	0.91	8007

Classification Report for SVM:

	precision	recall	f1-score	support
0	0.90	0.84	0.87	2093
1	0.88	0.94	0.91	3879
2	0.92	0.86	0.89	2035
accuracy			0.89	8007
macro avg	0.90	0.88	0.89	8007
weighted avg	0.89	0.89	0.89	8007

From the results, Random Forest performs has better f1-score and accuracy than Logistic Regression and SVM. It means that Random Forest will give more accurate results on new, unseen data.

Discussion

Comparison

In this project, three machine learning models have been used: Logistic Regression, Random Forest, and Support Vector Machine (SVM). Each model was selected for its unique strengths and how well it suited the data.

Logistic Regression was chosen as the starting point because it is straightforward and easy to understand. It works well when the relationships between features and outcomes are relatively simple. However, when faced with more complicated patterns in the data, it might not be as effective.

Random Forest offered a more advanced approach. This model builds multiple decision trees and averages their results, which often leads to better accuracy. It is particularly good at handling complex relationships and interactions within the data. While it is not as easy to interpret as simpler models, its robust performance makes it a valuable tool.

However, incorrect parameters may cause additional computational cost or unsatisfactory performance.

Support Vector Machine (SVM) stands out for its ability to handle high-dimensional data and complex patterns through different kernels. It is great for dealing with intricate datasets, but it requires careful adjustment of its parameters and can be demanding on computational resources. Finding the right kernel and tuning the hyperparameters are important for getting the best results.

To sum up, Logistic Regression served as a useful baseline, Random Forest showed impressive versatility and accuracy, and SVM delivered satisfactory performance with the right tweaks. By evaluating these models, the most accurate one was Random Forest which showed high performance and accuracy.

Challenges

Several challenges were encountered that required thoughtful approaches while developing this project.

One of the initial challenges was to clean up the data. In order to remove any and every invalid and null values, different exploratory techniques were used. Missing values can significantly impact the model's performance so cleaning up data before the training is one the crucial steps. The data was also visually examined to find any disproportionate distributions among classes. This is to ensure that there is no imbalance in the data before training the model.

Another major step was encoding the categorical data into numerical data. In order for the machine learning algorithms to interpret the data, precise techniques are required. LabelEncoder was used in this project for the encoding task. This process was crucial for preparing the data for modeling, as machine learning algorithms typically require numerical input. Another step before the training was to standardize the input data. This is also one of important steps as it impacts the performance of the model.

Hyperparameter tuning was a critical step for optimizing model performance. Various hyperparameters for each model were explored using GridSearchCV, which involved testing multiple combinations to find the most effective combination of parameters. Although this process is time consuming and utilizes resources significantly, however it is still an important process in order to fine tune a model for best performance.

Another key challenge was ensuring that the models performed well on new, unseen data. To achieve this, cross-validation was employed which splits the data into subsets. Some subsets are used for training and other subsets for testing. This technique helped to ensure that the models generalized well to new data and, hence providing reliable results.

Conclusion

Summary of Key Findings

This project aimed to predict the gaming behavior using a dataset with various features, including demographic information, gaming preferences, and engagement levels. Three machine learning models were employed, Logistic Regression, Random Forest, and Support Vector Machine (SVM), to analyze the data and identify patterns in gaming behavior.

The results indicated that Random Forest generally provided the best performance, followed by SVM and Logistic Regression. The performance metrics, such as accuracy, precision, and recall, varied across the models, highlighting the strengths and limitations of each approach.

Implications of the Results

The findings from this analysis have several implications for understanding gaming behavior:

- **Personalization of Gaming Experiences:** The ability to predict engagement levels accurately can help in designing personalized gaming experiences. For instance, game developers can tailor content and in-game features to enhance player engagement based on the predicted behavior. They can add reward system or another interactive feature to keep the engagement level high of players.
- **Marketing Strategies:** Insights into gaming preferences and engagement levels along with other features can help in developing marketing strategies. Targeted promotions and advertisements can be designed to appeal to different sectors of players keeping in view their interests and needs. For instance, developing ads and promotions featuring the most popular genres or hosting themed events to boost engagement and attract the new players.
- **Game Development:** Understanding the factors that influence player engagement can guide game design decisions and help to retain the players. Features that drive high engagement can be analyzed to study the patterns, while aspects that contribute to lower engagement can be reassessed and improved to satisfy existing player base as well as attracting new players.

References

Bialystok, E., 2006. Effect of bilingualism and computer video game experience on the Simon task. *anadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale,* 60(1), p. 68.

C. Shawn Green, A. P. D. B., 2010. Improved probabilistic inference as a general learning mechanism with action video games. *Current biology,* 20(17), pp. 1573-1579.

James W. Karle, S. W. J. M. S., 2010. Task switching in video game players: Benefits of selective attention but not resistance to proactive interference. *Acta Psychologica,* 134(1), pp. 70-80.

KHAROUA, R. E., 2024. *Predict Online Gaming Behavior Dataset.* [Online]
Available at: https://www.kaggle.com/datasets/rabieelkharoua/predict-online-gaming-behavior-dataset
[Accessed 2024].

Laurence, A. H. I. B. F. A., 2023. Video Game Engagement: A Passkey to the Intentions of Continue Playing, Purchasing Virtual Items, and Player Recruitment. *International Journal of Computer Games Technology.*

Yee, N., 2017. *Beyond 50/50: Breaking Down The Percentage of Female Gamers by Genre,* s.l.: s.n.