Player Profiling in World of Warcraft Using CTGAN-Generated Synthetic Data and Bartle's Archetypes

Fernando Bernardo dos Santos[1], **Augusto Cesar Espíndola Baffa**[1], **José Ricardo da Silva Junior**[2], **José Alberto Rodrigues Pereira Sardinha**[1]

[1]Departamento de Informática
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)
Rio de Janeiro, RJ – Brasil

fernandobsantos@aluno.puc-rio.br, {abaffa, sardinha}@inf.puc-rio.br

[2]Instituto Federal de Educação, Ciência e Tecnologia do Rio de Janeiro (IFRJ)
Rio de Janeiro, RJ – Brasil

jose.junior@ifrj.edu.br

Abstract. Introduction: *Player behavior analysis in MMORPGs is essential for game analytics, yet it remains constrained by limited access to large-scale player telemetry. Existing player classification approaches, such as Bartle's taxonomy, typically assign players to a single discrete category, overlooking mixed motivations.* ***Objective:*** *This work proposes a methodology for profiling World of Warcraft players using synthetic data generated by Conditional Tabular GANs (CTGAN), extending Bartle's archetypes to percentage-based representations that capture mixed motivations.* ***Methodology:*** *The World of Warcraft Avatar History (WoWAH) dataset was preprocessed and used to train four generative models (Gaussian Copula, CTGAN, TVAE, CopulaGAN) under identical conditions. CTGAN was selected based on superior correlation preservation ($\Delta = 0.056$), which is particularly relevant for downstream profiling tasks that rely on combinations of gameplay metrics. A percentage-based scoring framework maps gameplay features to archetype scores using conceptually defined weights grounded in Bartle's taxonomy and supported by empirical findings from prior work.* ***Results:*** *CTGAN preserved categorical distributions (Chi-Square $p > 0.99$) and inter-feature correlations with the highest fidelity among all evaluated models. The profiling methodology produced strong correlations between archetype scores and expected gameplay metrics ($r > 0.85$), with distributions consistent with known characteristics of the game. These results suggest that synthetic data combined with continuous player profiling enables behavioral analysis in data-scarce environments, reducing dependency on proprietary telemetry and supporting more accessible game analytics research.*
Keywords *Player profiling, Synthetic data, CTGAN, Bartle's archetypes, MMORPG, World of Warcraft.*

1. Introduction

Player behavior analysis in massively multiplayer online role-playing games (MMORPGs) has become increasingly important as these virtual environments attract millions of users with diverse motivations and playstyles. Understanding how players engage with complex virtual worlds is critical for game design, personalization, and broader research on human motivation in digital spaces.

However, AI-driven player analysis faces a critical dependency: it requires large volumes of behavioral data for training. This dependency tends to favor large studios. Companies such as Blizzard, Riot Games, and Electronic Arts accumulate years of continuous telemetry collection, while most player telemetry remains proprietary [Drachen et al. 2013]. Public datasets are rare exceptions, creating barriers for smaller studios and academic researchers.

Bartle's taxonomy [Bartle 1996] provides a foundational framework for classifying player motivations, identifying four archetypes: Killers (competition), Achievers (progress), Socializers (interaction), and Explorers (discovery). Although influential, it assumes that players fit strictly into a single category, overlooking the fluidity of human motivation. In practice, players exhibit mixed motivations that vary with context and game phase.

This research addresses both challenges by combining synthetic data generation with flexible player profiling. By using Conditional Tabular GANs (CTGAN) [Xu et al. 2019] to generate realistic synthetic player profiles from the World of Warcraft Avatar History (WoWAH) dataset [Lee et al. 2011], and by extending Bartle's taxonomy to support percentage-based profiling, this work presents a methodology that reduces data access barriers while providing a more nuanced player classification. The research is guided by two questions: **RQ1** How can synthetic data generated by CTGAN preserve both the statistical properties and the gameplay context of real player behavior?; **RQ2** How can Bartle's archetypes be quantified with percentage weights to represent players' mixed motivations more accurately?

This work makes three key contributions: (1) a synthetic-data-driven methodology for player profiling that removes the dependency on proprietary telemetry in game analytics; (2) a continuous reformulation of Bartle's taxonomy enabling interpretable and mixed-motivation player representations; (3) evidence that preserving inter-feature correlations is a critical requirement when using generative models for downstream behavioral profiling tasks. These contributions are particularly relevant for game analytics, where the lack of accessible player data often limits the applicability of data-driven methods. By combining synthetic data generation with interpretable profiling, this work provides a practical alternative for both academic research and small-scale game development contexts.

The remainder of this paper is organized as follows. Section 2 reviews related work on player motivation modeling and synthetic data generation. Section 3 describes the dataset, preprocessing pipeline, generative models, and the proposed percentage-based profiling framework. Section 4 presents the experimental results, including the evaluation of synthetic data quality and profiling outcomes. Finally, Section 5 concludes the paper, discusses the implications of the findings, along with limitations and directions for future work.

2. Related Work

2.1. Player Motivation Research

The study of player motivations has evolved significantly since Bartle's original taxonomy. Yee [Yee 2006] identified three primary motivation components (achievement, social interaction, and immersion) through survey data collected from thousands of MMORPG players, demonstrating that players exhibit combinations of motivational factors rather than fitting into discrete categories. Subsequent studies further reinforce that player motivations are multidimensional and context-dependent [Hamari e Tuunanen 2014].

The HEXAD framework [Marczewski 2015, Tondello et al. 2016] proposes six user types grounded in Self-Determination Theory (SDT) [Ryan e Deci 2000], supported by a validated psychometric scale. While HEXAD provides a theoretically robust and empirically validated model, its application requires survey-based instruments that are typically unavailable in standard MMORPG telemetry datasets. In contrast, Bartle's taxonomy remains more practical for telemetry-driven analysis, as it can be approximated using behavioral metrics derived from gameplay logs. Several works have explored the use of telemetry data to infer player types. Bicalho et al. [Bicalho et al. 2019] combined K-means clustering with decision trees to classify players into extended Bartle categories using singleplayer telemetry, achieving 75–80% accuracy when validated against survey data. Normoyle and Jensen [Normoyle e Jensen 2021] applied hierarchical Bayesian clustering to multiplayer telemetry and found that emergent play styles aligned with Bartle's archetypes, suggesting that the framework captures meaningful behavioral regularities even when not explicitly enforced.

Despite these advances, existing approaches present key limitations. Survey-based models require data that is typically unavailable in game telemetry, while clustering-based methods produce discrete categorizations that fail to capture the inherently mixed and dynamic nature of player motivations. These limitations motivate the need for continuous, telemetry-compatible profiling approaches.

2.2. Synthetic Data Generation

Generating realistic synthetic tabular data is a challenging problem due to the presence of mixed data types, complex dependencies, and imbalanced categorical distributions. CTGAN [Xu et al. 2019] addresses these challenges through three key mechanisms: (i) mode-specific normalization for modeling multimodal continuous variables, (ii) conditional generation to capture dependencies between categorical features, and (iii) training-by-sampling to mitigate class imbalance.

The field has since expanded beyond GANs, with diffusion-based approaches such as TabDDPM [Kotelnikov et al. 2023] and transformer-based architectures, surveyed comprehensively by Shi et al. [Shi et al. 2025]. CTGAN and related models have demonstrated strong performance in domains such as healthcare [Ahmed et al. 2025] and finance [Szymura 2024], where data privacy and limited access are critical concerns. However, their application to game analytics remains relatively unexplored, particularly in the context of player behavior modeling and profiling.

In the context of this work, synthetic data generation serves a dual purpose: (i) reducing dependency on proprietary telemetry datasets, and (ii) enabling controlled evaluation of profiling methodologies under realistic data distributions. Unlike prior work, which typically assumes access to real player data, this study explicitly investigates whether synthetic data can preserve both statistical properties and gameplay semantics relevant to downstream behavioral analysis.

Table 1 summarizes the positioning of this work relative to existing literature and highlights key differences between existing approaches and the proposed methodology. Prior work on player motivation either relies on survey data (e.g., Yee) or applies telemetry-driven clustering methods that produce discrete or emergent categorizations (e.g., Bicalho et al., Normoyle and Jensen). In contrast, this work combines synthetic data generation with a continuous, percentage-based extension of Bartle's taxonomy, enabling the representation of mixed player motivations directly from telemetry-like data. Additionally, while CTGAN has been studied primarily as a general-purpose tabular data generator, its integration with player profiling represents a novel application within game analytics. This positioning underscores the dual contribution of the proposed approach: addressing data accessibility through synthetic generation and enhancing expressiveness in player modeling through continuous profiling.

Table 1. Comparison with related work.

Work	Synth. Data	Framework	Profiling	Data
Yee (2006)	No	Custom	Percentage	Survey
Bicalho et al. (2019)	No	Bartle (ext.)	Clustering	Telemetry
Normoyle & Jensen (2021)	No	Bartle (emerg.)	Categorical	Telemetry
Xu et al. (2019)	CTGAN	N/A	N/A	N/A
This work	**CTGAN**	**Bartle (ext.)**	**Percentage**	**Telemetry**

3. Methodology

3.1. Dataset and Preprocessing

This research utilizes the WoWAH dataset [Lee et al. 2011], which contains temporal records of 91,065 unique avatars from World of Warcraft's The Burning Crusade (TBC) era. The preprocessing pipeline consisted of three steps: (1) temporal aggregation, reducing each character to a single record using the most recent snapshot; (2) random sampling of 10,000 profiles with a fixed seed to ensure reproducibility; and (3) feature engineering to derive gameplay metrics relevant to Bartle's archetypes. The engineered features include 7 numerical variables (`playtime_hours`, `gear_score`, `players_killed`, `pvptime_hours`, `dungeons_completed`, `raids_completed`, `profession_level`) and 4 categorical variables (`faction`, `specialization`, `role`, `profession`). In addition, auxiliary attributes such as character level were retained to support feature normalization and constraint enforcement. The final dataset consists of 10,000 player profiles with mixed numerical and categorical attributes, suitable for tabular generative modeling.

The engineered features fall into three groups according to how they were obtained. First, `level`, `race`, `charclass`, and `guild` are extracted directly from WoWAH. Second, `faction` is derived deterministically from race (TBC race–faction rules), while `specialization`, `role`, and `profession` are assigned through class-conditioned rules (e.g., role follows from class and specialization; professions use class-weighted probability distributions). Third, the numerical behavioral features absent from WoWAH (`gear_score`, `profession_level`, `players_killed`, `pvptime_hours`, `dungeons_completed`, `raids_completed`, and `playtime_hours`) are inferred through level-bracketed stochastic rules calibrated to TBC progression: each feature is sampled from ranges defined per level bracket (1–10, 11–30, 31–60, 61–70), with interdependencies enforced for coherence — `gear_score` grows with playtime at max level, `raids_completed` is zero below level 60 and increases with guild membership, `pvptime_hours` is bounded by kills and total playtime, and PvP-oriented classes receive higher kill ranges. All derived values respect TBC-specific rules (e.g., level cap 70 and profession cap 375), and the rules are fully specified in the replication package.

3.2. Synthetic Data Generation with CTGAN

Rather than selecting CTGAN a priori, four generative models were compared under controlled conditions (200 epochs, batch size 500, and 10,000 generated samples per model) using the SDV library [Patki et al. 2016]: **Gaussian Copula**: a statistical baseline modeling dependencies via a Gaussian copula; **CTGAN / TVAE** [Xu et al. 2019]: Conditional Tabular GAN designed for mixed-type tabular data; Tabular Variational Autoencoder with similar preprocessing to CTGAN; **CopulaGAN** [DataCebo, Inc. 2023]: a hybrid approach provided by the SDV ecosystem [Patki et al. 2016], combining copula transformations with GAN training.

CTGAN addresses key challenges of tabular data that standard GANs (designed for images) cannot handle: (i) mixed categorical and numerical data types; (ii) multimodal continuous distributions (e.g., player levels concentrated at both low values and the level cap); and (iii) imbalanced categorical distributions. Mode-specific normalization models each continuous variable using a Gaussian Mixture Model (GMM). Each value x is associated with a mixture component k with mean μ_k and standard deviation σ_k, and is normalized within that mode as $\alpha_j = (x - \mu_k)/\sigma_k$, where j indexes the feature. This representation is augmented with a one-hot encoded mode indicator $\boldsymbol{\beta}_j$. The conditional generator samples conditioned on categorical values to capture dependencies (e.g., Blood Elf $\rightarrow$ Horde). Training-by-sampling ensures balanced exposure to categorical values during training.

Models were evaluated using three criteria: (i) numerical fidelity, measured by the mean Kolmogorov-Smirnov (KS) statistic; (ii) categorical fidelity, assessed via Chi-Square tests; and (iii) correlation preservation, measured as the mean absolute difference between correlation matrices (Δ). Post-processing was applied to enforce domain constraints specific to TBC, including level bounds, raid eligibility, gear progression consistency, and faction–race compatibility. On average, approximately 30% of records per numerical feature required range adjustments, while dependent categorical attributes were re-derived from their parent columns to ensure logical consistency.

3.3. Percentage-Based Bartle Profiling

Each archetype score is computed as a weighted combination of min–max normalized gameplay metrics. Let $\hat{x} = (x - x_{\min})/(x_{\max} - x_{\min})$ denote the normalized value of a feature. The archetype scores are defined as:

$$S_{killer} = 0.70 \cdot \hat{x}_{kills} + 0.30 \cdot \hat{x}_{pvptime} \tag{1}$$

$$S_{achiever} = 0.50 \cdot \hat{x}_{gear} + 0.30 \cdot \hat{x}_{raids} + 0.20 \cdot \hat{x}_{profession} \tag{2}$$

$$S_{socializer} = 0.70 \cdot \hat{x}_{dungeons} + 0.30 \cdot \hat{x}_{raids} \tag{3}$$

$$S_{explorer} = 0.60 \cdot \hat{x}_{playtime} + 0.40 \cdot \hat{x}_{level} \tag{4}$$

Let $\mathcal{S} = \{\text{killer}, \text{achiever}, \text{socializer}, \text{explorer}\}$ denote the set of archetypes. The percentage-based profile is obtained by normalizing the scores $P_a = \frac{S_a}{\sum_{b \in \mathcal{S}} S_b} \times 100, \quad a \in \mathcal{S}$. This formulation produces mixed profiles (e.g., 40% Achiever, 30% Explorer), where the dominant archetype corresponds to the maximum P_a. This formulation ensures that each archetype score is a function of multiple behavioral dimensions, making the profiling process inherently dependent on inter-feature relationships. As a result, preserving correlations in the synthetic data becomes essential for maintaining the semantic validity of the resulting player profiles.

The weights were defined based on a combination of theoretical grounding and empirical evidence. Bartle's taxonomy [Bartle 1996] provides conceptual associations between gameplay behaviors and archetypes, while prior empirical studies such as Yee [Yee 2006] support the relevance of specific behavioral metrics (e.g., playtime, combat activity, and group participation). In this work, weights are treated as a heuristic but interpretable mapping between gameplay metrics and archetypal tendencies. While data-driven optimization of weights would require external ground truth labels (e.g., survey-based annotations), which are not available in WoWAH, the proposed formulation prioritizes interpretability and alignment with established theory.

Regarding sensitivity to the weight constants, two structural properties mitigate the impact of moderate variations. First, the metrics combined within each formula are strongly positively correlated with one another (e.g., `gear_score`, `raids_completed`, and `profession_level` all grow with progression), so shifting weight among them produces highly correlated score variants. Second, the dominant archetype assignment depends on the relative ordering of the four scores, which is driven primarily by *which* metrics enter each formula rather than by their exact coefficients. Consequently, the aggregate archetype distribution should remain stable under plausible weight perturbations, although individual borderline players near decision boundaries may switch dominant archetypes. A systematic sensitivity analysis quantifying this effect is left as future work.

3.4. Validation Strategy

Synthetic data quality (RQ1) is evaluated through multiple complementary metrics: (i) Chi-Square tests for categorical distributions; (ii) Kolmogorov-Smirnov (KS) tests for numerical distributions, interpreted with caution due to high statistical power at $n = 10,000$; (iii) relative differences in descriptive statistics (means and standard deviations); and (iv) correlation preservation (Δ). Profiling plausibility (RQ2) is evaluated using correlations between archetype raw scores and their associated gameplay metrics, as well as empirical validation of the weighting structure via multiple linear regression. Raw scores (prior to percentage normalization) are used to avoid compositional data artifacts [Aitchison 1986].

3.5. Artifact Availability

To support verifiability, a replication package is publicly available[1]. It contains: (i) the preprocessing scripts (temporal aggregation, sampling, and type definitions); (ii) the complete feature engineering rules, including the level-bracket ranges and class-conditioned assignment tables; (iii) the configuration and training code for all four generative models; (iv) the post-processing code enforcing TBC domain constraints; (v) the Bartle scoring functions and validation analyses (statistical tests, correlations, regression, and the K-means baseline); and (vi) instructions to reproduce all figures and tables. The WoWAH dataset itself is publicly available [Lee et al. 2011].

4. Results

All experiments were conducted using Google Colab's cloud environment (CPU: Intel Xeon @ 2.20GHz; no GPU/TPU in the selected runtime) and a fixed random seed (`random_state=42`) for dataset sampling to ensure reproducibility. The implementation was based on the SDV library, following the experimental protocol described in Section 3. Each generative model was trained under identical conditions (200 epochs, batch size 500), and evaluated on 10,000 generated samples; CTGAN training took 942.6s (15.7 minutes) for 200 epochs in this environment.

4.1. Baseline Model Comparison

TVAE achieved the best numerical distribution fidelity (mean KS = 0.069), while CTGAN achieved the best correlation preservation ($\Delta = 0.056$), approximately half of TVAE's error (0.101). All models preserved categorical distributions (6/6 features passing Chi-Square tests). CTGAN was selected for subsequent analysis due to its superior preservation of inter-feature relationships. This criterion is particularly relevant for Bartle profiling, where each archetype score is computed as a weighted combination of multiple metrics. Distortions in correlations may lead to incoherent profiles even when marginal distributions are well preserved. Table 2 summarizes the baseline comparison results.

Table 2. Baseline comparison of synthetic data generation models.

Model	Mean KS $\downarrow$	Corr. $\Delta\downarrow$	Cat.	Time
Gaussian Copula	0.298	0.321	6/6	5s
CTGAN	0.125	**0.056**	6/6	1134s
TVAE	**0.069**	0.101	6/6	177s
CopulaGAN	0.093	0.087	6/6	995s

Figure 1 presents per-column similarity scores, providing a more granular comparison. For numerical columns, TVAE and CopulaGAN achieved slightly higher fidelity, while CTGAN remained competitive. The Gaussian Copula showed consistently lower performance, particularly on zero-inflated features such as `raids_completed`. For categorical columns, all models achieved near-perfect p-values (≈ 1.0), indicating that categorical fidelity does not significantly differentiate the models.

[1]`https://github.com/fernandosaints/wow-ctgan-bartle-profiling`

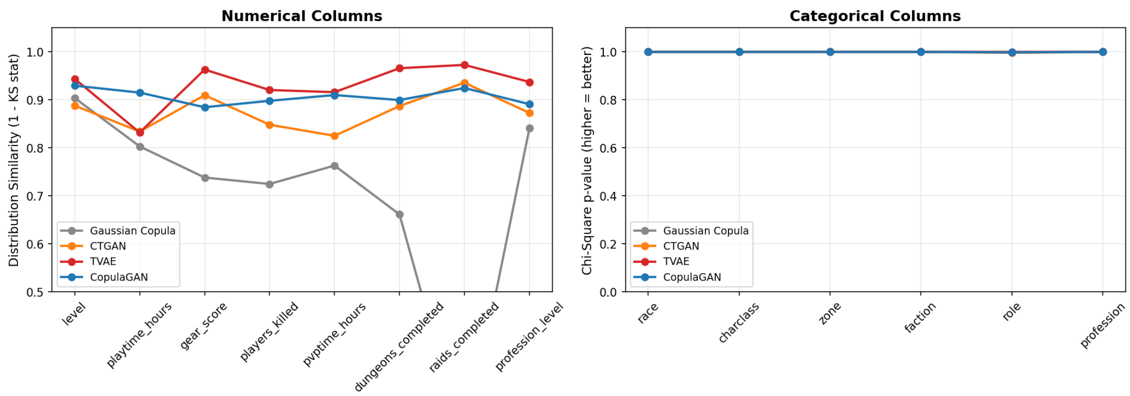

Figure 1. Per-column distribution similarity scores for numerical (left) and categorical (right) features across all four models. Higher values indicate better preservation.

4.2. Unsupervised Baseline (K-means, $k = 4$)

To provide a non-parametric profiling baseline, we applied K-means with $k = 4$ to the standardized numerical features of the post-processed synthetic dataset. The clustering achieved a silhouette score of 0.562, indicating well-separated groups. Cluster proportions were 54.6%, 15.2%, 13.0%, and 17.1%, and the resulting centroids were interpretable in terms of gameplay intensity (e.g., a low-level low-engagement cluster and an endgame high-activity cluster). This baseline complements the conceptual Bartle scoring by offering a data-driven grouping for comparison.

4.3. CTGAN Validation

Categorical distributions All categorical features achieved Chi-Square $p > 0.99$, indicating strong preservation of discrete distributions, including structured dependencies such as race–faction mappings and class–specialization combinations.

Numerical distributions KS tests showed statistical significance ($p < 0.001$) for all features, which is expected given the large sample size ($n = 10{,}000$) [Shi et al. 2025]. KS statistics ranged from 0.062 to 0.214, indicating relatively small distributional deviations in practical terms. Figure 2 illustrates the distribution comparison for player levels, showing that CTGAN captures the characteristic bimodal structure (concentration at low levels and at the level cap). Descriptive statistics indicate that standard deviations are preserved within 7.4% on average, while means differ by 13.3% on average. The mean shift is more pronounced in level-dependent features (level: 24.2%, playtime: 25.1%), whereas level-independent features exhibit stronger preservation (pvptime: 1.4%, raids: 0.0%). Since the profiling framework relies on min–max normalization (range-based rather than mean-based), these shifts are expected to have limited impact on the resulting archetype scores.

Correlation preservation A mean absolute difference of $\Delta = 0.056$ indicates strong preservation of inter-feature relationships, outperforming all other evaluated models. Figure 3 compares correlation matrices for real and synthetic data, showing that CTGAN maintains key structural dependencies between gameplay metrics.

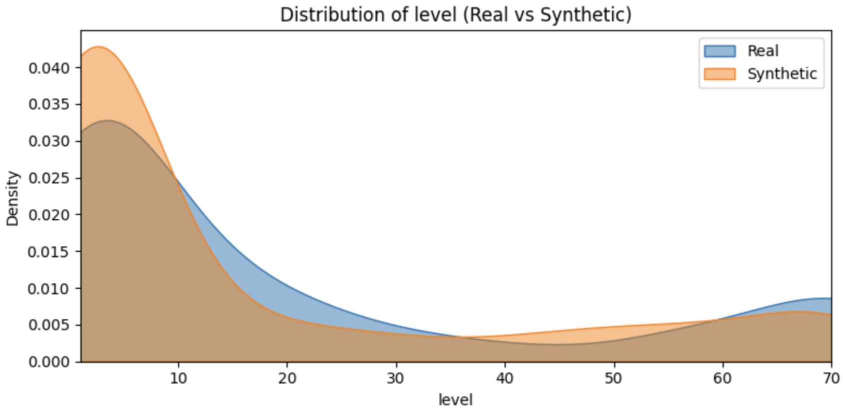

Figure 2. Distribution comparison of player levels between original and synthetic datasets. The bimodal structure is preserved.

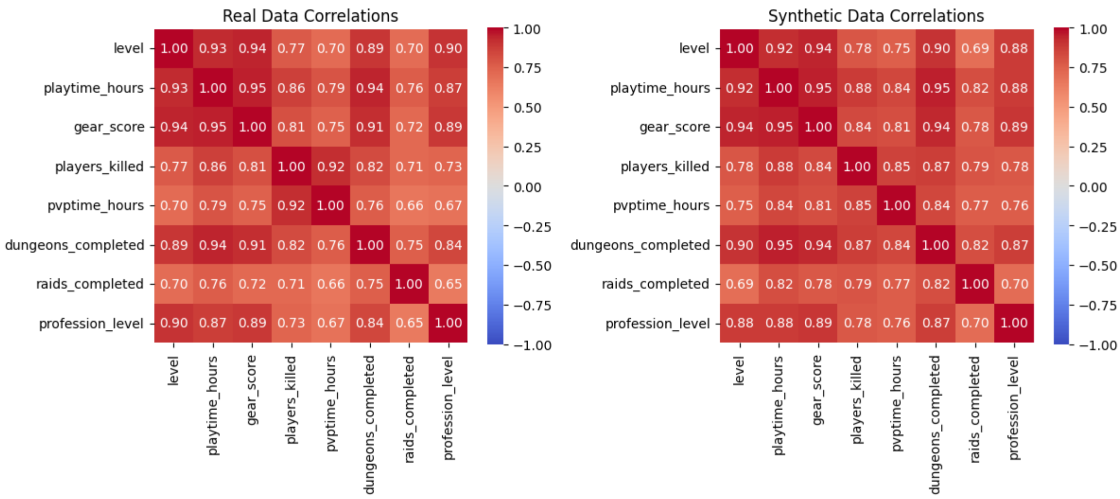

Figure 3. Correlation matrix comparison between original (left) and synthetic (right) datasets. Mean absolute difference $\Delta = 0.056$.

4.4. Bartle Profiling Results

The resulting archetype distribution (Achiever: 46.2%, Explorer: 33.9%, Killer: 13.3%, Socializer: 6.6%) is consistent with the TBC era's emphasis on progression and raid-oriented gameplay, aligning with prior findings on the prevalence of achievement-driven behavior in MMORPGs. Table 3 presents the correlation between archetype scores and their primary gameplay metrics.

Table 3. Correlation validation between Bartle archetype scores and primary metrics.

Archetype	Primary Metric	Pearson r	p-value
Killer	players_killed	0.92	<0.001
Achiever	gear_score	0.87	<0.001
Socializer	dungeons_completed	0.91	<0.001
Explorer	playtime_hours	0.85	<0.001

Table 4 reports the full correlation matrix between Bartle percentage scores and all numerical gameplay metrics.

Table 4. Pearson correlations between Bartle percentage scores and numerical gameplay metrics.

Score	Level	Playtime	Gear	Kills	PvP time	Dungeons	Raids	Prof.
Killer_pct	0.166	0.287	0.242	0.477	0.428	0.312	0.319	0.183
Achiever_pct	-0.687	-0.557	-0.571	-0.445	-0.428	-0.530	-0.355	-0.436
Socializer_pct	0.923	0.872	0.906	0.724	0.705	0.914	0.689	0.789
Explorer_pct	0.308	0.148	0.161	0.032	0.035	0.090	-0.007	0.079

These results reinforce that the proposed profiling framework captures consistent behavioral patterns across multiple gameplay dimensions, rather than relying on isolated features. This is particularly important in MMORPG contexts, where player behavior emerges from the interaction of multiple systems. Killer_pct aligns most strongly with PvP-oriented activity (players_killed: 0.477; pvptime_hours: 0.428), consistent with the intended interpretation of the archetype. Socializer_pct shows high correlations with cooperative and progression-related metrics (dungeons_completed: 0.914; playtime_hours: 0.872; gear_score: 0.906), reflecting that the available telemetry proxies social behavior primarily through group activities rather than direct interaction logs. Explorer_pct exhibits weaker correlations overall, as expected from its formulation being driven mainly by playtime and level. Finally, negative correlations for Achiever_pct should be interpreted with caution because percentage scores are compositional (they sum to 100%), which can induce trade-offs across archetypes when one dimension increases.

This "closure" effect [Aitchison 1986] deserves explicit treatment, as it directly shapes the patterns in Table 4. Because the four percentages are constrained to sum to 100%, an increase in any one archetype's share mechanically forces a decrease in the others. This explains the systematically negative row for Achiever_pct: as overall activity grows, the Socializer raw score (driven by dungeons and raids, which grow steeply with engagement) rises faster than the Achiever raw score, compressing the Achiever *share* even though the Achiever raw score itself increases with the same metrics ($r = 0.87$ with gear score, Table 3). For the interpretation of mixed motivations, this implies that percentage profiles should be read as *relative* emphases within a player's activity portfolio, not as absolute intensities: a decreasing Achiever percentage does not indicate less achievement-oriented behavior, but a portfolio shifting toward group content. For analyses of absolute behavioral intensity, the raw scores (used throughout the validation) are the appropriate representation. This dual reading also clarifies archetype separability in the telemetry space: separability is defined by the raw-score dimensions, while the percentages provide a normalized, player-centric summary of their balance.

All archetypes exhibit strong positive correlations ($r > 0.85$) with their designated primary metrics, indicating that the scoring functions behave consistently with their intended design. Multiple linear regression recovers the predefined weights with $R^2 = 1.0$, as expected given the linear formulation of the scoring functions. Additionally, regression using all gameplay features shows that non-designated variables have negligible coefficients, suggesting limited cross-archetype interference. It is important to state the scope of these results explicitly: because the scores are linear combinations of the metrics, the high correlations with primary metrics and the perfect regression fit are mathematical consequences of the scoring formulas. They therefore constitute *internal consistency* checks — verifying that the formulas were implemented correctly, that min–max normalization did not distort the intended weighting, and that each score responds to its designated metrics and not to others — rather than empirical evidence that the profiles match players' self-perceived motivations. External validity would require ground-truth labels (e.g., player surveys), as discussed in Section 5. From a game design perspective, the predominance of Achiever profiles (46.2%) reflects the progression-oriented structure of the TBC era, where endgame content such as raids plays a central role. The relatively lower proportion of Socializers may be partially explained by the absence of direct social interaction metrics (e.g., chat logs), highlighting the limitations of telemetry-only approaches.

These observations reinforce the importance of feature selection in player modeling and suggest that incorporating richer social signals could improve the identification of cooperative playstyles.

5. Conclusion

This paper presented a methodology for profiling World of Warcraft players using synthetic data generated by CTGAN, classifying them into Bartle's archetypes with percentage weights reflecting mixed motivations. A baseline comparison of four generative models empirically supported the selection of CTGAN due to its superior correlation preservation ($\Delta = 0.056$). While TVAE achieved better marginal distribution fidelity, the results highlight that model selection should be guided by downstream task requirements. For profiling tasks where features are combined through weighted formulas, preserving inter-feature relationships is critical.

CTGAN preserved categorical distributions (Chi-Square $p > 0.99$) and generated 10,000 synthetic profiles that approximate the statistical properties of the original WoWAH dataset. The percentage-based profiling produced strong correlations between archetype scores and expected gameplay metrics ($r > 0.85$), with an archetype distribution consistent with the TBC era's gameplay emphasis.

The proposed profiling framework extends traditional categorical Bartle classifications by enabling continuous representations of player behavior, capturing the multidimensional nature of player motivation identified by Hamari and Tuunanen [Hamari e Tuunanen 2014]. The approach is also conceptually aligned with Self-Determination Theory (SDT) [Ryan e Deci 2000], with Explorers associated with autonomy, Achievers with competence, and Socializers with relatedness.

Regarding applicability to other games and genres, the pipeline separates game-agnostic and game-specific components. The generative stage (model comparison, CTGAN training, statistical validation) operates on any tabular behavioral dataset and requires no game knowledge. The game-specific coupling is concentrated in two places: (i) the post-processing rules that enforce domain constraints (e.g., level caps, faction–race compatibility), and (ii) the mapping from available telemetry to archetype formulas. Applying the methodology to another title therefore requires its own telemetry sample and a one-time specification of its domain rules and metric-to-archetype mapping — a design effort rather than a methodological change. For genres with different motivational structures (e.g., competitive shooters without crafting systems), the archetype formulas would draw on different proxies, while the selection criterion identified here — prioritizing correlation preservation for downstream profiling — remains transferable. Reducing this coupling further, for example through declarative constraint specifications reusable across games, is a promising engineering direction.

Several limitations should be acknowledged. Although external validation with player-reported data is not available, we partially mitigate this limitation by: (1) comparing distributions with known MMORPG behavior patterns, (2) validating consistency with prior literature [Yee 2006, Hamari e Tuunanen 2014], and (3) analyzing alignment with expected gameplay semantics. The WoWAH dataset covers only *The Burning Crusade era* (2006–2009), limiting generalizability: modern MMORPGs feature higher-dimensional telemetry and different progression loops (e.g., Battle Pass systems, World Quests). We note, however, that the CTGAN methodology itself is agnostic to the semantics of the columns — higher dimensionality primarily increases training cost, and new progression mechanics would enter as additional features and post-processing rules; what ages is the behavioral snapshot, not the generative technique. The dataset lacks direct metrics for certain behavioral dimensions, such as social interaction (e.g., chat logs) and exploration achievements, requiring the use of proxy variables. This proxy limitation is most acute for the Socializer archetype: because its score is derived from group combat activities (dungeons and raids), the metric cannot distinguish a genuinely social player from a “high-intensity Achiever” who groups instrumentally for better loot, and thus risks labeling efficient progression as social motivation. Under the available telemetry, the Socializer score is best interpreted as a measure of *group-content engagement* — a behavioral proxy for social play — rather than of social motivation itself; disambiguating the two would require interaction data such as chat frequency, friend networks, or repeated grouping with the same players. The archetype weights were defined conceptually rather than learned from data, and the validation of the profiling framework is internal: correlations confirm consistency of the formulation but do not provide external evidence of alignment with player self-perception. Additionally, the absence of labeled archetype data prevents the use of evaluation protocols such as Train on Real, Test on Synthetic (TRTS) and Train on Synthetic, Test on Real (TSTR), which would provide stronger evidence of synthetic data utility. Guild membership saturation (99.6%) further limits its discriminative power for Socializer identification, and CTGAN may underrepresent extreme values in heavy-tailed distributions.

Future work should focus on strengthening both the generative and profiling components. Promising directions include validation with player surveys to establish external ground truth (e.g., using the compact Hexad-12 instrument [Krath et al. 2023]), empirical optimization of archetype weights using supervised data, a systematic sensitivity analysis of the weight vectors, and comparison with alternative profiling approaches such as GMM and PCA-based methods. On the generative side, more advanced models such as diffusion-based TabDDPM [Kotelnikov et al. 2023] and transformer-based approaches may further improve fidelity. Extending the methodology across multiple games and incorporating temporal dynamics would enable the study of how player profiles evolve over time.

A particularly actionable direction is a small-scale TSTR protocol: if even a modest labeled subset became available (e.g., a few hundred players completing a Bartle or Hexad questionnaire alongside telemetry collection), one could train an archetype classifier on synthetic profiles labeled by the scoring formulas and test it on the real labeled players. Classification performance approaching that of a model trained directly on real data would provide external evidence of both synthetic data utility and scoring validity, while a large gap would localize where the pipeline diverges from real player perception. In the same spirit, following a reviewer's suggestion, the full pipeline (generation, post-processing, and profiling) could be replicated end-to-end on a smaller publicly available telemetry dataset from another game, providing an additional validation of the method beyond WoWAH.

Overall, this work demonstrates that synthetic data generation and continuous player profiling can be jointly leveraged to enable robust behavioral analysis without access to proprietary datasets. Beyond the specific case of MMORPGs, the proposed approach provides a general framework for applying generative models to game analytics tasks where data availability is limited.

References

Ahmed, H. A., Nepomuceno, J. A., Vega-Márquez, B., e Nepomuceno-Chamorro, I. A. (2025). Synthetic data generation for healthcare: Exploring generative adversarial networks variants for medical tabular data. *International Journal of Data Science and Analytics*, 20(6):5739–5754.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, London.

Bartle, R. (1996). Hearts, clubs, diamonds, spades: Players who suit muds. *Journal of MUD Research*, 1(1):19.

Bicalho, L. F., Baffa, A., e Feijó, B. (2019). A game analytics model to identify player profiles in singleplayer games. In *2019 18th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*, pages 11–20. IEEE.

DataCebo, Inc. (2023). CopulaGAN synthesizer — SDV documentation. `https://docs.sdv.dev/sdv/single-table-data/modeling/synthesizers/copulagansynthesizer`. Accessed: 2026-07-15.

Drachen, A., El-Nasr, M. S., e Canossa, A. (2013). *Game Analytics: Maximizing the Value of Player Data*. Springer, London.

Hamari, J. e Tuunanen, J. (2014). Player types: A meta-synthesis. *Transactions of the Digital Games Research Association*, 1(2):29–53.

Kotelnikov, A., Baranchuk, D., Rubachev, I., e Babenko, A. (2023). TabDDPM: Modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 17564–17579. PMLR.

Krath, J., Altmeyer, M., Tondello, G. F., e Nacke, L. E. (2023). Hexad-12: Developing and validating a short version of the gamification user types hexad scale. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI)*, pages 1–18. ACM.

Lee, Y.-T., Chen, K.-T., Cheng, Y.-M., e Lei, C.-L. (2011). World of warcraft avatar history dataset. In *Proceedings of the Second Annual ACM Conference on Multimedia Systems (MMSys)*, pages 123–128. ACM.

Marczewski, A. (2015). *Even Ninja Monkeys Like to Play: Gamification, Game Thinking and Motivational Design*. Gamified UK.

Normoyle, A. e Jensen, S. T. (2021). Bayesian learning of play styles in multiplayer video games. arXiv preprint arXiv:2112.07437.

Patki, N., Wedge, R., e Veeramachaneni, K. (2016). The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410. IEEE.

Ryan, R. M. e Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78.

Shi, R. et al. (2025). A comprehensive survey of synthetic tabular data generation. arXiv preprint arXiv:2504.16506.

Szymura, A. (2024). Synthetic financial data: A case study regarding polish limited liability companies data. *Econometrics. Ekonometria. Advances in Applied Data Analysis*, 28(2):1–17.

Tondello, G. F., Wehbe, R. R., Diamond, L., Busch, M., Marczewski, A., e Nacke, L. E. (2016). The gamification user types hexad scale. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY)*, pages 229–243. ACM.

Xu, L., Skoularidou, M., Cuesta-Infante, A., e Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Yee, N. (2006). Motivations for play in online games. *CyberPsychology & Behavior*, 9(6):772–775.