Barcelona School of Economics

Master's Degree in Data Science
Specialization in Data Science for Decision Making

"Exploring User Retention in Enhance VR: A Comprehensive Analysis using Predictive Models and Clustering Techniques"

Authors: Catalina Odizzio and Agostina Pissinis

Supervisors: Hannes Mueller, Jesús Cerquides

Date: July 2023

ABSTRACT IN ENGLISH (100 words)**:**

This study delves into understanding and predicting user engagement in Enhance VR, a virtual reality cognitive training application, through data-driven approaches. The dataset encompasses de-identified user data including demographic characteristics, mood and session related variables. Initial data exploration involves descriptive statistics, data visualization, and inferential statistics, assessing correlations between attributes and their effects on engagement and performance. Machine learning models including Random Forests and Gradient Boosting are developed to predict user engagement levels. K-Prototypes clustering is employed for segmentation, identifying distinct user groups based on behavioral and demographic attributes. This research informs the strategic design and content delivery of Enhance VR by identifying distinct user groups and predicting engagement patterns.

ABSTRACT IN CATALAN/ SPANISH (100 words)

Este estudio profundiza en la comprensión y predicción del compromiso del usuario en Enhance VR, una aplicación de entrenamiento cognitivo de realidad virtual, a través de un enfoque basado en datos. El conjunto de datos abarca usuarios no identificados, incluyendo características demográficas, de ánimo y relacionadas con sesiones. La exploración inicial de datos comprende estadísticas descriptivas, visualizaciones y estadísticas inferenciales, evaluando correlaciones entre atributos y sus efectos en el compromiso y rendimiento. Se desarrollan modelos de aprendizaje automático, Random Forest y Gradient Boosting entre otros, para predecir el nivel de compromiso del usuario. Empleamos K-Prototypes para la segmentación, identificando grupos distintos de usuarios basados en atributos conductuales y demográficos. Esta investigación informa el diseño estratégico y la entrega de contenido de Enhance VR al identificar distintos grupos de usuarios y predecir patrones de compromiso.

KEYWORDS IN ENGLISH (3)**:** Predictive Modeling, Machine Learning, User Segmentation.

KEYWORDS IN CATALAN/ SPANISH (3)**:** Modelado predictivo, Aprendizaje automático, Segmentación de usuarios.

Exploring User Retention in Enhance VR: A Comprehensive Analysis using Predictive Models and Clustering Techniques

Authors:
Catalina Odizzio
Agostina Pissinis

Supervisors:
Hannes Mueller
Jesús Cerquides

Teaching Assistant:
Elliot Motte

A thesis submitted in partial fulfillment of the requirements for the degree

Master in Data Science for Decision Making
Barcelona School of Economics

July 4, 2023

Contents

Abstract

This study delves into understanding and predicting user engagement in Enhance VR, a virtual reality cognitive training application, through data-driven approaches. The dataset encompasses de-identified user data with variables including demographic characteristics, mood and session related variables. Initial data exploration involves descriptive statistics, data visualization, and inferential statistics, assessing correlations between attributes and their effects on engagement and performance. Machine learning models including Random Forests and Gradient Boosting are developed to predict user engagement levels. K-Prototypes clustering is employed for segmentation, identifying distinct user groups based on behavioral and demographic attributes. By applying statistical analyses and machine learning techniques, insights into user demographics, engagement, and performance are explored. This research informs the strategic design and content delivery of Enhance VR by identifying distinct user groups and predicting engagement patterns.

Keywords: Virtual Reality, Cognitive Training, Predictive Modeling, Machine Learning, User Segmentation, Clustering, K-Prototypes, Profiling.

1 Introduction

In an era where the global population is aging rapidly, cognitive decline has emerged as a pressing health issue. Cognitive abilities, including memory, attention, information processing, and problem-solving, are essential for maintaining the autonomy and well-being of individuals (Brugada-Ramentol et al., 2022). To address the growing need for effective cognitive training, companies like Virtuleap have dived into developing applications like Enhance VR, seeking to make an impact on cognitive health. Virtuleap is a company that specializes in harnessing the power of Immersive Virtual Reality (IVR) for cognitive training and monitoring. Enhance VR, their flagship application, comprises a collection of games designed to engage users in cognitively demanding tasks within a multisensory virtual environment. What makes Enhance VR distinctive is its ability to incorporate proprioceptive and visuomotor information, creating an embodied experience for users. This stands in stark contrast to conventional screen-based applications that often suffer from poor transferability to activities of daily living due to their more abstract and detached nature.

Enhance VR is an application that consists of 15 games that are designed to train and monitor a range of cognitive domains, including memory, attention, task flexibility, information processing, orientation, problem-solving, and motor control. The company has taken measures to ensure that the games are rooted in scientific principles while maintaining a gamified and engaging experience for the users. The app is designed to be used regularly in the form of short workouts, comprising three randomly chosen games, which conform the Workout of the Day (WOD), or for the player to choose their own game. The adaptive difficulty of the app adjusts the challenge level in tandem with the user's progression. This ensures that users are continually stimulated, fostering sustained cognitive growth.

The application collects data on demographic variables, given by the user when creating their account. Additionally, the users have to report their mood and sleep hours once every day that they initialize the application. When playing, the information on each of their sessions is tracked, enabling an in-depth analysis of various usage related variables, such as how many

sessions they have been playing and for how long, or even which games they play more often. From all of this information one is able to obtain a description of the user and their behaviour within the game.

Another key aspect of Enhance VR is performance tracking. The application features the Enhance VR Performance Index (EPI), which aggregates weighted performance across all cognitive categories. This allows users to monitor their progress over time along with the score they obtain after playing each of the games, which is improved periodically after they play one of the games and break their previous record.

Considering Virtuleap's objective of enhancing the cognitive assessment and training industry through the utilization of cutting-edge technologies like virtual reality and artificial intelligence, the retention and engagement of users hold great importance. Evaluating the impacts of these technologies on user progression within the application over the course of sessions, ought to be accompanied by a strong emphasis on user retention. The concept of "retention" has emerged as a crucial metric that game developers closely monitor to ensure the long-term success of their products. While the term "churn" traditionally refers to the loss of customers or players, the emphasis on retention in the gaming context reflects the industry's dedication to fostering strong and lasting relationships with its user base.

Retaining users is of great importance for game developers, as it directly impacts revenue generation and player engagement. According to industry statistics (SuperData, 2020), a significant portion of revenue in the game industry in the realm of free-to-play mobile games, is derived from users who continue to actively engage with the game beyond the initial download. Conversely, users who churn early on often fail to cover the costs associated with acquiring and onboarding them. By focusing on retention rather than churn, game developers can gain valuable insights into user behavior and preferences, allowing them to design targeted strategies to retain and engage players.The shift towards retention-focused strategies represents a paradigm change in the gaming industry, where the primary objective is no longer solely acquiring new users but rather cultivating a loyal and engaged player base. Armed with this information, they can then implement tailored interventions, such as personalized offers, rewards, or gameplay enhancements, to incentivize players to remain engaged with the game.

In this context, the aim of this project is to analyze the behavior and profile of Enhance VR users and identify the variables that influence their retention. This objective will be accomplished by leveraging the collected app data and developing predictive models capable of forecasting user retention probabilities. Furthermore, clustering techniques will be employed to segment users according to their characteristics and behavioral patterns, facilitating a deeper understanding of the drivers behind user retention.

This study follows a structured approach to investigate the topic of user retention in the context of Enhance VR. In Section 2, the Literature Review explores retention rate calculation, predictive models, and clustering techniques, drawing insights from a range of authoritative sources. Section 3 provides an overview of the dataset used, including data transformations, aggregated insights, and retention rate measures. The adopted approach for predictive modeling and clustering analysis is outlined in Section 4, Methodology. Section 5 presents the results, offering a detailed analysis of the findings obtained from the methodology employed. Finally, our work concludes with a summary of the key findings and possible expansions of our research.

2 Literature Review

As the methodology presented in this analysis consists of two main components, namely retention prediction and user profiling through clustering, we will structure the related work section into two distinct parts to address each aspect separately. The first part will focus on retention prediction, examining prior research and techniques employed. The second part will delve into user profiling and behavior analysis, exploring studies that utilize clustering techniques to gain insights into user characteristics and preferences.

2.1 Retention prediction

Churn prediction is a well-explored area across multiple industries, including credit scoring, gaming, telecommunications, subscription-based services, e-commerce, and online platforms. In the specific context of the gaming industry, predicting and addressing user retention is of importance in this study. As the first stage of our study, we had to establish the definition of retention within the context of our research and determined the approach for its measurement. Retention can be defined in various ways, either as a longitudinal measure or a specific moment in time prediction. For our analysis, we specifically focused on two distinct levels, the user level and the session level, basing our decision in other studies performed in the area of study. On Milošević et al. (2017) the authors propose a two-stage approach that leverages machine learning models trained on behavioral data collected during the first day of user activity. By focusing on early churn prediction, the system enables the implementation of personalized push notifications to retain users and prevent them from leaving the game. This paper's methodology aligns with our own research objective of analyzing user activity at an early stage in our user level analysis to gain insights into churn behavior and develop effective preventive strategies. In Tekin et al. (2023) they tackle the problem by combining fuzzy clustering and ensemble learning to predict user retention in the gaming industry, exclusively utilizing the first 24 hours of data and aggregating it at the user level.

Other approaches such as survival analysis and heuristic modeling are also employed in understanding user retention. The paper by Drachen et al. (2016) investigates the problem of rapid retention prediction in the context of mobile Free-to-Play games. The authors explore heuristic modeling approaches to build simple rules for predicting short-term retention based on player activity in the first session, day, and week. The study defines retention as any game activity during the second week after installation and examines different prediction periods and classification strategies. In Periáñez et al. (2016) the authors develop a survival ensemble model specifically for retention prediction in mobile social games. Their model predicts the probability of retention as a function of time, enabling the identification of different loyalty profiles among players. Similarly, our analysis also considers different time periods to examine user retention in the context of the Enhance VR game.

As a next step in our study, we focused on extracting comprehensive information from the data by creating new features to enhance the performance of our models. In Hadiji et al. (2014) the authors model and predict player retention in Free-to-Play games. They introduce a range of generic features, such as playtime, session length, session intervals, and virtual economy-related factors. In Lee et al. (2016) the researchers extracted relevant features from player logs that captured the characteristics of retention, such as the number of played days, number of

purchases, number of log-ins per day and daily level distribution. We also utilized the paper by Tekin et al. (2023) to incorporate additional features into our models. The features created by the authors include session count, session length, maximum level number, maximum session length and average session length. They highlighted the creation of relevant features capturing installation information and gameplay patterns, which are essential in modeling user behavior accurately.

Additionally, we explored the concept of feature importance to gain insights into the relative significance of each feature in predicting user retention, allowing us to prioritize the relevant features. In a similar way Drachen et al. (2016) evaluates the importance of various features, including installation information and gameplay patterns, to understand their relevance in predicting retention. In Lee et al. (2016), the feature importance analysis using the random forest model revealed that the Number of Purchases feature was the most important among the behavioral features.

In order to determine the models to be employed in our analysis, we reviewed relevant studies in the field that conducted similar investigations. Specifically, we examined the paper by Tekin et al. (2023) in which the authors adress the limitations of traditional algorithms on high-dimensional and imbalanced datasets. The authors demonstrate the effectiveness of boosting-type algorithms like XGBoost, Catboost, and LightGBM and highlight the importance of considering different clusters of users for personalized predictions. Particularly XGBoost, was found to be the most accurate for predicting retention. In (Lee et al., 2016) the researchers adopted several classification algorithms, namely decision tree, random forest, and support vector machine. To address the challenge of imbalanced datasets, they applied the Synthetic Minority Over-sampling Technique (SMOTE). In Drachen et al. (2016) the authors utilize three popular machine learning classifiers: Logistic Regression, Support Vector Machines, and Random Forest, to develop the short-term prediction model. Fine-tuning of the models is carried out and the hyperparameters for the SVM and RF models are optimized using a grid search method with cross-validation error as the evaluation metric. Milošević et al. (2017) explore the use of different models including Logistic Regression, Naive Bayes, Decision Tree, Gradient Boosting, and Random Forest. The results indicate that the Gradient Boosting model achieved the highest performance, according to the AUC, Precision, Recall, and F1 Score, the same metrics were used in our analysis to compare different models for their predictions. In Bekkar et al. (2013) the authors specifically addresses the assessment of models on imbalanced datasets, such as ours. The paper emphasizes the limitations of using accuracy as an evaluation measure in such scenarios. It highlights the need for alternative evaluation measures that provide a more comprehensive and reliable assessment.

To calibrate the probabilities in our research, we based our study in two papers: Niculescu-Mizil and Caruana (2005) and Martino et al. (2019). The first paper focuses on the relationship between the predictions made by different learning algorithms and the true posterior probabilities. It identifies that certain algorithms, such as boosted trees and boosted stumps, tend to push probability mass away from 0 and 1, resulting in a sigmoid-shaped distortion in the predicted probabilities. On the other hand, models like neural networks and bagged trees do not exhibit such biases and produce well-calibrated probabilities. The paper explores two calibration methods to correct biased probabilities: Platt Scaling and Isotonic Regression. The second paper focuses on the importance of calibrating a classification system to ensure proper probability estimates, reviewing three calibration techniques: Platt's Scaling, Isotonic Regression, and SplineCalib.

2.2 Clustering methods

Proceeding to the second part, we will focus on our aim to analyze user profiling. Here, we navigate through various studies that harness the power of clustering techniques to review the tapestry of user attributes and inclinations.

In the analysis, we experimented with different models for clustering. Initially, we employed K-means clustering as our baseline model. This method was first introduced in Hartigan and Wong (1979), and has since become one of the most widely used clustering techniques in various fields. The authors lay the foundation for the algorithm that partitions n observations into k clusters, in which each observation belongs to the cluster with the nearest mean. Transforming categorical features was crucial as K-means relies on calculating the mean which is not defined for categorical data. However, doing this can sometimes lead to loss of information and high-dimensional sparse data. This is supported by Huang (1998) in which the authors address this critical limitation of the original k-means algorithm. Two novel extensions to the k-means algorithm are presented, namely k-modes, which is adept at clustering categorical data, and k-prototypes, a hybrid of the previous ones.

K-Prototypes is designed to handle data with mixed attributes, the algorithm combines the K-means method for numerical attributes and the K-modes method for categorical attributes. By using K-Prototypes, we were able to cluster the data without the need for encoding the categorical variables, thus retaining more information and potentially gaining more insights from the data. In the paper titled Ranti et al. (2019) the authors delve into the practical application of the k-prototypes algorithm in the domain of video game analytics, maintaining that the algorithm is consistent with the nature of gaming data. This paper highlights the importance of gathering player behavioral data through user telemetry and how it can be leveraged by game developers.

In the paper by Bauckhage et al. (2014), the authors tackle the rapidly growing volume of behavioral data within the gaming industry. This explosion of data requires the development of techniques to extract meaningful insights, particularly since these datasets can be vast, time-dependent, and high-dimensional. The authors emphasizes the utility of clustering techniques to explore and make sense of behavioral data, thereby reducing its complexity. Specifically, the paper highlights clustering as a potent tool for discerning player profiles and analyzing play styles, areas that have been gaining traction in the rapidly evolving field of game analytics, which is our next focus in this analysis. Within this context, profiling is understood as the process of describing and characterizing the behavior patterns of players in each cluster.

Dutra (2022) focuses on customer profiling as a tool for lifestyle visualization. It combines creative imagination techniques with storytelling and statistical methods to create easily interpretable representations of the consumer. A key aspect of the paper's methodology is the use of exploratory factor analysis. It's a technique used to reduce the number of variables and identify latent constructs within the data. The second part of the results section of the paper focuses on cluster analysis and customer profiling. Here, the authors group the data into clusters and create profiles for each cluster.

Another way of enhancing the comprehensibility of clustering results for interpretation can be achieved through visualization techniques. In our endeavor to create lucid visual representations of the clustering output, we evaluated three distinct methods: PCA (Principal Component

Analysis), t-SNE (t-Distributed Stochastic Neighbor Embedding), and UMAP (Uniform Manifold Approximation and Projection).

Jolliffe and Cadima (2016) discuss how PCA is used for dimensionality reduction in large datasets. The technique operates by identifying new uncorrelated variables that maximize the variance in the data. These new variables, known as principal components, are not predefined and are created based on the dataset. In Van der Maaten and Hinton (2008) the t-SNE technique is introduced. The authors present it as a variation of Stochastic Neighbor Embedding. They explain how t-SNE significantly improves visualizations by reducing the crowding of points in the center of the map, a common issue in high-dimensional data visualization. The technique excels in revealing structures at various scales, which is particularly crucial for high-dimensional data lying on several different but related low-dimensional manifolds. Wang et al. (2020) delves into the inner workings of several dimension reduction techniques, including t-SNE and UMAP. The authors focus on understanding the trade-offs between the preservation of global structure and local structure. Unlike PCA, t-SNE and UMAP are both non-linear techniques, on one side UMAP is known for being faster and more scalable compared to t-SNE, but t-SNE tends to create more distinct clusters, which can be more interpretable and valuable for certain applications, particularly in exploratory data analysis.

3 Data

3.1 Data description

The data provided by the company consists of three tables: users, moods, and sessions. The users table comprises 56,670 observations, the sessions table includes 326,232 observations, and the moods table contains 98,869 observations. These tables collectively span the time period from August 25, 2020, to June 8, 2023.

The **users** table contains information about registered users on the app, including a unique identifier for each of them. It also includes demographic data such as the user's date of birth, self-reported gender, self-reported profession, country of registration, selected language, highest education level achieved, and the date and time of user registration. The specific categories for each categorical variable can be found in Table 1.

The **moods** table contains information about the moods and slept hours of users. These are registered once a day when the user enters the app. Each entry in this table is associated with a unique mood identifier and the user identifier. The self-reported mood of the user is recorded on a scale of 1 to 5, ranging from "very sad" to "very happy." The self-reported sleep hours are recorded on a scale from 5 to 9. The date and time of mood and slept hours registration are also recorded.

The **sessions** table contains information about the game sessions played by users. Each session is identified by a unique session identifier and is associated with the user identifier. The table includes data such as the game played in the session, the workout of the day identifier, indicating whether if the game was selected by the user from the library or it was a random set of games provided by the app, the session status, the start time, and end time of the session. Additionally, it records whether the session involved a tutorial and whether it was a benchmark session. The

Gender	Profession	Language	Education
1: Male	1: Art & Design	1: English	1: No formal education
2: Female	2: Education	2: Chinese	2: Primary education
3: Other	3: Engineering	3: Spanish	3: Secondary education
4: Prefer not to say	4: Finance	4: Portuguese	4: Bachelor's degree
	5: Healthcare	5: Japanese	5: Master's degree
	6: Law		6: Doctorate or higher
	7: Management		
	8: Media		
	9: Military		
	10: Research		
	11: Technology		
	12: Other		

Table 1: Users table variables' categories

level reached by the user during the session and the score obtained are also included. The specific categories for each categorical variable can be found in Table 2.

Game	Session status
1: React (Flexibility)	1: Playing
2: Hide and Seek (Orientation)	2: Finished
3: Memory Wall (Memory)	3: Abandoned
4: Balance (Motor Control)	4: Playing Tutorial
5: Pizza Builder (Attention)	5: Abandoned Tutorial
6: Magic Deck (Memory)	
7: Odd Egg (Problem-solving)	
8: Assembly (Processing)	
9: Whack-a-mole (Attention)	
10: Harmonize (Processing)	
11: Maestro (Memory)	
12: Slinger (Motor Control)	
13: Stacker (Problem-solving)	
14: Orbital (Orientation)	
15: Shuffled (Attention)	

Table 2: Sessions table variables' categories

3.2 Data preprocessing

In order to create a unified dataset adaptable for each specific task, we conducted an initial phase of data preprocessing. First, we performed data cleaning on the users table based on the guidelines provided by the company. The objective was to remove unreliable user entries based on their demographic characteristics. The following criteria were used to identify and remove such users: individuals who reported being too young for their reported education level or profession. For example, users younger than 12 who reported being in high school/secondary,

users younger than 17 with a reported bachelor's degree, users younger than 20 with a reported master's degree, users younger than 25 with a reported doctorate or higher. This resulted in the removal of less than 3% of the users, which were also eliminated from the other two tables.

Next, we addressed missing values in the dataset. Initially, we decided to remove users who had missing data on variables such as date of birth, gender, profession, and education level. These missing values accounted for less than 2% of the total users. Additionally, the country variable, which had a 9% of missing values, was removed from the dataset. The provided values for this variable were numeric codes that could not be reliably interpreted or matched with specific country identifiers.

After completing the data cleaning process, we merged the sessions data with the users and moods data. This merging was performed based on the user identifier and session date, with the aim of creating a unified dataset that could then be tailored to specific tasks. Subsequently, we removed sessions that lacked mood data due to a mismatch between the session date and the date of mood registration. These sessions with missing mood data represented approximately 3% of the total sessions.

3.3 Data exploration

3.3.1 Variable visualization

After preparing the dataset, we conducted an exploration of session characteristics to gain insights into user behavior and inform our predictive and clustering analysis. As part of this exploration, we examined the session status, as depicted in Figure 1, and found that 70% of the sessions were finished, while the remaining 30% were abandoned.

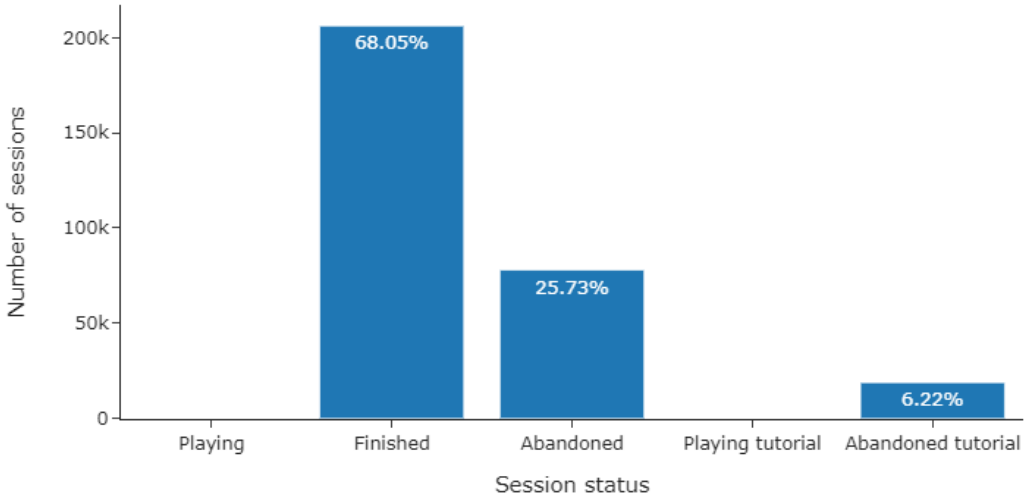

Figure 1: Number of sessions by status

Regarding the finished sessions, we further investigated the distribution of the number of sessions played by users throughout the data period, illustrated in Figure 2. The mean number of sessions played by user was approximately 7, with a median of 3. This median indicates that

half of the users participated in 3 sessions or fewer. The upper quartile value of 6 implies that 75% of the users engaged in 6 sessions or fewer. Notably, for better visualization, the figure does not include the values after the upper fence, that ranges from 14 to 1434 sessions, which highlights an unusually high number of sessions undertaken by at least one user.

Additionally, we analyzed the number of days in which users played their sessions, as depicted in Figure 3. It was observed that 70% of the players completed their sessions within a single day, while 15% did it in two days. The remaining users spread their sessions across three or more days. This finding indicates a significant proportion of users discontinuing their engagement after one or two days playing.

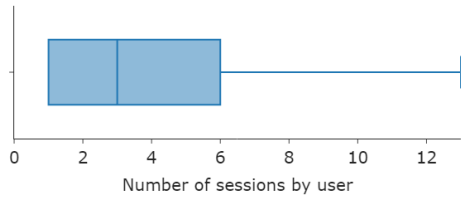

Figure 2: Distribution of number of sessions played by user

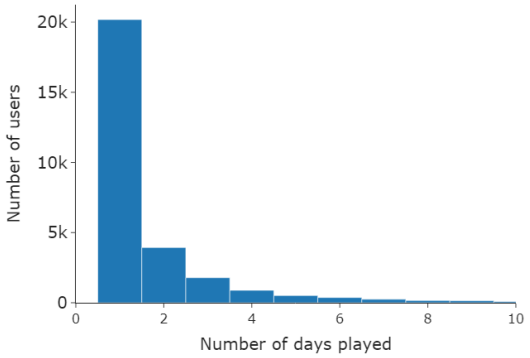

Figure 3: Number of days played by users

At the session level, we further examined the distribution of finished sessions based on the users' days of gameplay, shown in Figure 4. It was observed that 37% of the finished sessions occurred on the users' first day of gameplay, while 15% took place on their second day. The remaining sessions occurred from the third day of gameplay onwards, progressively decreasing as the number of days increased. This finding highlights that the majority of sessions are concentrated within the initial 10 days of gameplay.

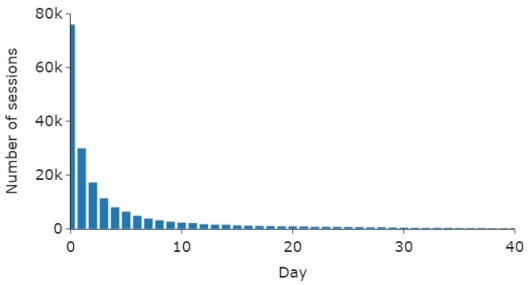

Figure 4: Number of sessions played by users' gameplay day

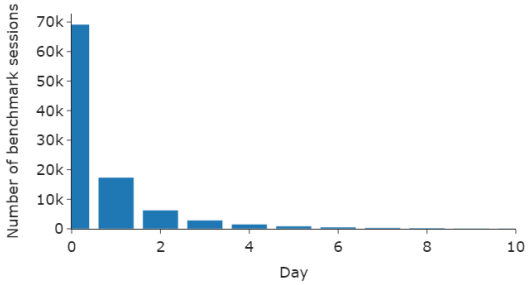

Figure 5: Number of benchmark sessions played by users' gameplay day

Furthermore, we analyzed the distribution of the benchmark sessions. A benchmark session is a first longer session to establish users' baseline performance in each game. We found that a 70% occurred on the users' first day of gameplay, while 17% took place on their second day (see Figure 5). The remaining sessions occurred from the third day of gameplay onwards, showing

a significant decrease in frequency as the number of days increased. As expected, the majority of benchmark sessions are concentrated within the initial 2 days of gameplay.

3.3.2 Retention rate calculation

In this study, we computed the retention rates using the data provided by the company as it stands up to the June 8,2023. Through a general comprehension of the prevailing retention rates, we not only elucidate the current retention landscape within the company but also acts as a reference for evaluating and forecasting future retention trends. We can see in the literature authors referring to these metrics to measure user engagement and ascertain the efficacy of interventions. In Su et al. (2021) the authors portray retention as central metric in game publishing analytics, which refers to the use of data analysis tools and techniques in the context of video game publishing. Viljanen et al. (2016) elucidate Retention, Rolling Retention, and Lifetime Retention as prominent population-level retention metrics, which are inherently implicated by their model.

Retention reflects the ability of users to continuously engage with the product and find value in it (GoPractice, 2023). Calculating retention involves determining the number of people who return to the product on a specific day or week after their initial usage, and it is expressed as a percentage. To calculate retention, we utilized the following formula, similar to the approach taken by Debeauvais et al. (2014), where they measured retention based on the active users at a specific moment in N time:

$$\text{Day } N \text{ Retention} = \frac{\text{Active Users}}{\text{Total Users}} = \left(\frac{\text{Users on Day } N}{\text{Users on Day 0}}\right) \times 100$$

This formula represents what in the industry is known as Day N retention. Common day N retention rates that are often considered in the industry include Day 1, Day 7 and Day 30 retention. These metrics provide insights into the short-, medium- and long-term retention of players after their initial engagement with the game. Day 1 retention measures the percentage of players who continue to play the game on the first day after installation, Day 7 and Day 15 represent the medium-term horizon, and long-term retention is essentially retention after Day 30. In our study, we have chosen to calculate retention by strict calendar dates. This means that regardless of the specific time a user performed their initial action, their subsequent retention days are determined by the calendar date Amplitude (2020). It allows us to compare and analyze retention rates across different time periods and provides a clear understanding of user behavior on specific days.

Figure 6 and Table 3 collectively offer a portrayal of our observations. The graph depicted in Figure 6 illustrates the retention trend over a span of 40 days. Meanwhile, Table 3 enumerates key metrics that serve as a benchmark for comparison against prevalent market standards. The 'Day N' column denotes specific days after a player's initial interaction with the game, while the 'Retention' column indicates the percentage of players who remained engaged with the game as of that day. On the same day of installation we observe that 91.97% of the users initialized a session. This rate decreases to 14.95% after one day, to 6.75% after 7 days and 5.23% after 15 days. Finally it stabilizes to an approximate rate of 5.01% by Day 30.

Day N	Retention (%)
D0	91.97
D1	14.95
D7	6.75
D15	5.23
D30	5.01

Table 3: App Day N Retention

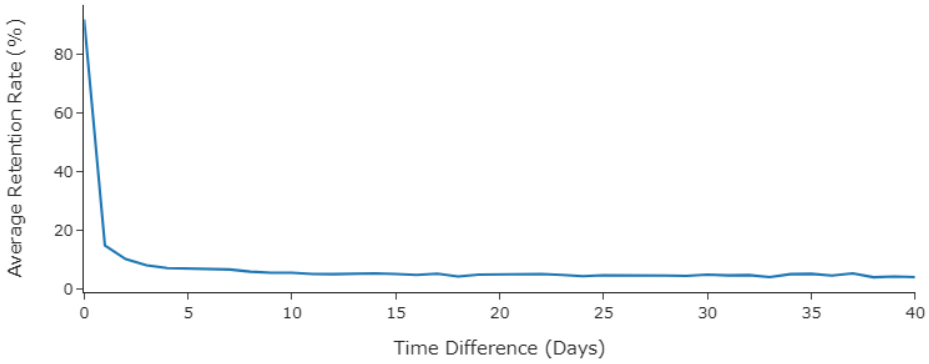

Figure 6: Day N retention across time

We expanded our analysis by calculating retention rates for each distinct game. In a parallel manner, we evaluated the percentage of players who maintained their engagement with a specific game through to the Nth day. This methodology illuminates the retention trends and allegiance of players that are particular to each game. However, it is imperative to recognize that these figures are also influenced by the volume of users actively participating in a specific game since retention is computed as a ratio relative to the total user base of the app. Consequently, the values represent the proportion of overall retention attributable to players engaged in that particular game. For example we can see here that the most popular games are the first and the third, with higher retention rates in day 1 and 7. But in general all games follow the same tendency, similar to the aggregate one previously mentioned.

Day N	Game 1	Game 2	Game 3	Game 4	Game 5	Game 6	Game 7	Game 8
D0	32.81	21.8	29.18	19.58	21.24	19.66	20.76	23.6
D1	8.7	7.01	8.45	7.16	7.83	7.43	7.12	6.89
D7	5.19	4.91	5.51	4.95	5.2	4.99	5.05	4.88
D15	4.9	5.03	4.04	5.04	4.74	4.13	4.67	3.94
D30	6.36	4.7	5.29	3.38	3.77	3.49	4.19	3.44

Table 4: Day N Retention for Games 1-8

Day N	Game 9	Game 10	Game 11	Game 12	Game 13	Game 14	Game 15
D0	27.82	11.22	12.92	14.71	15.97	14.12	18.17
D1	7.93	5.67	5.89	6.17	6.55	7.01	7.44
D7	4.66	4.59	5.07	4.82	4.98	5.01	6.22
D15	4.65	5.31	4.61	3.92	4.45	5.44	6.57
D30	4.32	3.8	4.14	3.8	5.89	5.44	4.52

Table 5: Day N Retention for Games 9-15

4 Methodology

In the subsequent sections, we present the methodology employed for both retention prediction and clustering tasks. This includes a description of the feature selection process, the models utilized for prediction, the evaluation metrics employed to assess performance, as well as the features and techniques used for clustering.

4.1 Retention prediction

4.1.1 Retention definition

We begin by formalizing the retention prediction task as a binary classification problem, which is an approach widely using in the study of the gaming industry. It is employed by Milošević et al. (2017) to predict churn, and Drachen et al. (2016) to predict retention.

In this study, the retention definition is established at two levels. Firstly, at the user level, each individual user is classified as either retained (1), indicating a likelihood to continue playing, or churned (0). For this approach, we defined retention as the occurrence of any game activity after the benchmark session. Specifically, a player is labeled as retained only if they register at least one session after the benchmark. The benchmark session serves as an approximation of the first gameplay day since the majority of such sessions are observed within the initial day of gameplay. This concept is consistent with (Tekin et al., 2023) and (Milošević et al., 2017), that predict customer retention based on data from day one.

Secondly, at the session level, we analyze data from the entire period. Here, we define retention as the presence of a following session after the last played session. To incorporate a temporal aspect, we introduce a window of 10 consecutive days, following a similar approach as Periáñez et al. (2016). Accordingly, a user will be considered retained (1) if they engage in a gameplay session within the 10-day window, and churned (0) otherwise. This session-level approach allows us to capture retention patterns that extend beyond the immediate aftermath of the benchmark session.

User level

For the user-level approach, we utilized the user-level data for benchmark sessions, specifically focusing on finished sessions. This involved aggregating the benchmark session data of each

user, using the available variables provided by the company, as well as additional variables found in the literature (Tekin et al., 2023; Hadiji et al., 2014).

Consequently, we performed feature engineering and constructed the following variables by user: number of benchmark sessions, average score achieved across sessions, average playtime per session, average playtime rate per session (minutes played relative to the entire day), average proportion of sessions where the daily workout was selected, and the period between the user's registration date and the session's date. We also incorporated demographic variables such as age, profession, education level, and selected language.

As part of the data preparation process, we transformed categorical variables into dummy variables to effectively capture their categorical information. We also conducted a random train and test split of the data, with the train set containing 23,147 observations and the test set containing 5,787 observations. Additionally, we standardized the features before training our models.

The dataset exhibited an imbalanced class distribution, with approximately 27% of the observations corresponding to the positive class (retained) and the remaining observations to the negative class.

Session level

In the session-level approach, we utilized the data at the session level to capture the user's progression throughout their gameplay sessions. To achieve this, we created session-specific features for each user using the expanding mean technique, which calculates the average value of a variable considering all preceding sessions for that user. This allowed us to construct features such as the average score, average playtime per session, average slept hours, and average proportion of sessions where the daily workout was selected, providing insights into the user's performance and engagement over time.

Additionally, we incorporated other relevant features, including the period between the user's registration date and the session's date, demographic characteristics, and the most frequently reported mood across the cumulative sessions. Categorical variables were transformed into dummy variables as part of the data preparation process.

To evaluate our models, we split the dataset into a training set and a test set, ensuring that data from the same user was included in either the training set or the test set, but not in both. The training set consisted of 164,054 observations, while the test set comprised 41,982 observations. This division allowed us to assess the performance of our models on unseen data and validate their predictive capabilities.

Regarding the class distribution, this dataset also exhibits an imbalance, with approximately 83% of the observations corresponding to the positive class (retained) and the remaining observations to the negative class.

4.1.2 Modeling

For our binary classification problem, we employ a set of widely used models in the literature, including Logistic Regression (Drachen et al., 2016; Hadiji et al., 2014), Decision Trees (Miloše-

vić et al., 2017; Runge et al., 2014), Random Forest (Tekin et al., 2023), XGBoost Tekin et al. (2023), and Gradient Boosting (Milošević et al., 2017).

Logistic Regression is a probabilistic binary classification algorithm that utilizes a weighted linear combination of features to estimate the probability of instances belonging to the positive or negative class. It offers interpretability and simplicity.

The **Decision Tree** algorithm constructs a rule-based tree using training data and traverses the tree to predict the class label of instances. It relies on the properties of the tree and the features of the current instance to make predictions, making it an intuitive and straightforward model for classification tasks.

Random Forest, on the other hand, is an ensemble of Classification and Regression Trees (CART) that work collectively to form a robust model. It comprises multiple decision trees trained on bootstrapped datasets generated by random resampling. The prediction in Random Forest is based on the majority vote of individual trees.

Gradient Boosting is a machine learning algorithm that is based on an ensemble of decision trees. It utilizes the gradient boosting method to iteratively train a sequence of weak models, where each subsequent model is trained to correct the mistakes made by the previous ones. This iterative process enables the model to learn complex patterns.

XGBoost improves upon traditional gradient boosting by incorporating regularization principles and employing a set of sophisticated techniques. It optimizes the size of each decision tree in the ensemble to prevent overfitting and enhance generalization.

A summary of the specific models used for each approach can be seen in Table 6.

Approach	Task	Model
User level / Session Level	Binary classification Churned - 0 Reteined - 1	Logistic Regression Decision Tree Random Forest XGBoost Gradient Boosting

Table 6: Predictive models

In order to address the class imbalance, we adjusted the class weights in the Decision Tree, Random Forest, and XGBoost models. This weighting mechanism assigns higher weights to the minority class and lower weights to the majority class, thus giving more importance to the predictions of the minority class during training. This approach helps mitigate the bias towards the majority class and enhances the models' ability to capture patterns and make more accurate predictions for both classes.

As part of the model optimization process, some hyperparameters were tuned to improve the performance of the models based on decision trees. Specifically, we focused on tuning the maximum depth of the decision tree and the minimum number of samples required to split a node, as well as the number of estimators.

Additionally, although we defined the task as a binary classification problem, we considered

that predicting probabilities can provide valuable information for decision-making. The goal of predicting whether a user will be retained or not is to take actions to prevent users from churning. Therefore, retention strategies should primarily focus on users with intermediate probabilities of continued engagement rather than cases where there is a high probability of retention or churn. To address this, we decided to employ two methods, namely Platt Scaling and Isotonic Regression, to calibrate the predicted probabilities generated by the classifier. By applying these calibration techniques, we intend to refine the predicted probabilities to align them better with the actual likelihood of user retention, enabling more informed decision-making and tailored retention strategies.

4.1.3 Evaluation metrics

When evaluating the performance of the predictive models, we employed a set of metrics that were well-suited for our specific task and aligned with the characteristics of our dataset, such as Precision, Recall, AUC, and F1 score. Below, we provide concise explanations of each metric.

Precision is defined as the proportion of correctly predicted positive observations to all predicted positive observations. It measures the accuracy of the positive predictions made by the model.

Recall, also known as sensitivity, is the ratio of correctly predicted positive observations to all observations in the actual positive class. It shows the effectiveness of the algorithm in identifying the positive instances correctly.

AUC represents the area under the ROC curve. The ROC curve is a graphical representation of the model's performance across different classification thresholds. The AUC is a metric that summarizes the overall performance of the model by considering the entire range of classification thresholds.

The **F1 score** is a harmonic mean of Precision and Recall. It considers both false positives and false negatives and provides a balanced measure of the model's performance, especially in imbalanced datasets.

For evaluating the calibration performance of the classifier, we employed the Brier score and Log-Loss score, as suggested by previous studies (Martino et al., 2019). These metrics assess the calibration quality of the predicted probabilities. The **Brier score** measures the average squared difference between the predicted probabilities and the actual outcomes, with lower scores indicating better calibration. On the other hand, **Log-Loss score** measures the performance of the predicted probabilities by comparing them with the true labels, with lower scores indicating better calibration.

4.2 Clustering methods

4.2.1 Feature engineering

To apply clustering techniques and identify the main user profiles, the data preparation and feature engineering process focused on structuring the data at user-level. The first step involved

filtering the data to include only finished sessions.

Various features were then engineered to capture different aspects of user behavior, as done for retention prediction. These features included the number of sessions, average score, average playtime per session, average slept hours, and the average proportion of sessions where the daily workout was selected. Additionally, the most frequent mood and frequent game played across all sessions by each user were considered as features. The demographic variables were also included in the analysis.

To ensure the clustering algorithm produced meaningful results, outlier observations that could potentially disrupt the cluster definitions were addressed. The interquartile range (IQR) method was applied to identify and handle outliers. For instance, cases with average session duration atypical values, which could be indicative of incorrect data entry or system malfunction, were considered outliers and treated accordingly.

Lastly, numerical variables were standardized to normalize their scale and prevent biases that could be introduced by using variables with different scales, ensuring a fair representation of each feature in the clustering process.

4.2.2 Modelling

In this section, we outline the modeling techniques employed for clustering analysis. Specifically, we discuss the K-Prototypes algorithm used for clustering following Ranti et al. (2019), t-Distributed Stochastic Neighbor Embedding (t-SNE) for visualization of the clusters (Van der Maaten and Hinton, 2008), and finally, the profiling of clusters to gain insights into the behavior of different user segments.

K-Prototypes clustering

One of the essential aspects of clustering analysis was the ability to handle different types of data. In our dataset, we have both numerical and categorical variables, and therefore, a clustering algorithm that can handle this mixed data type is necessary. The K-Prototypes algorithm is an extension of the K-Means algorithm, which is particularly suited for clustering datasets that have mixed numeric and categorical features. In contrast to K-Means, which performed inadequately due to the loss of insights when encoding categorical variables, K-Prototypes retains the essence of categorical variables without the need for encoding.

The K-Prototypes algorithm combines the K-Means and K-Modes algorithms to form clusters by minimizing a combined distance measure. The objective function is defined as:

$$V = \sum_{i=1}^{k} \left[\sum_{x \in S_i} (x - \mu_i)^2 + \gamma \sum_{x \in S_i} d(x, m_i) \right]$$

In the formula, k denotes the number of clusters. S_i symbolizes the i^{th} cluster, while μ_i represents the mean of the numeric features in that same cluster. On the other hand, m_i indicates the mode of the categorical features within the i^{th} cluster. γ, acts as a weighting factor that harmonizes the relative influence of numeric and categorical features in the distance

measure. Lastly, $d(x, m_i)$ signifies a measure of dissimilarity between the data point x and the mode m_i of the categorical features.

The K-Prototypes algorithm operates through an iterative process which commences by randomly selecting k initial centroids and modes. Subsequently, each data point is allocated to the nearest centroid according to the distance metric. Thereafter, the centroids and modes for each cluster undergo an update based on the means of the numeric features and modes of the categorical features, respectively. This cycle of assigning data points and updating centroids and modes is repeated until either there are no changes in the centroids and modes, or a predetermined maximum number of iterations is reached. An important decision in the K-Prototypes clustering is selecting the appropriate number of clusters, k. The Elbow Method was employed to ascertain the optimal number of clusters. This method involves plotting the cost (value of the objective function V) against the number of clusters, and identifying the point at which adding more clusters does not significantly improve the fit.

Visualization using t-SNE

After obtaining clusters using the K-Prototypes algorithm, our next step was to visualize these clusters to understand the structure and separation among them. We chose to use t-Distributed Stochastic Neighbor Embedding (t-SNE), which works by constructing a probability distribution over pairs of high-dimensional objects in such a way that similar objects have a higher probability of being picked, while dissimilar points have an exponentially smaller probability.

The t-SNE algorithm then aims to minimize the divergence between a distribution that measures pairwise similarities of the input objects and a distribution that measures pairwise similarities of the corresponding low-dimensional points in the embedding. It does this by minimizing the Kullback-Leibler divergence of the two distributions with respect to the locations of the points in the map.

Profiling of clusters

Once clusters are formed and visualized, the final step involves profiling these clusters to understand the characteristics and behaviors of the users within them. Profiling involves analyzing the central tendencies of the features within each cluster. Through profiling, we can identify common patterns and characteristics, which can be useful for tailoring strategies and interventions to different segments of users. This information can be instrumental in guiding decision-making processes, such as crafting personalized marketing campaigns, optimizing user recommendations, or developing targeted user engagement strategies, such as push notifications.

We analyzed each cluster separately. For numerical variables, averages were computed within each cluster. For categorical variables, the percentage of each category within each cluster was calculated and reported. In addition to the analysis of central tendencies, the segmentation was enriched through the creation of graphical representations, which depicted the distributions of numerical variables within each cluster. To complement this, an encompassing plot was constructed to showcase the distribution of each numerical variable across all clusters, wherein the percentage of values from that variable within each cluster was calculated.

5 Results

In this section, we present the results of the predictive models, clustering analysis, and profiling conducted in our study on user retention in Enhance VR Data.

5.1 Retention prediction

5.1.1 User level

Table 7 presents the metrics of the predictive models for user retention. First, the Random Forest model has the highest AUC score of 0.85 among all the models evaluated. However, when selecting the best model, it is important to consider precision and recall in addition to the AUC score.

In the context of predicting user retention and preventing churn, the primary goal is to identify users at high risk of churning and take appropriate actions to prevent it. By minimizing false positives, which are cases where users who are likely to churn are wrongly predicted as retained, the risk of not addressing the needs of these users is effectively reduced. Therefore, precision becomes an important metric to prioritize.

Model	AUC	Class	Precision	Recall	F1 score
Logistic Regression	0.79	1	0.82	0.63	0.71
		0	0.87	0.95	0.91
Decision Tree	0.83	1	0.63	0.86	0.73
		0	0.94	0.80	0.87
Random Forest	0.85	1	0.73	0.81	0.77
		0	0.92	0.89	0.91
XGBoost	0.82	1	0.83	0.70	0.76
		0	0.89	0.95	0.92
Gradient Boosting	0.82	1	0.83	0.70	0.76
		0	0.89	0.95	0.92

Table 7: Predictive models performance at user level

Both the XGBoost and Gradient Boosting models demonstrate the highest precision values of 0.83. This indicates their greater effectiveness in minimizing false positives. These models also show similar values across other metrics. However, when considering precision as a key factor, the Gradient Boosting model slightly outperforms XGBoost. Hence, the Gradient Boosting model may be the preferred choice for predicting user retention in this particular analysis.

Figures 7 and 8 depict the Precision-Recall and ROC curves for the Gradient Boosting model, respectively. The behaviour of these curves is aligned with the previously analyzed metrics and suggest that the model is good.

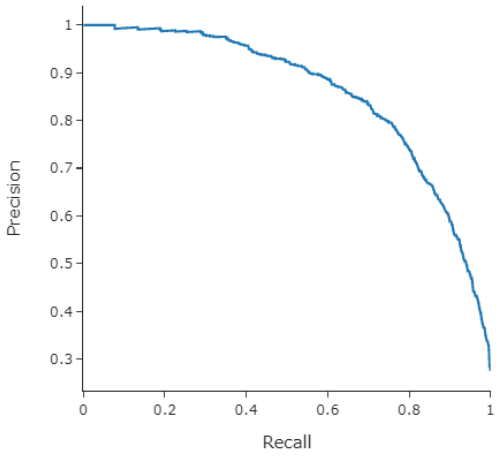

Figure 7: Gradient Boosting Precision-Recall curve

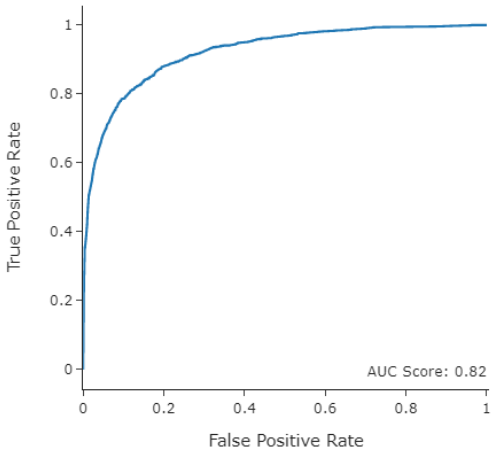

Figure 8: Gradient Boosting AUC-ROC curve

Finally, the confusion matrix of the model depicted in Figure 9 indicates that 95% of the instances belonging to churned class were correctly classified as churned class (true negatives), while 5% were mistakenly classified as retained class (false positives). Moreover, 30% of the instances belonging to retained class were misclassified as churned class (false negatives), while 70% of the instances were correctly identified as retained class (true positives).

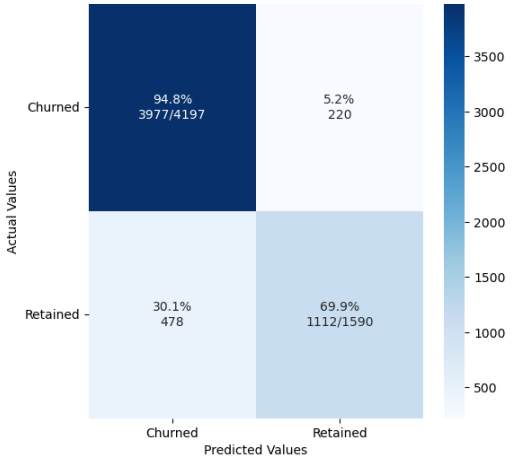

Figure 9: Gradient Boosting Confusion Matrix

Regarding the features of the Gradient Boosting model, Figure 10 indicates that the number of sessions and the days after user creation, that represents the days between the user registration and session, are the two most important variables for predicting user retention. This implies that these two features strongly influence the model's ability to predict whether a user will continue using the app or not.

The high importance assigned to the number of sessions suggests that the quantity of user sessions has a significant impact on their likelihood of staying engaged with the app. This feature reflects the level of user interaction and can be seen as a measure of user commitment

and interest. The importance of days after using creation highlights the temporal aspect of user behavior. On the other hand, the duration between user registration and the session represents the user's initial experience with the app.

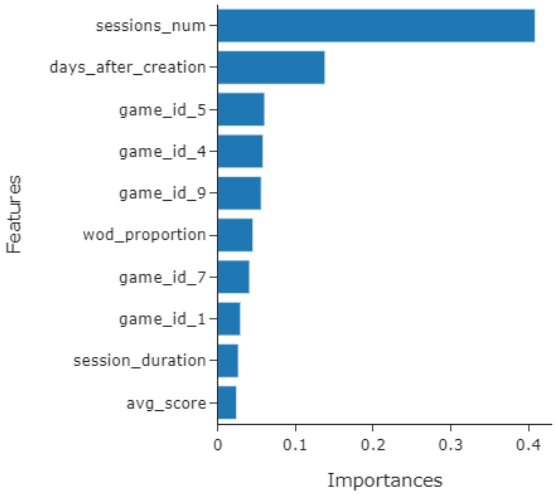

Figure 10: Feature Importances Gradient Boosting

Figure 11 depicts the calibration curve for the Gradient Boosting model, along with the curves for the Gradient Boosting calibrated using Platt's (Sigmoid) and Isotonic methods. These curves illustrate the relationship between the average predicted probability for each bin and the fraction of positive classes in each bin. It can be observed that the base model approximates perfect calibration, with some points lying below the diagonal, indicating over-forecasting, while others lie above the diagonal, indicating under-forecasting.

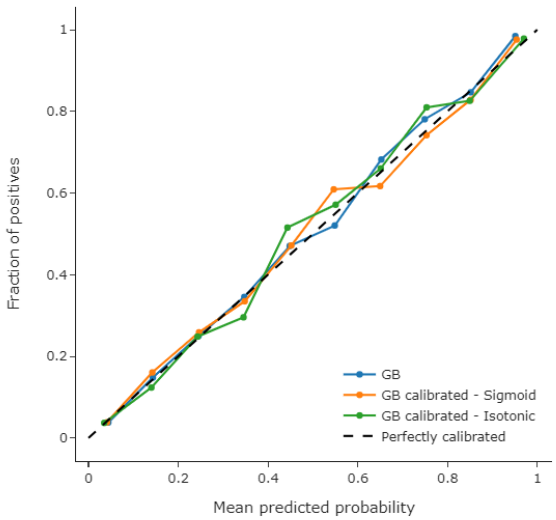

Figure 11: Gradient Boosting Classifier Calibration

Upon observing the histograms presented in Figure 12, which display the distribution of predicted probabilities for each model, we can note that there is only a slight difference among

them. This suggests that the calibration methods applied did not result in significant changes to the distribution of the predicted probabilities.

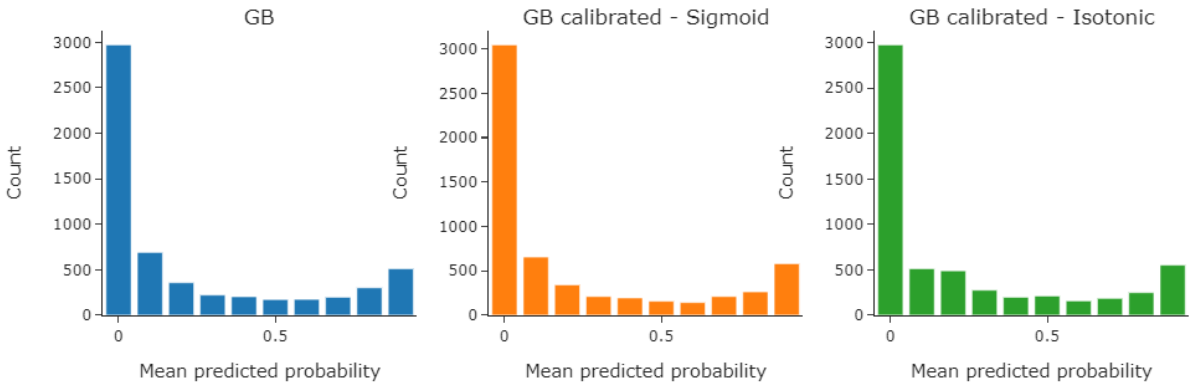

Figure 12: Gradient Boosting Classifier Predicted Probabilities

In terms of calibration performance, Table 8 indicates that the use of the Sigmoid method slightly improves both the Brier score and Log loss compared to the base Gradient Boosting model. On the other hand, the isotonic method does not offer any significant improvement. Although the model with the Sigmoid method is not perfectly calibrated, the predicted retention probabilities are better calibrated. This means that the predicted probabilities of user retention are more accurate, providing a more solid foundation for decision-making and designing actions to prevent user churn.

Classifier	Brier score	Log loss
Gradient Boosting	0.090	0.299
Gradient Boosting - Sigmoid	0.089	0.298
Gradient Boostin - Isotonic	0.090	0.299

Table 8: Calibration evaluation

5.1.2 Session level

Table 9 presents the metrics of the predictive models for retention at the session level. Overall, the models demonstrate relatively low AUC scores compared to the models at the user level, indicating poorer discrimination between positive and negative instances. This can be attributed to the minor representation of the negative class in the dataset, which affects the model's ability to distinguish between the two classes.

In addition to the lower AUC scores, there are some variation in precision, recall, and F1 scores among the models, highlighting differences in their abilities to correctly identify instances, particularly in the negative class.

Model	AUC	Class	Precision	Recall	F1 score
Logistic Regression	0.50	1	0.84	0.99	0.91
		0	0.75	0.00	0.01
Decision Tree	0.69	1	0.92	0.70	0.79
		0	0.30	0.68	0.42
Random Forest	0.69	1	0.92	0.70	0.79
		0	0.30	0.68	0.42
XGBoost	0.69	1	0.94	0.53	0.68
		0	0.25	0.84	0.39
Gradient Boosting	0.51	1	0.84	0.99	0.91
		0	0.62	0.01	0.02

Table 9: Predictive models performance at session level

Similar to the approach used to compare the models at the user level, precision is given particular attention when evaluating the models at the session level. Considering this metric and its interaction with the other performance measures, the Random Forest model may be the preferred choice for predicting user retention for this specific approach, given that it slightly outperforms the Decision Tree and XGBoost models.

The precision-recall and ROC curves of the final model are presented in Figure 13 and Figure 14, respectively. Comparing the AUC score of this model with the one presented for the Gradient Boosting model in the previous section, it is evident that the performance of the model at the session level is poorer.

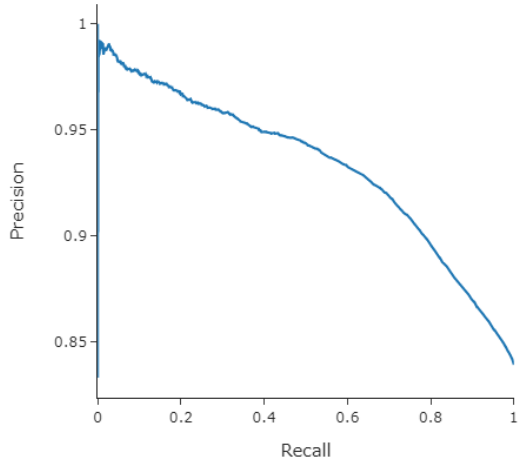

Figure 13: Random Forest Classifier Precision-Recall curve

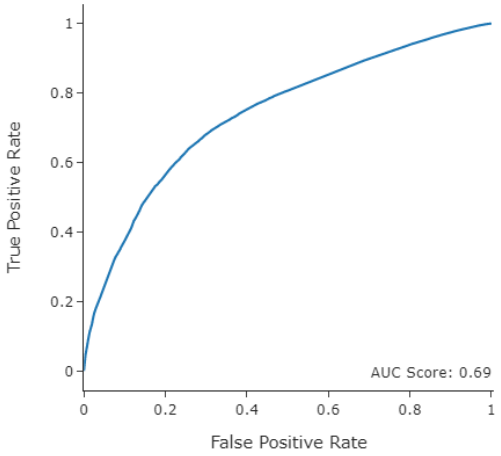

Figure 14: Random Forest Classifier ROC curve

The confusion matrix shown in Figure 15 provides insights into the model's classification performance. It indicates that 68% of the instances belonging to the churned class were correctly classified as churned class (true negatives), while 32% were mistakenly classified as the retained class (false positives). Moreover, 30% of the instances belonging to the retained class were misclassified as the churned class (false negatives), while 70% of the instances were correctly

identified as the retained class (true positives).

These results highlight the need for further investigation and improvement in the models' performance for predicting retention at the session level and utilizing the predicted probabilities for decision-making. This improvement process may involve further fine-tuning the model parameters to optimize their predictive capabilities. Additionally, incorporating additional relevant features that provide a more comprehensive understanding of users' behavior and their interactions with the app could enhance the accuracy and reliability of the predictions.

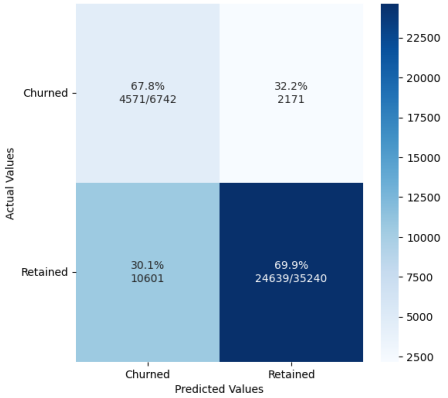

Figure 15: Random Forest Classifier Confusion Matrix

Regarding the feature importances of the model, Figure 22 indicates that average session duration, day of the session, days after creation, and whether the session is not a benchmark are some of the most relevant features that impact retention prediction at the session level. By examining the individual decision trees from the model and following the decision path, we can interpret the positive or negative effect of these variables on the prediction. The visualization of one of that decision trees (see Appendix B) suggests that shorter sessions are associated with a higher likelihood of user retention. On the other hand, if the session is marked as a benchmark, there is a higher likelihood of user churn.

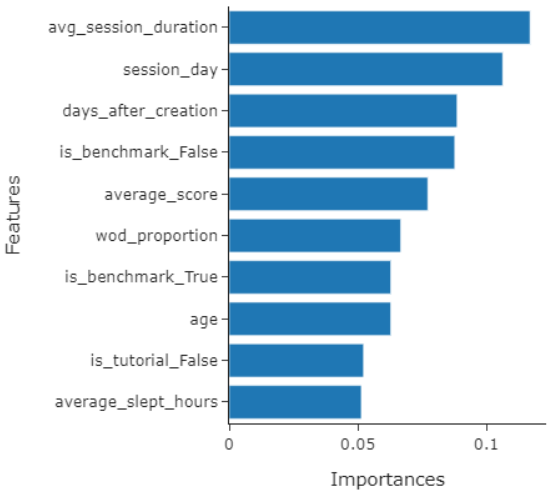

Figure 16: Feature importances Random Forest Classifier

5.2 Clustering methods

5.2.1 Modelling

In this section, we will present the results obtained from the K-Prototypes clustering analysis conducted on the given dataset. Through this analysis, we sought to segment the data into distinct groups based on the similarities in their features. As we presented in the methodology section, we utilized the Elbow Method. By plotting the cost against different values of k, we could visually inspect for an "elbow point" beyond which the decrease in cost became marginal or linear,indicating that adding more clusters beyond this point would not substantially improve the fit. This is depicted in Appendix A, where the elbow point to calculate the clusters is to be found between the values 4 and 6. Consequently, we selected 5 as the optimal number of clusters.

To better comprehend the segregation and structure of the 5 clusters formed, we show in Figure 17 below our results when using the T-SNE technique on the data.

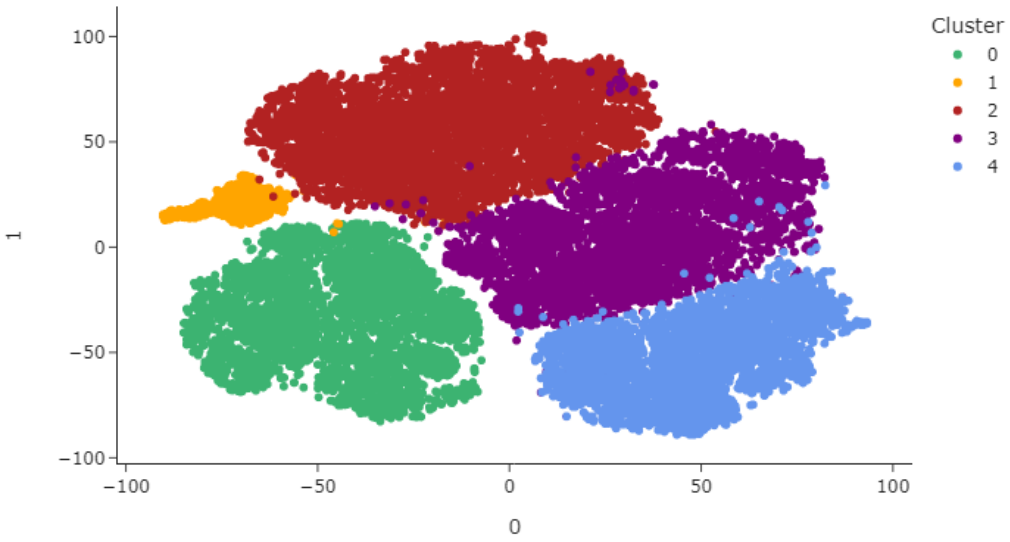

Figure 17: Cluster Visualization using T-SNE

A cursory glance at the plot indicates that the clusters are relatively segregated, suggesting that the K-Prototypes algorithm effectively identified distinct groups within the high-dimensional data space. Particularly, we can discern that Cluster 2 has the most substantial representation among all the clusters. Conversely, Cluster 1 seems to have the smallest representation. This is also supported in Table 10 below which outlines the number of users in each cluster.

Cluster	Number of Users
0	6859
1	830
2	8125
3	7493
4	5353

Table 10: Number of Users in Each Cluster

5.2.2 Profiling

This step provides an in-depth understanding of the characteristics of each cluster, helping to uncover the patterns and behaviors exhibited by users within these segments. With the help of the next graphs, we can detect the key features that make each cluster unique and form hypotheses about their nature and possible real-world interpretations.

In Figure 18 we provide a visualization into the distribution of the numerical variables across the different user clusters. Here we facilitates an analysis of the clusters in terms of age, proportion of Workout of the Day, amount of sessions, average score, average hours slept, and average session duration. Each box plot within the subplots represents a cluster and displays the distribution of the corresponding attribute within that cluster.

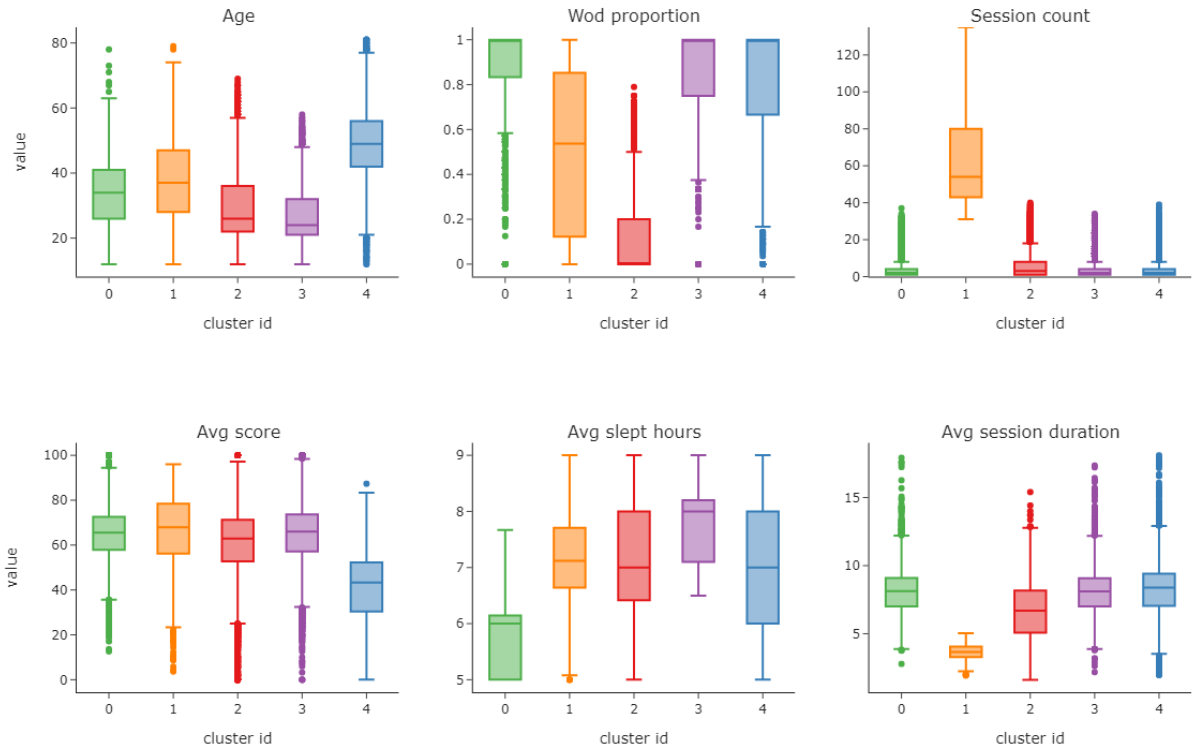

Figure 18: Distribution of characteristics of users by cluster

In addition to examining numerical variables, we explored the distribution of categorical variables among different user clusters (see Figures 19 and 20). Each bar plot provides insights into a specific categorical variable, including Gender, Language, Education, and Frequent Mood. It is worth mentioning that we also analyzed two additional categorical variables, Profession and Game. However, due to the large number of categories involved and the limited insights gained from visualizations, we decided not to display these variables graphically. Each of these plots employs horizontal bars, where the y-axis is used for the categorical variable's classes, and the x-axis represents the proportion of total users in each category. Each color represents a cluster and the numbers within them represent the total amount of users within that cluster and category.

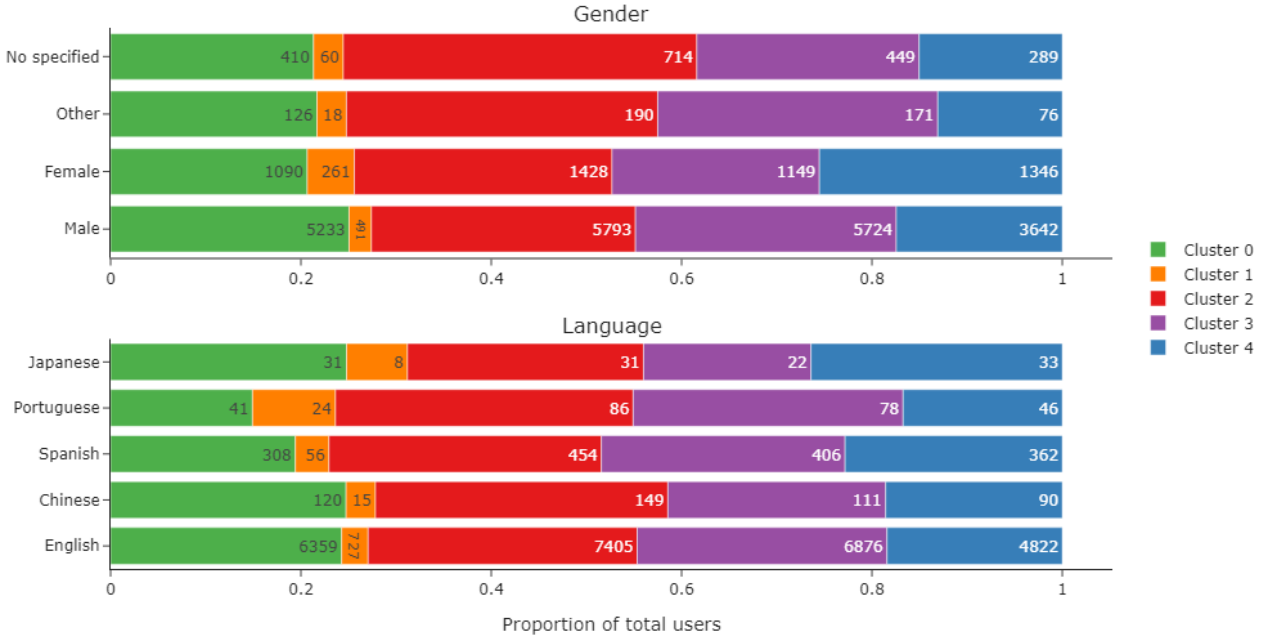

Figure 19: Distribution of characteristics of users by cluster

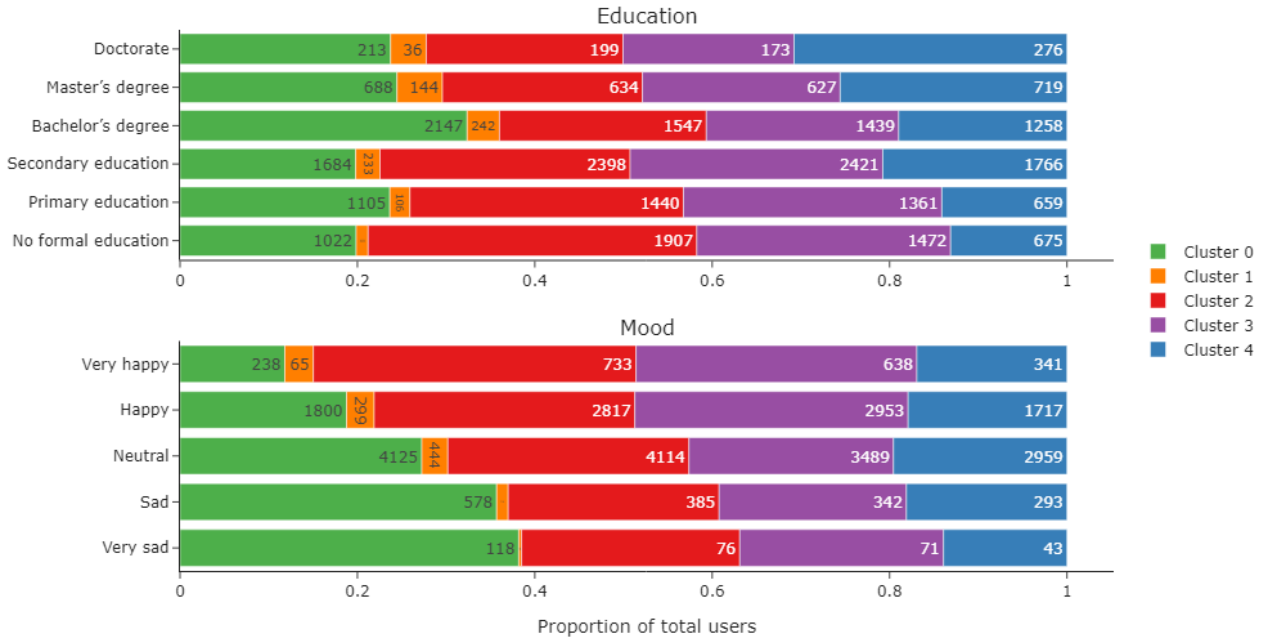

Figure 20: Distribution of characteristics of users by cluster

After examining the visualizations and delving into the various categorical characteristics, we can now consolidate the information garnered into concise summaries for each cluster. These summaries effectively create a profile that captures the attributes and tendencies of the categories within each cluster.

Cluster 0: The Scholarly Enthusiasts

This cluster encompasses 6859 users who average at 34 years of age, laying within the common age spectrum compared to other clusters. We can see that these users have a strong preference for the Workout of the Day game combination. This is evident as the lower quartile for the WOD proportion is notably high at 0.8. However, the frequency of play is sparse among them, exhibiting a low session count. When these users do immerse themselves in play, they tend to have longer sessions, averaging about 8.13 minutes. Albeit the relatively low engagement frequency, the sessions' scores paint a picture of regular performance. With an average of 6 hours of sleep, these users sleep the least in comparison to other clusters.

Turning to the demographic tapestry, males dominate this cluster, and English prevails as the primary language, as in all other clusters. Education-wise, a majority have attained education up to a bachelor's degree, and have the highest representation within this category. However, a deeper dive reveals that this cluster harbors the highest representation within the sad mood categories.

Synthesizing the above, Cluster 0 is composed of occasional but involved players, who indulge in extended sessions when they do play, and show an inclination toward the Workout of the Day. They maintain similar performance levels as other clusters despite lower engagement frequency and shorter sleep duration.

Cluster 1: The Dedicated Explorers

Users in this cluster have a median age of 34 and with only 830 they conform the smallest cluster. They show a wide range in their preference for the WoD, with the interquartile range spanning from around 0.2 to 0.8. They have the highest frequency of usage, with an average of 54 sessions per user, but a low duration of each of these session. Having these low session lengths might indicate that these users are exploring different games and features within the application. They may benefit from new content and challenges due to their high engagement. They engage with the Enhance VR application more regularly than any other group. This indicates a high level of commitment to cognitive training, obtaining the highest scores these users are consistent in their performance.

The educational background of the majority is either a bachelor's degree or secondary education. As for the mood, individuals in this cluster report a mix of neutral and happy as their most frequent moods. Their name is attributed due to their high engagement and versatile approach to the Enhance VR application.

Cluster 2: The Young Selectives

Cluster 2 encompasses 8125 users, who are primarily on the younger side with an average age of 26 years. They are discerning in their game choices, as reflected in their lukewarm interest towards the Workout of the Day. Engaging in an average number of sessions with moderate session durations and scores, their involvement seems balanced rather than performance-driven. Predominantly holding secondary education qualifications, this cluster's mood profile is commendable, with a notable proportion reporting being happy or very happy. This positive disposition, coupled with their selective approach to gaming, suggests that they might be seeking enjoyment and aligning their choices with personal preferences. For Enhance VR, this cluster's youthful energy and selectiveness might present an opportunity.

Cluster 3: The Young Achievers

Cluster 3, consists of 7493 users who have the youngest distributio among all the clusters with an average age of 24 years. These players exhibit a marked preference for the Workout of the Day, as indicated by the mean and median of WoD proportion at 1. They engage in a standard number of sessions and boast the second-highest scores, reflecting a commendable performance. A distinctive trait of this cluster is their healthier sleep patterns, averaging 8 hours, which might be linked to their youthful age. Additionally, possibly linked to their age, the cluster is predominantly composed of individuals who have attained secondary education. Mood-wise, this cluster mirrors Cluster 2 with a significant number of users reporting being happy or very happy.

It is evident that Cluster 3 represents a young and enthusiastic group who relish the Workout of the Day, maintain healthy sleep patterns, and derive positive experiences from the application.

Cluster 4: The Veteran Navigators

Cluster 4 encompasses 5353 users and stands out due to the higher average age of its members, around 49 years. These mature players, who are actively engaged in work life, tend to be meticulous in their approach. They don't engage as frequently, but when they do, they indulge in substantial session lengths. They exhibit a fondness for the Workout of the Day, albeit with

more variation compared to the younger clusters. On average, they sleep around 7 hours a night. Performance-wise, they have a tendency to score lower across all games compared to other clusters. The demographic primarily comprises males who are English speakers, and most have a background in secondary education. In terms of mood, the Veteran Navigators report a balance of neutral and happy moods. With their age and meticulous engagement in mind, Enhance VR might find it beneficial to tailor the experience for these veteran users by using elements such as goal-setting and progress tracking.

In an effort to provide further insight into user engagement within the different clusters, retention rates were calculated for each cluster. Retention, in this context, is defined as the proportion of active users who returned to play after N days from the time the app was installed. The retention rate was computed by dividing the number of active users of each cluster on day N by the total number of users who installed the app on day 0, across all clusters. This metric is critical as it indicates the fraction of the total active user base that continues to engage with the games after N days, belonging to a specific cluster. Table 11 summarizes the retention rates for clusters 0 to 4 over the various time frames we have used previously. As time progresses, there is a general decrease in retention rates, although some clusters exhibit more stability or even growth. Given the number of users in each cluster is different, direct comparisons among clusters cannot be made since the measurement is based on the total number of users who installed the app on day 0. Instead, to better interpret retention within each cluster, it is more appropriate to observe its evolution over the measured days.

Day N	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
D0	22.99	6.04	27.62	24.69	17.74
D1	5.53	4.92	7.45	5.90	5.30
D7	3.97	4.66	4.39	4.01	4.87
D15	3.47	4.66	3.85	3.29	4.64
D30	2.92	4.28	3.56	3.06	6.51

Table 11: Day N Retention for Clusters 0 to 4

Detailed Clustering profiles

In the preceding sections, we have presented general visualizations that give an overview of the various characteristics of the user clusters. While these visualizations provide a good introduction to the differences and similarities among clusters, Tables 12 and 13 below, delve into the specifics by presenting a comprehensive breakdown of each cluster. These tables serve as a reference map. Table 12, contains information on demographic variables such as gender, age, profession, education and language, and the user proportion. Table 13 delves into games, game categories, mood, average amount of sessions, average score, average slept hours and the WOD proportion.

Table 12: Clusters Table - Part 1

Variables	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Users Proportion	23.93	2.9	28.35	26.14	18.68
Gender	Male: 76% Female: 15% Other: 1% Not Spec.: 5%	Male: 59% Female: 31% Other: 2% Not Spec.: 7%	Male: 71% Female: 17% Other: 2% Not Spec.: 8%	Male: 76% Female: 15% Other: 2% Not Spec.: 5%	Male: 68% Female: 25% Other: 1% Not Spec.: 5%
Age	33	37	28	26	49
Profession	Other: 38% Tech.: 18% Engin.: 8% A & D: 7% H & Med.: 5% Educ.: 5% Mng.: 4% R&D: 2% Fin.: 2% Media: 2% Military: 1% Law: 1%	Other: 40% Tech.: 13% Educ.: 10% H&Med.: 9% Engin.: 7% A & D: 4% Mng.: 4% R&D: 3% Fin.: 2% Media: 1% Military: 1% Law: 0%	Other: 47% Tech.: 12% Educ.: 7% Engin.: 7% A & D: 6% H&Med.: 5% Mng.: 3% Media: 2% Fin.: 2% R & D: 2% Military: 1% Law: 1%	Other: 45% Tech.: 14% Educ.: 7% A & D: 7% Engin.: 7% H&Med.: 4% Mng.: 3% Media: 2% Fin.: 2% R & D: 1% Military: 1% Law: 1%	Other: 39% Tech.: 15% Engin.: 7% H&Med.: 7% A&D: 6% Educ.: 6% Mng. 6% Fin.: 2% Media: 2% R & D: 2% Law: 1% Military: 1%
Education	BS: 31% HS: 24% PE: 16% NF Educ.: 14% Master: 10% Doc.: 3%	BS: 29% HS: 28% Master: 17% PE: 12% Doc.: 4% NF Educ.: 8%	HS: 29% BS: 19% PE: 17% Master: 7% NF Educ.: 23% Doc.: 2%	HS : 32% BS: 19% NF Educ.: 19% PE: 18% Master: 8% Doc.: 2%	HS: 32% BS: 23% Master: 13% NF Educ.: 12% PE: 12% Doc.: 5%
Language	Eng.: 92% Sp.: 4% Ch.: 1%	Eng.: 87% Sp.: 6% Por.: 2% Ch.: 1%	Eng.: 91% Sp.: 5% Ch.: 1% Por.: 1%	Eng.: 91% Sp.: 5% Ch.: 1% Por.: 1%	Eng.: 90% Sp.: 6% Ch.: 1%

Table 13: Clusters Table - Part 2

Variables	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Games	React: 32% MW: 20% H&S: 11% Balance: 5% PB: 5% Assembly: 5% WaM: 5% Odd Egg: 4% MagD: 3% Slinger: 1%	React: 25% PB: 16% WaM: 9% MW: 8% Balance: 5% H&S: 5% Odd Egg: 5% Stacker: 5% MagD: 4% Harmonize: 4% Maestro: 3% Assembly: 2% Slinger: 2%	React: 27% WaM: 16% PB: 14% Assembly: 10% H&S: 4% MW: 6% Balance: 4% Odd Egg: 4% Slinger: 3% Maestro: 1% Harmonize: 1% Stacker: 2% MagD: 2%	React: 35% MW: 17% H&S: 11% Balance: 6% PB: 6% WaM: 6% Odd Egg: 4% Assembly: 4% MagD: 2% Slinger: 1% Stacker: 1%	React: 31% H&S: 15% MW: 12% Odd Egg: 8% PB: 7% Assembly: 4% MagD: 3% Balance: 3% WaM: 4% Harmonize: 3% Stacker: 2% Slinger: 1% Maestro: 1%
Game Category	Flex.: 32% Memory: 24% Att.: 12% Or.: 11% Proc.: 6% MC and Att.: 5% PS: 5% MC: 1%	Att.: 26% Flex.: 25% Memory: 16% PS: 10% Proc.: 7% Or.: 6% MC and Att.: 5% MC: 2%	Att.: 31% Flex.: 27% Memory: 10% Proc.: 11% PS: 6% MC: 3% MC and Att.: 4% Or.: 4%	Flex.: 35% Memory: 20% Or.: 12% Att.: 13% MC and Att.: 6% Proc.: 5% PS: 5% MC: 1%	Flex.: 31% Memory: 17% Or.: 15% Att.: 12% PS: 10% Proc.: 7% MC and Att.: 3% MC: 1%
Mood	Neutral: 60% Happy: 26% Sad: 8% Very happy: 3% Very sad: 1%	Neutral: 53% Happy: 36% Very happy: 7% Sad: 2% Very sad: 0%	Neutral: 50% Happy: 34% Very happy: 9% Sad: 4% Very sad: 0%	Neutral: 46% Happy: 39% Sad: 4% Very happy: 8% Very sad: 0%	Neutral: 55% Sad: 5% Happy: 32% Very happy: 6% Very sad: 0%
Avg. Amount of Sessions	3	67	6	3	3
Avg. Score	64	65	60	64	40
Avg. Slept Hours	5	7	7	7	6
WOD Proportion	89%	50%	11%	89%	81%

6 Conclusion

Based on our study, focusing on the benchmark session data, rather than aggregating data across all sessions, was more effective in predicting user retention. The first impression and initial interaction with the game play a critical role in determining whether users will continue playing and remain engaged over time. Among the features analyzed, session-related variables such as the number of sessions, gameplay duration, scores achieved, proportion of games played using the workout of the day, and whether it was the benchmark session or not consistently emerged as significant predictors. This suggests that prioritizing strategies based on these session-related aspects can enhance user engagement and increase the likelihood of long-term retention.

Our clustering analysis uncovered distinct user segments characterized by diverse behaviors and characteristics, which influenced user retention rates. Notably, session length, engagement frequency, game preferences, emerged as key variables that differentiated these clusters. Interestingly, it was observed that the gameplay characteristics exhibited more dissimilarities among clusters compared to the demographic characteristics. This emphasizes the importance of understanding and leveraging gameplay-related factors to tailor retention strategies effectively.

To enhance session-level retention prediction, potential extensions can be explored. This includes parameter tuning and additional feature analysis to improve the accuracy of predictive models. Another avenue worth considering is the incorporation of survival ensemble trees, as discussed in existing literature, which can provide insights into the likelihood of user retention over longer periods.

Integrating the predictive model and clustering sections by offering retention predictions for each cluster can greatly enhance the practical utility of the study. By estimating the retention probability of users within each cluster, tailored retention strategies can be developed based on the specific characteristics associated with each segment. Furthermore, expanding the analysis from clusters to individual games can provide more precise insights into user behavior within each game.

In conclusion, our study demonstrated the superiority of user-level predictive models over session-level models for retention prediction in this specific context. We identified significant features that influence user retention and emphasized the value of clustering analysis in understanding user segments and their impact on retention rates. While there are areas for improvement and potential extensions, our findings serve as a starting point for enhancing user retention and customizing strategies to different user segments.

References

Amplitude (2020). Why n day retention is the metric that matters for mobile games.

Bauckhage, C., Drachen, A., and Sifa, R. (2014). Clustering game behavior data. *IEEE Transactions on Computational Intelligence and AI in Games*, 7(3):266–278.

Bekkar, M., Djemaa, H. K., and Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl*, 3(10).

Brugada-Ramentol, V., Bozorgzadeh, A., and Jalali, H. (2022). Enhance vr: A multisensory approach to cognitive training and monitoring. *Frontiers in Digital Health*, 4:916052.

Debeauvais, T., Lopes, C. V., Yee, N., and Ducheneaut, N. (2014). Retention and progression: Seven months in world of warcraft. In *FDG*.

Drachen, A., Lundquist, E., Kung, Y., Rao, P., Sifa, R., Runge, J., and Klabjan, D. (2016). Rapid prediction of player retention in free-to-play mobile games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 12, pages 23–29.

Dutra, M. J. M. (2022). Customer profiling in the ambit of gaming: portraying lifestyles. *IROCAMM: International Review of Communication and Marketing Mix, 5 (2), 95-118.*

GoPractice (2023). Retention: how to understand, calculate, and improve it.

Hadiji, F., Sifa, R., Drachen, A., Thurau, C., Kersting, K., and Bauckhage, C. (2014). Predicting player churn in the wild. In *2014 IEEE Conference on Computational Intelligence and Games*, pages 1–8.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Lee, S.-K., Hong, S.-J., Yang, S.-I., and Lee, H. (2016). Predicting churn in mobile free-to-play games. In *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1046–1048.

Martino, A., De Santis, E., Baldini, L., Rizzi, A., et al. (2019). Calibration techniques for binary classification problems: A comparative analysis. In *IJCCI*, pages 487–495.

Milošević, M., Živić, N., and Andjelković, I. (2017). Early churn prediction with personalized targeting in mobile social games. *Expert Systems with Applications*, 83:326–332.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. ICML '05, page 625–632, New York, NY, USA. Association for Computing Machinery.

Periáñez, , Saas, A., Guitart, A., and Magne, C. (2016). Churn prediction in mobile social games: Towards a complete assessment using survival ensembles. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 564–573.

Ranti, K. S., Salim, K., and Girsang, A. S. (2019). Clustering steam user behavior data using k-prototypes algorithm. In *Journal of Physics: Conference Series*, volume 1367, page 012018. IOP Publishing.

Runge, J., Gao, P., Garcin, F., and Faltings, B. (2014). Churn prediction for high-value players in casual social games. In *2014 IEEE Conference on Computational Intelligence and Games*, pages 1–8.

Su, Y., Backlund, P., and Engström, H. (2021). Comprehensive review and classification of game analytics. *Service Oriented Computing and Applications*, 15:141–156.

SuperData (2020). 2020 year in review: Digital games and interactive media.

Tekin, A. T., Cebi, F., and Kaya, T. (2023). Retention prediction in the gaming industry: Fuzzy machine learning approach. In Calisir, F., editor, *Industrial Engineering in the Age of Business Intelligence*, pages 103–117, Cham. Springer International Publishing.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Viljanen, M., Airola, A., Pahikkala, T., and Heikkonen, J. (2016). Modelling user retention in mobile games. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8. IEEE.

Wang, Y., Huang, H., Rudin, C., and Shaposhnik, Y. (2020). Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *CoRR*, abs/2012.04456.

Appendices

A Elbow Method Visualization

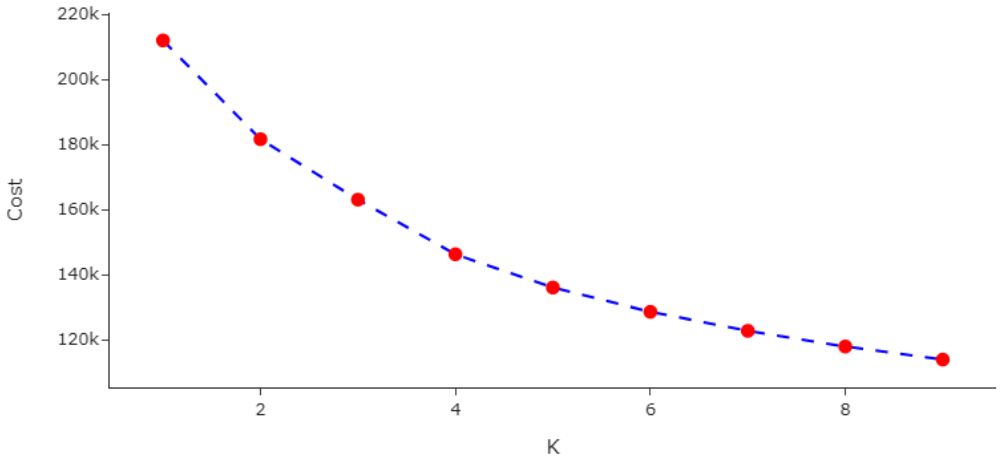

Figure 21: Optimal amount of clusters using Elbow method

B Random Forest Visualization

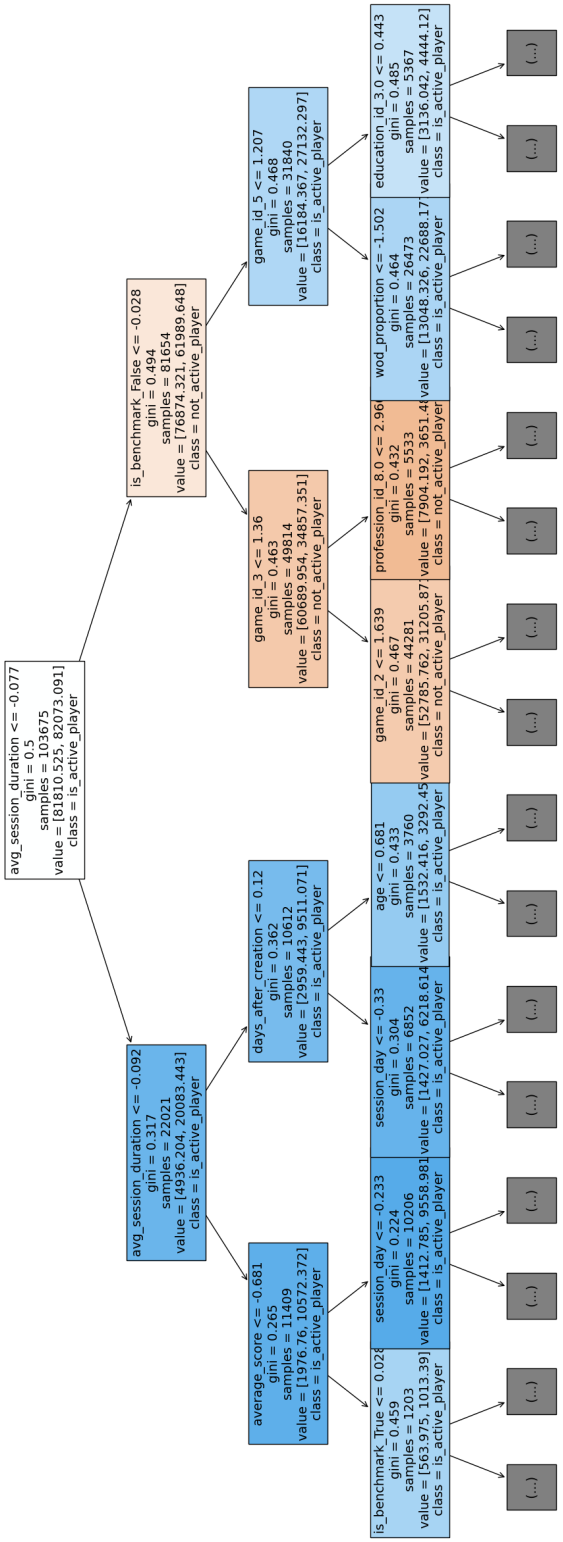

Figure 22: Individual decision tree from Random Forest