July 2025

What Kinds of Users Engage with Different Types of Motivational Design: A Machine Learning Study in GitHub

Hairui Tang
Tampere University, hairui.tang@tuni.fi

Nannan Xi
Tampere University, nannan.xi@tuni.fi

Juho Hamari
Tampere University, juho.hamari@tuni.fi

What Kinds of Users Engage with Different Types of Motivational Design: A Machine Learning Study in GitHub

Completed Research Paper

Hairui Tang
Tampere University
Kalevantie 4, 33100 Tampere
hairui.tang@tuni.fi

Nannan Xi
Tampere University
Kalevantie 4, 33100 Tampere
nannan.xi@tuni.fi

Juho Hamari
Tampere University
Kalevantie 4, 33100 Tampere
juho.hamari@tuni.fi

Abstract

Motivational information system design, such as the use of gamification, has become a prominent dimension of IS design and research especially in user-facing systems where user engagement is crucial such as in crowdsourcing, electronic-commerce or online education. A major bottleneck currently is the inability to tailor or personalize motivational design to more accurately match users' behavioral patterns, background traits or personality. Towards this end we investigate how user behavior logs, collected from the last 15 years comprised of 10,000 users of GitHub, can be used to predict user engagement with different types of motivational design through machine learning. The results show that behavior-related metrics are the most predictive, followed by social network-related metrics, while user profile-related metrics are less influential. These findings challenge static user classification models, emphasizing the need for adaptive, network-aware gamification strategies. This study offers theoretical insights into gamification engagement and practical recommendations for personalized gamification design.

Keywords: Gamification, motivational design, social network analysis, machine learning, user engagement

Introduction

In today's constantly expanding digital space, nearly all systems and services that support individual and organizational activities and decision-making across various domains are increasingly being designed and refined to offer motivational benefits, rather than merely providing hedonic or utilitarian value (Hamari, 2019; Koivisto & Hamari, 2019). By incorporating motivational design elements, digital systems are charged with supporting engagement, eliciting contributions, and fostering the sense of community. Given its effectiveness, practical feasibility, and user familiarity, gamification commonly referring to the use of game-related elements, principles and techniques to create game-like experiences that fulfill users' motivational needs (Xi & Hamari, 2019; Klock et al., 2020), has emerged as one of the most prominent and widely discussed motivational design approaches.

Although motivational design represented by gamification has been developed in both practice and academia for over 15 years, the key challenge remains in effectively tailoring motivational elements to diverse user needs and contexts. Previous research in motivational design is rooted in generic frameworks and does not account for user engagement behavior heterogeneity (Klock et al., 2020). Traditional frameworks such as Self-Determination Theory (SDT) and Player Motivation Taxonomies have attempted to classify players in accordance with predetermined prototypes for motivation, for instance, achievement-motivated, socially motivated, or exploration-motivated players (Yee, 2006). The frameworks are rooted in self-reported surveys and not actual behavior analytics and are consequently susceptible to social desirability biases and are limited in application in fluid online environments (Santos et al., 2024). The majority of classification frameworks are individualist in nature and overlook dimensions related to network and social aspects of user behavior, particularly in collaboration-based environments (Takaoka & Jaccheri, 2024).

This research attempts to bridge these research gaps based on machine learning and social network analysis in examining user engagement in varied motivational design features. Relying on a rich database on GitHub for a period of 15 years and involving 10,000 users, this research applies some machine learning models (e.g., Random Forest, Support Vector Machines, and Logistic Regression) in predicting user engagement in motivational design features (Breiman, 2001; Strobl et al., 2008). The key research question guiding this research is as follow:

RQ: *What are user characteristics best predicting engagement with different motivational design features, more specifically, gamification elements which provide motivational benefits?*

To address this question, we use a GitHub big dataset and perform machine learning to identify key user features for predicting gamification engagement. By combining user demographics information (i.e., location, biography) and behavior metrics (i.e., repository activity, repository settings) and network features (i.e., centrality, clustering, connectivity), we are proposing a network-aware gamification system in which game mechanisms can be adapted based on user behavior and network position. Unlike prior user classification schemes, our approach presents a more active and responsive system for modeling gamification engagement in collaboration networks.

This research makes contributions in several different ways to theory and practice. *Theoretical contribution:* It presents a theory based on data for personalized motivational design, blending network analysis and machine learning in order to move beyond generic static classification schemes. *Methodological contribution:* By leveraging behavior-based information, network topologies, and predictive modeling, this research introduces a reproducible methodology for user engagement typologies identification in research in motivational design. *Practical contribution:* The findings in this research provide practical implications for online learning environments, digital platforms, and crowd-sourced systems for achieving maximum engagement based on network-sensitive and adaptive motivational strategies.

In essence, this research highlights the necessity to move away from generic motivational frameworks and towards data-based and adaptive practices in motivational design. By bridging behavior analytics and network-based engagement dynamics, this research aims to contribute towards designing more engaging, individualized, and sustainable motivators in online collaboration arenas.

Background

Motivational Design in Digital Systems

Motivational design is the combination of systematic strategies for engaging, persisting, and inducing desired user behavior in digital environments (Ryan & Deci, 2000). As distinct from usability design, in which efficiency and usability are mainly emphasized, motivational design applies psychological and behavior principles to ensure extended user engagement. Self-Determination Theory (SDT) (Ryan & Deci, 2000), Flow Theory (Csikszentmihalyi, 1990), and Expectancy-Value Theory (Eccles & Wigfield, 2002) have provided sound conceptual frameworks for research in motivational design and have highlighted intrinsic and extrinsic motivation in shaping engagement.

Among research assumptions in motivational design is that users act in digital systems based on autonomy, competence, and relatedness (Deci & Ryan, 1985). Perceived control in one's behavior, experience of mastery in overcoming obstacles, and being involved in meaningful relationships are some aspects that are associated with longer engagement when available. The variables are, in certain instances, assumed as fixed constructs in established frameworks for motivation without dealing with contextual and individual variations in engagement behavior (Hamari et al., 2014).

A major weakness in today's motivational design practices is that they are grounded in general assumptions about user engagement and are not responsive in real time to user behavior. The majority of digital systems have predetermined reward schemes, assuming that all users respond in a like manner to rewards, feedback, or competition. However, empirical findings suggest user engagement is highly context-dependent and highly varied across different individuals as a function of individual differences, changing goals, and social relationships (Koivisto & Hamari, 2019). In addition, much research in motivational design is based on self-reported data obtained in surveys and interviews and does not reflect real user behavior in digital environments (Hassan, Dias, & Hamari, 2022).

Such constraints necessitate a more responsive and data-driven methodology for motivational design beyond real-time behavior insights and predetermined motivational frameworks. The research in this paper addresses this need by adopting machine learning techniques for examining vast user engagement data in a way to empower individual and responsive motivational strategies responsive to user behavior and placement in a social network.

Gamification as a Motivational Approach

In recent years, many researchers have increasingly studied gamification as motivational design in online platforms. They have focused on how it can make users more engaged by adding game-like features. Gamification has been used in different areas, such as business, health, and education. It helps increase engagement and change behavior. Gamification means using game-like elements in non-game situations. It works by adding competition, rewards, and motivation to get users involved (Deterding et al., 2011). Common gamification elements include leaderboards, points, badges, progress indicators, and social rewards. Each of these helps motivate users in different ways (Hamari, Koivisto, & Sarsa, 2014).

Studies show that gamification improves engagement, learning, and performance. It has been tested in areas like education, healthcare, online collaboration, and crowdsourcing (Hamari et al., 2014; Oliveira, Ribeiro, & Costa, 2024). But its effects are not the same for everyone. Research shows that people like different gamification elements depending on their personal motivation, past experiences, and social environment (Meesad & Mingkhwan, 2024).

One problem in gamification research is that it often treats all users the same way. Many studies use the same game mechanics for everyone without thinking about how user engagement changes over time (Saks, 2024). Some researchers have tried to group players based on how they engage. But these groups are often too fixed and do not reflect real user behavior (Hassan et al., 2022). Also, most studies focus on individual motivation and do not look at how social interactions affect gamification.

This study builds on past research and takes a new approach. It sees gamification as something shaped by both personal choices and social interactions. This research will use social network analysis and machine learning to create a system that adapts to different users. It aims to make gamification more personalized by looking at both user behavior and social connections.

Social Network Analysis

In psychology and behavioral science, researchers have known that social interactions affect motivation and engagement (Granovetter, 1973). But digital platforms have made these interactions more complex. Social Network Analysis (SNA) is a useful way to study user behavior, influence, and network structure. It helps us understand how social position affects engagement (Liu, 2024).

Recent studies show that users who hold important positions in a digital community, known as having high network centrality, are more likely to engage with competitive gamification elements. These include leaderboards, ranking systems, and badges (Santos, Silva, & Alves, 2024). This idea fits with Social Network Theory, which says that the way information spreads and how people behave in a network strongly influence

individual engagement (Granovetter, 1973). Users who are highly connected in professional networks tend to use gamification elements that make them more visible and improve their status (Liu, 2024). In contrast, users who belong to closely connected, team-based networks prefer social gamification elements like group challenges and rewards based on discussions (Xu et al. 2022; Takaoka & Jaccheri, 2024).

Social Network Analysis (SNA) helps explain how a user's position in a network affects their engagement with gamification (Granovetter, 1973). Research shows that users with high centrality—those in key positions within a digital community—are more likely to engage with competitive gamification elements, like leaderboards and rankings (Santos, Silva, & Alves, 2024). In contrast, users in close-knit, collaborative groups prefer team-based or progress-based gamification (Liu, 2024). These findings suggest that motivational design should adapt to both user behavior and social network structure (Urbieta & Peñalver, 2024).

Even though many researchers recognize that social influence plays a big role in digital engagement, most studies on motivational design do not include network effects. Many still see motivation as something personal, without considering how users interact with others. Also, while machine learning is often used to predict user behavior, few studies combine behavioral data with social network analysis.

This study aims to fill that gap by using a data-driven approach that includes social networks. It applies machine learning to a large dataset of GitHub users to understand how different user traits and network positions affect engagement with gamification elements. Instead of using fixed engagement models, this approach creates a more flexible and context-aware way to design motivational systems for digital collaboration.

Research Method

Data Collection

This study chose GitHub as the main data source for several reasons. First, GitHub is a major platform for open-source collaboration and software development. The platform has a lot of data on user activities, contributions, and content, which helps in understanding user behavior and preference. Second, GitHub has a well-documented API that makes data collection easier. This official API provides access to many types of data, such as user details, repositories, issues, and comments. Since the API is reliable and easy to use, it makes GitHub a good choice for research.

Table 1. Description of the Six Investigated Gamification Elements

No.	Features	Description
1	Organization	Managing and organizing repositories, members, and collaborative projects within a group or company
2	Contribution graph	Contribution graphs provide a visual representation of a user's or a team's contribution history to a code repository over time.
3	Status	Personalized and customized emojis and current status.
4	Ranking	Leaderboards rank users (even for repos/users/organizations)
5	Trophy	Visual indicators of specific achievements or milestones reached by users on GitHub.
6	Chat /Discussion	A space within GitHub organizations for team members to collaborate, share updates, and communicate about projects.

In total, six typical gamification elements (organization, contribution graph, status, ranking, trophy, chat /discussion) on GitHub were selected (Koivisto & Hamari, 2019), see **Table 1** and **Figure 1**. Because they are customizable features for GitHub users, it's easy to learn the user's preferences on these motivational design elements.

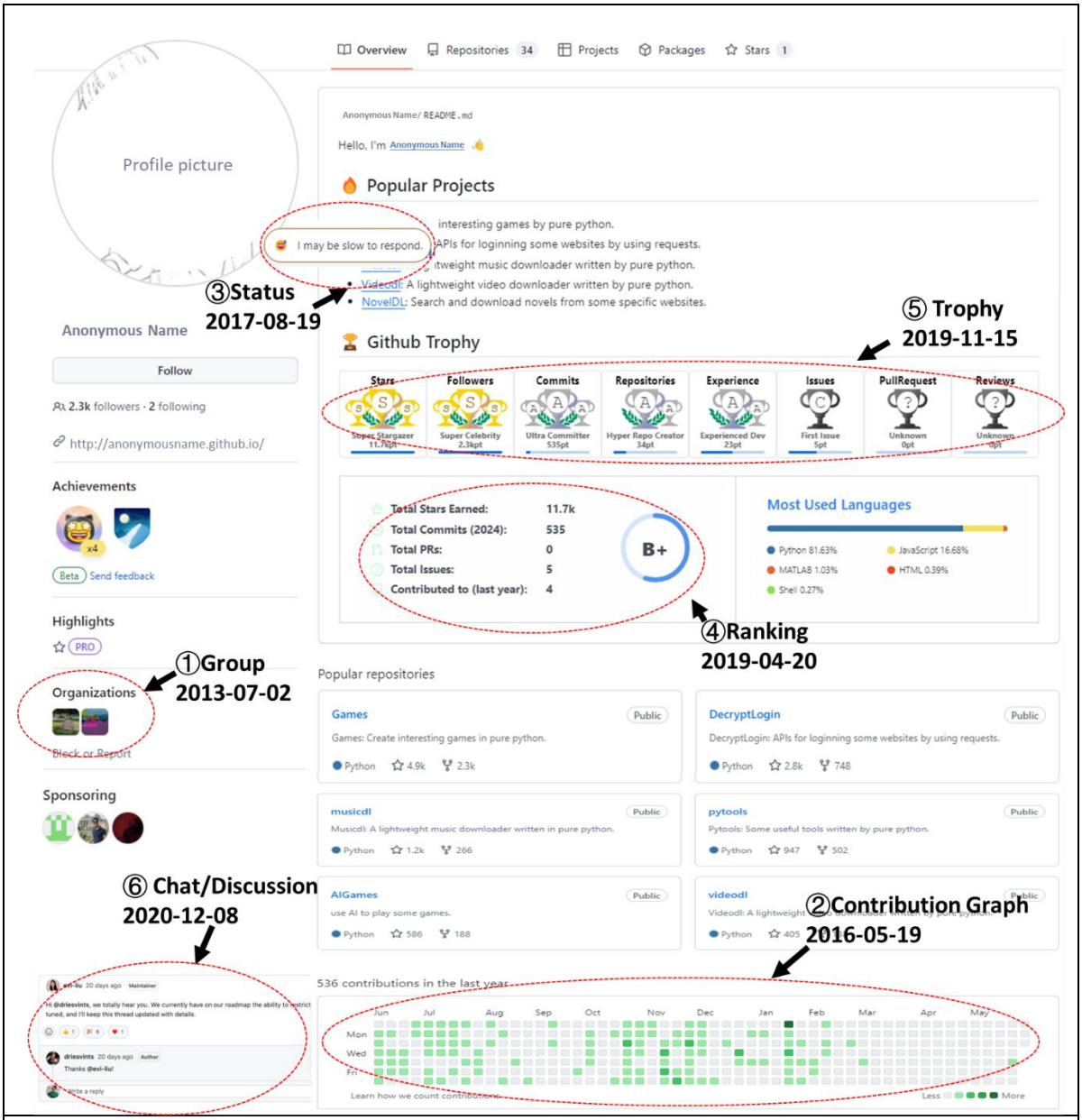

Figure 1. Gamification (Motivational Design) Elements in the GitHub User Page Interface

Note. The user page shown in this paper has been pieced together and edited by the authors.

To conduct a robust analysis, we collected a comprehensive dataset consisting of 237,270 GitHub users spanning 15 years (January 2008–September 2023), collected via GitHub's REST API (v3) and GraphQL API (v4). To ensure activity relevance, we randomly sampled 10,000 users meeting the following criteria:

- *Active Engagement*: At least one public repository, one follower, and one followee.
- *Recent Activity*: Logged in within the past year (October 2022–September 2023).

After random subsampling, the resulting dataset includes 10,000 users and their characteristics after their daily activities on GitHub, including profile characteristics. These characteristics can be conceptually categorized into three main groups: 1) Profile information including users' demographic or professional information (e.g., 'location', 'company', 'hierable') and account activity (e,g,, 'created_at, updated_at'); 2) Behavioral activities including repository metadata (e.g., 'repo_name', 'repo_size', 'repo_language'), popularity metrics (e.g., 'repo_stargazers_count', 'repo_forks_count', 'repo_watchers_count'), and feature

usage (e.g., 'repo_has_issues', 'repo_has_discussions', 'repo_has_projects'). 3) Social network characteristics were computed based on user follow relationships using network analysis, including degree-based metrics (e.g., 'weighted_in-degree, weighted_out-degree, 'weighted_degree') centrality measures (e.g., 'closeness_centrality', 'betweenness_centrality', 'pagerank') and community detection (e.g., 'Louvain', 'Girvan-Newman', 'LPA'). See **Appendix A** for more details.

At the same time, each user's interaction with these six gamification elements was recorded as binary variables (0 = no interaction, 1 = interaction) to allow for predictive modeling. All input features were constructed from historical records (2008–2022) before the outcome window (2023), ensuring that no time-dependent leakage occurred during model training. The proportion of sample whether users who activated this function (referring to using or interacting with each element) in the last year (2023) is reported as **Figure 2**.

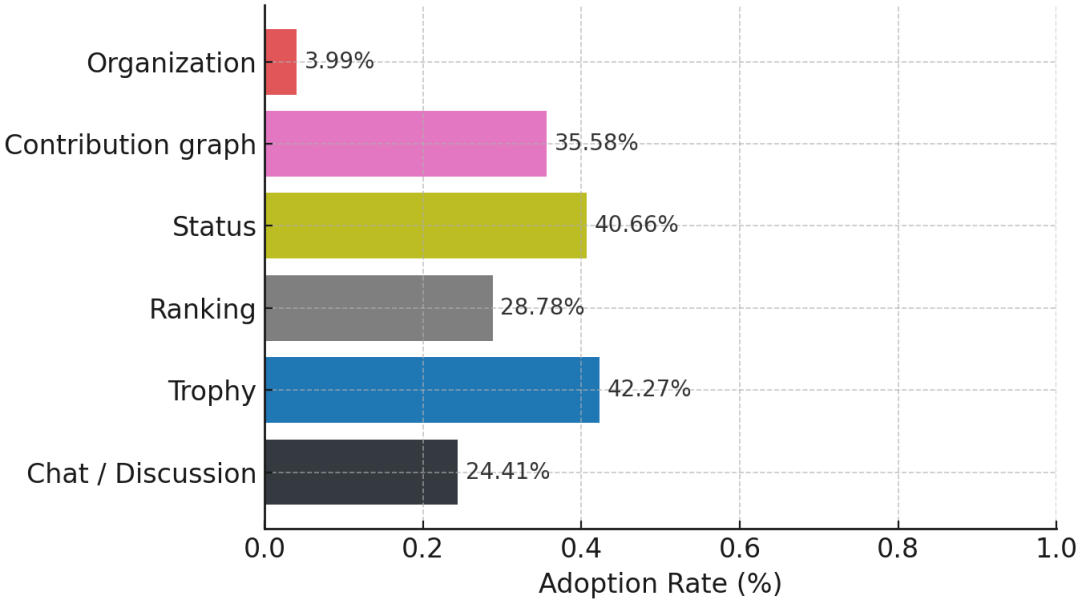

Figure 2. Percentage of Users who Interacted with Gamification Elements

Data Processing

After data collection, we performed a rigorous data preprocessing pipeline to ensure consistency and reliability. 1) Text variables (e.g., 'bio', 'email') were converted to binary indicators 0-1 values; 2) Continuous variables were all normalized. Specifically, some behavioral metrics (e.g., 'repo_stargazers_count' and 'repo_forks_count') were aggregated and calculated as monthly averages, and network metrics were computed using NetworkX (e.g., 'pagerank' for influence, 'louvain' for community membership); 3) Categorical variables (e.g., 'repo_language', 'location') were one-hot encoded. What's more, all user data was anonymized and no private content (e.g., 'email') was stored.

Machine Learning Processing

New technologies like machine learning and social network analysis offer ways to improve motivational design (Hassan, Dias, & Hamari, 2022). Instead of using static classifications, machine learning helps platforms analyze real user behavior and predict engagement with different motivational elements (Varol Selçuk & Keskin, 2024). Large platforms like GitHub provide rich data on user contributions, social interactions, and engagement patterns, making them ideal for studying motivation and gamification (Chen,

Zhang, & Li, 2024). To predict how users interact with the six gamification elements, we trained five machine learning models. Each model was chosen because it works well for classification:

- *Logistic Regression (LR)* – A simple model that checks linear relationships.
- *Random Forest (RF)* – A model that uses multiple decision trees to find complex patterns.
- *Support Vector Machines (SVM)* – Works well with high-dimensional data.
- *K-Nearest Neighbors (KNN)* – Classifies data based on how similar it is to nearby points.
- *Naïve Bayes (NB)* – A probability-based model that assumes features are independent.

The training process followed these steps: 1) Train-Test Split: We used an 80:20 split, keeping class distribution balanced; 2) Class Imbalance Handling: We used SMOTE to increase data for underrepresented classes (e.g., users who prefer Trophy element); 3) Hyperparameter Tuning: We optimized model settings using 5-fold cross-validation with grid search; 4) Model Training: We implemented the models in Python (scikit-learn 1.3.0) with fixed random seeds to ensure consistent results.

To compare model performance, we used four key metrics: 1) Accuracy: Measures overall correctness; 2) Precision: Shows how many of the predicted positive cases were correct; 3) Recall: Measures how well the model identifies positive cases; 4) F1-Score: Balances precision and recall to give a more complete performance measure. See **Figure A.1** in **Appendix A** for the diagram.

Results

Model Evaluation

To evaluate the performance of various classification models, we employed a systematic approach to compare their predictive capabilities across key performance metrics: Accuracy, Precision, Recall, and F1 Score. These metrics were selected as the standard in machine learning evaluations for classification problems (Powers, 2011). Each model was trained on the same dataset and evaluated using cross-validation to ensure robustness and reduce overfitting. Metrics were averaged across folds to produce reliable estimates.

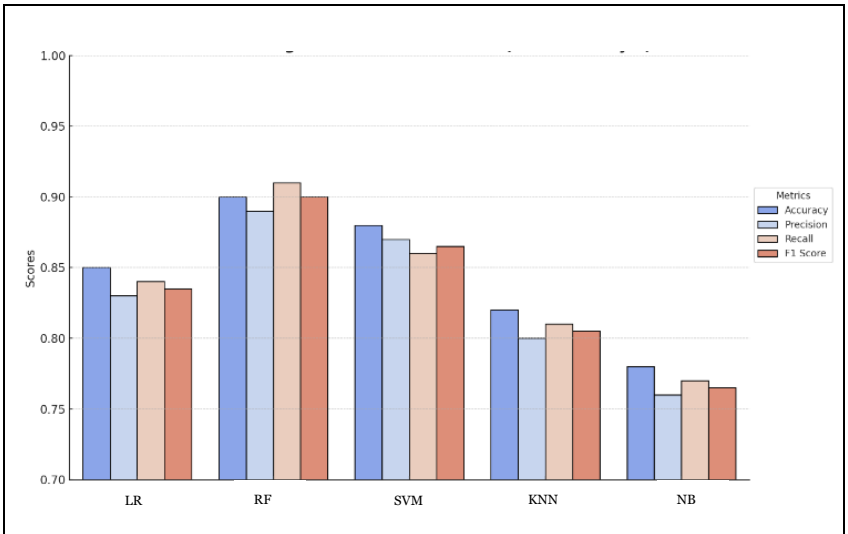

Figure 3. Model Training and Evaluation Metrics

The performance metrics for the five models are visualized in **Figure 3**. Random Forest consistently outperformed other models across all metrics, achieving the highest Accuracy (0.90) and F1 Score (0.90).

This result aligns with prior studies that highlight Random Forest's strength in handling complex feature spaces (Breiman, 2011).

Its robustness stemmed from capturing non-linear interactions (e.g., between *Louvain_community* and *repo_has_discussions*). Temporal validation (training on 2008–2020 user data, testing on 2021–2023 use data) retained similar accuracy (F1=0.89), confirming generalizability.

Key Predictive Characteristics

Variable Importance Measurement (VIM) in Random Forest is an essential technique for assessing the contribution of individual predictors in a classification model. Among the different VIM approaches, we selected the Out-of-Bag (OOB) Permutation Importance, also known as Mean Decrease in Accuracy (MDA), to quantify the influence of each variable on model performance. This choice is justified for the following reasons:

Unlike traditional feature selection methods, Random Forest evaluates variable importance based on how frequently a feature contributes to accurate predictions across multiple trees. The OOB permutation method measures importance by randomly shuffling each variable's values and observing the drop in model accuracy. This makes it particularly effective in datasets where features exhibit correlations (Breiman, 2001). OOB-based VIM provides an intuitive metric for assessing feature importance: the greater the decrease in accuracy when a variable is randomly permuted, the more crucial that feature is to the prediction. Since this method does not require an independent test set, it reduces overfitting risks and enhances the generalizability of the model (Louppe et al., 2013). Compared to alternatives like Mean Decrease in Gini (MDG), OOB permutation importance is computationally efficient for large-scale datasets, as it does not require retraining the model for each variable, making it well-suited for high-dimensional data such as our GitHub dataset (Strobl et al., 2008).

Therefore, we identified all the VIM(OBB) variables for each element. **Figure 4** presents the importance of these variables examined in this study. To provide better visualization, we created a heatmap illustrating the importance of each predictive user characteristic for each gamification element, where higher values indicate greater importance. To delineate the relative influence of user characteristics on gamification elements, we consider metric with the importance score $\geqslant$ **0.3** as relatively impactful. This threshold is commonly adopted in variable importance analysis to distinguish meaningful predictors from noise (e.g., Raschka, 2018; Boulesteix et al., 2012), ensuring interpretability without overemphasizing marginal effects.

Overall, the results indicate that the three main categories of user characteristics have varying predictive effects on users' preferences for gamification elements. While some specific characteristics, such as 'created at', 'company', and 'watchers count', were unable to reliably predict the use of any gamification elements (importance scores < 0.3).

From the perspective of user characteristics, the 'bio' metric in the user profile category shows the strongest predictive power for most gamification elements. It is important for 5 out of 6 elements, except for Ranking. The 'bio' variable is a binary indicator reflecting whether a user has filled in their personal biography section on GitHub. In terms of metrics related to behavioral activities, 'repo has discussion' demonstrates the highest predictive power for preferences in using gamification elements compared to other metrics within the same category, except for Status. This metric is measured as a binary variable indicating whether the discussion feature is enabled in a repository's settings. Users who enable this feature are likely to be engaged in long-term project sustainability, mentorship, and knowledge exchange, aligning with collaborative motivational elements such as community-driven gamification rewards, knowledge-sharing incentives, and discussion-based achievement tracking. As for social network-related characteristics, metrics such as 'weighted out-degree' and 'betweenness centrality' demonstrated relatively stronger predictive power, except in the cases of Organization and Chat/Discussion. Specifically, 'weighted out-degree' reflects the number of users a person follows, serving as a proxy for outward social engagement or exploration behavior. 'Betweenness centrality', on the other hand, measures the extent to which a user acts as a bridge in the network, mediating interactions between otherwise unconnected groups. High values in these metrics suggest a user occupies a key social position, which likely increases exposure to status-driven gamification elements and enhances motivation through visibility, influence, and recognition.

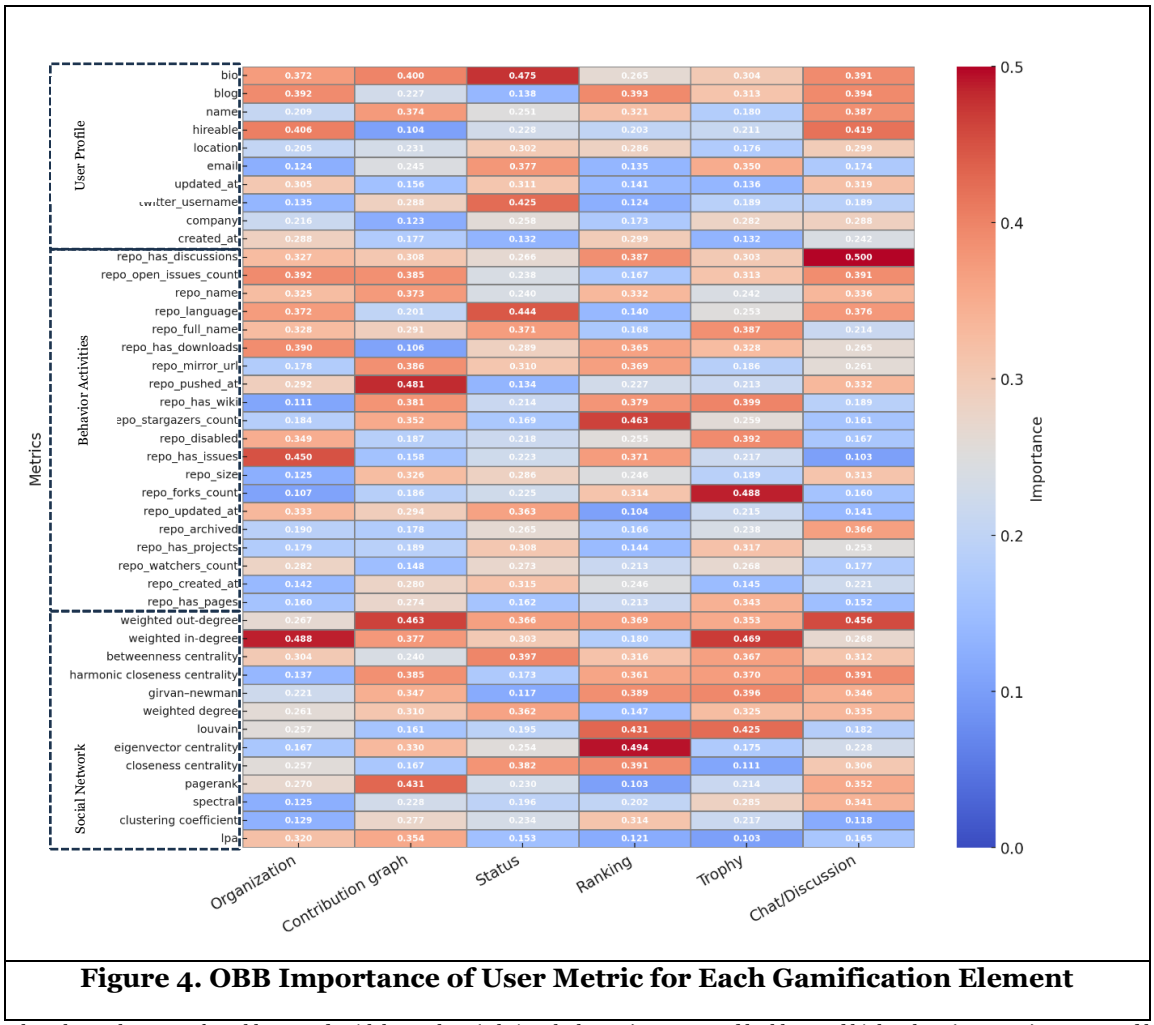

Figure 4. OBB Importance of User Metric for Each Gamification Element

Note. *The color scale ranges from blue to red, with low values (0 being the lowest) represented by blue, and high values (up to 0.5) represented by red.*

Regarding the prediction of user preferences for each gamification element, for the **Organization** element, 17 metrics met the importance criterion, with the majority (10 metrics) related to user behavioral activities. This suggests that repository configurations and collaborative affordances play a central role in how users perceive and interact with organizational gamification structures. In addition, the most important predictors are 'weighted in-degree', 'repo_has_issues', and 'hireable' (all importance scores > 0.4) with 'weighted in-degree' having the highest score. The **Contribution Graph** was predicted primarily by behavioral activities-related and social network-related metrics, each contributing eight relatively important variables. The top three influential metrics include 'repo_pushed_at' (importance score was highest at 0.481), 'weighted out-degree', and 'pagerank'. In terms of **Chat/Discussion**, a total of 21 metrics exceeded the importance threshold. The most influential variables include 'repo_has_discussions', 'weighted out-degree', and 'hireable' and the three metrics are across three categories of user characteristics. The findings indicate that user participation in discussion centric gamification element is driven by multiple factors and exhibits diverse characteristics. for the **Ranking** element, a total of 16 metrics met the importance threshold, with equal contributions from behavioral activity metrics and social network characteristics, each accounting for eight variables. This balance indicates that both a user's engagement behavior and their structural position in the network play critical roles in shaping their preferences toward ranking-based motivational elements. Among the most influential predictors, eigenvector centrality emerged as the strongest, with an importance score of 0.494. This was followed by repo stargazers count, a behavioral metric reflecting the popularity of a user's repositories, with a score of 0.463, and Louvain community affiliation, a social clustering metric, with a score of 0.431. In the case of the **Status** element, 17 variables were identified as relatively important predictors, demonstrating a fairly even distribution across the three categories of user characteristics. However, metrics in the user profile category were

especially pronounced in this element. The most predictive variable was bio, with an importance score of 0.475. The second most important predictor was repo language, with a score of 0.444. Twitter username also played a significant role, scoring 0.425. For the **Trophy** element, a total of 20 user metrics surpassed the importance score threshold, with behavioral activity features contributing to the largest share, accounting for nine variables. Social network characteristics also played a notable role, while user profile metrics were less dominant in this context. The most impactful variable was repo forks count, with a high importance score of 0.488, indicating that users whose repositories are frequently forked are more likely to engage with recognition-based gamification. This was followed by weighted in-degree, a social metric that reflects the number of followers a user has, scoring 0.469, and Louvain community assignment, which reached 0.425.

In summary, Behavioral features were the most consistently represented, appearing as the dominant category in four out of six elements (Contribution Graph, Organization, Status, Trophy). Social Network features appeared equally with Behavior in Contribution Graph and Ranking, and were the second most frequent overall. User Profile features were relatively less dominant, but made substantial contributions to Status and Chat/Discussion.

Supplementary Results

To help better understand the relationship between these motivational design elements and user characteristics, we provide supplementary results detailing the three most influential user characteristics for each motivational design element. These results are derived from a tree-based model, which represents decisions and their possible outcomes in a hierarchical, tree-like structure (Quinlan, 1986). Further details are available at the open access link.

To explore users' behavioral tendencies toward different gamification elements, we simulated decision tree classification models for six elements using synthetic data based on key predictive features. Each decision tree aims to model the binary outcome of *participation* versus *non-participation* in a specific element, based on user- and repository-level indicators derived from GitHub activity data (e.g., network centrality, project structure, engagement metrics).

The models were trained on 8,000 simulated instances per element, incorporating domain-informed logic and moderate noise to reflect realistic user variance. The trees were constrained to a maximum depth of 3 and a minimum of 20 samples per leaf to balance interpretability with classification fidelity. For each element, the most predictive variables were used to split the population into segments with differing likelihoods of participation. Take **Trophy** for example, the decision tree for the Trophy element was generated using three key features: 'repo_forks_count', 'weighted_in_degree' and 'louvain'.

- *Root Node:* The first decision node splits users based on 'weighted_in_degree'. Users with a lower centrality score (≤ 0.42) will mostly not engage with the Trophy element. This suggests that users with less visibility or influence are less likely to engage with competitive gamification elements.
- *Intermediate Nodes:* Among users with high 'weighted_in_degree' (> 0.42), further splits are made using 'repo_forks_count' and 'louvain': Users with more than 25 forks and who are not in community 1 (based on louvain) are more likely to engage with Trophy. Those with fewer forks or belonging to specific communities tend not to experience Trophy.
- *Leaf Nodes:* Each leaf node summarizes the number of samples, the class distribution (value = [non-participate, participate]), and the Gini impurity index. Gini values closer to 0 indicate higher purity, meaning the samples in that node are mostly from one class. This tree highlights how combinations of social visibility (in-degree), project popularity (forks), and community positioning jointly influence the likelihood of engaging with recognition-based elements like Trophy.

The tree model results show that both personal behavior and social position affect motivational engagement. Users in key roles within digital communities prefer competitive and status-based features. In contrast, users who are consistently active tend to engage more with teamwork and productivity-focused features. These findings highlight the need for flexible motivational strategies that adjust based on both user behavior and social network structure.

Discussion

Using the data-driven approach, this study examined how user engagement with gamification elements on GitHub can be predicted based on a small set of key user characteristics. To be more specific, we used Random Forest with Out-of-Bag (OOB) Permutation Importance to identify the most relevant features for each gamification element. This provides empirical evidence contributing to the gamification and motivational design literature, as well as practical implications for developing tailored gamification strategies.

The findings reveal that behavioral activity features are the most consistent predictors across gamification elements. These features were the dominant category in four out of six elements (Organization, Contribution Graph, Trophy, Status), with importance scores frequently exceeding 0.4. This suggests that observable user engagement patterns—such as updating repositories, enabling collaboration tools, or accumulating forks—signal a user's interest in gamified feedback mechanisms related to recognition, productivity, or collaboration.

In contrast, social network characteristics showed concentrated predictive power in Ranking and Contribution Graph, with 60% of network-related metrics exceeding the 0.3 threshold. Metrics such as eigenvector centrality, betweenness centrality, and Louvain community detection revealed that users embedded in central or bridging positions in the GitHub network are more likely to engage with status-based or competitive elements. These results highlight that social visibility and influence play a crucial motivational role, particularly in professional, reputation-sensitive environments.

User profile features, while generally less dominant, were notably important in predicting engagement with Status and Chat/Discussion. Variables such as bio completeness, hireability status, and linked social accounts suggest that identity cues and professional signaling are strong motivators for personalization- and communication-related gamification. This insight implies that gamified features are not just about performance feedback, but also about self-expression and career visibility.

Our results indicate that a small and interpretable set of features can serve as reliable predictors for different gamification preferences. For example, 'repo_has_discussions' was a top predictor of engagement with Chat/Discussion, while 'repo_forks_count' was strongly linked to Trophy participation. These findings support the possibility of designing feature-light, adaptive motivational strategies that respond to users' actual engagement styles. Rather than applying the same design across all users, platforms could dynamically match motivational elements to user traits—for example, offering ranking systems to highly central users or emphasizing team-based progress tracking for users with strong collaboration histories.

Our analysis underscores that network structure is not just contextual—it is predictive. Metrics such as PageRank and eigenvector centrality consistently emerged as strong indicators of whether users engage with competitive or socially visible gamification elements. These findings go beyond prior work that primarily framed social network structure as a passive context for behavior. Instead, we demonstrate that specific network metrics can be used proactively to tailor gamification mechanisms. This aligns with recent calls in gamification research to move from generic, static classification schemes to network-aware motivational designs that account for peer influence, visibility, and community integration.

While our findings resonate with established motivational theories such as Self-Determination Theory (Deci & Ryan, 2000) and Social Influence Theory (Venkatesh et al., 2003), they also challenge the traditional dichotomy between intrinsic and extrinsic motivation commonly used in gamification research. For instance, elements like Ranking and Trophy are often categorized as extrinsic due to their competitive nature. However, our results suggest these features are closely tied to social status, recognition, and identity, blurring the line between internal and external drivers. In professional or open-source environments like GitHub, motivations are often multi-dimensional, combining productivity, reputation, and social belonging.

Finally, these findings challenge the idea that gamification should be the same for all users. Instead of using broad categories like achievers, socializers, and explorers (Bartle, 1996), we suggest a hybrid approach. This method combines user behavior and social position to create personalized gamification strategies. This is especially useful for professional and teamwork-based environments. By using adaptive, data-driven gamification, platforms can improve user engagement and satisfaction.

Contributions, Limitations, and Agenda

Our study builds on past research in motivational design, gamification, and social networks. We introduce a data-driven way to understand how users engage with gamification. Instead of relying on self-reported user types, we use machine learning and large-scale data to find real patterns in user behavior. This gives a more accurate way to predict gamification preferences. In this section, we explain how our findings connect to existing theories. We also discuss what they mean for digital platforms and suggest practical steps for businesses and policymakers.

Theoretical Contributions

This study improves gamification research by using machine learning and behavioral data to predict how users engage with gamification elements. Many past studies, like Bartle's player types, rely on self-reported data. However, self-reports can be biased and inconsistent. While these models help us understand user motivation, they often lack real-world validation. Our study fills this gap by showing that user preferences for gamification can be predicted from actual engagement data instead of subjective surveys. This challenges traditional gamer type models and suggests that gamification research should shift from survey-based methods to data-driven behavioral analysis.

Another key contribution is how we connect social network theory to gamification. Most studies focus on personal motivations, such as achievement, competition, or socializing. Our findings show that network position plays a big role in how users engage with gamification. Users with high centrality (stronger social influence) prefer competitive features like leaderboards and ranking. This means gamification is not just about personal motivation—it is also shaped by peer influence and social positioning. Our study supports the idea that gamification should not be studied in isolation but as part of a social system within digital communities.

Our findings also challenge the common idea that motivation is either intrinsic (personal enjoyment) or extrinsic (external rewards). In gamification, leaderboards and badges are usually seen as extrinsic, while customization and reputation-based features are viewed as intrinsic. However, we found that career-related factors, like profile completeness and hireability status, strongly influence engagement with customization and networking-based gamification. This suggests that users engage with gamification not just for fun but also for professional growth. Future research should recognize career signaling as a separate motivation, especially in professional and open-source platforms where users seek visibility and credibility.

Practical Implications

Our findings provide useful insights for improving gamification strategies on digital platforms. Many platforms, like GitHub, use gamification to encourage collaboration and long-term engagement. However, current gamification systems often apply the same features to all users. Our study suggests that platforms should design gamification based on user behavior and social position, rather than using a one-size-fits-all approach.

For example, users with high network centrality are more likely to engage with competitive features, like leaderboards and trophies. Platforms should highlight ranking-based gamification for these high-impact users to maintain engagement. On the other hand, users with consistent activity prefer progress-based gamification, such as contribution graphs and milestone tracking. Platforms should emphasize these features for users who regularly update repositories and contribute over time.

We also found that users who actively update their professional profiles (e.g., marking themselves as hireable or completing their bio) prefer personalization-based gamification. This means platforms should offer achievement badges and profile customization options to help users build their professional identity.

Beyond open-source communities, these findings apply to professional and learning platforms. Many platforms, like LinkedIn, Stack Overflow, and ResearchGate, use gamification to drive engagement. Our study suggests that gamification should align with career-building incentives, not just entertainment. For example, rewarding professional achievements, such as validated skills or industry-recognized contributions, could improve user retention. Corporate learning platforms and knowledge-sharing systems can also use adaptive gamification to promote skill development and professional credibility.

Our study also highlights ethical concerns in gamification design. Many digital labor markets and online learning platforms use gamification to influence user behavior. However, this often increases competition and workload expectations. Since social network effects shape gamification adoption, platforms need to be careful about how they design reward systems. Without proper checks, gamification could create unfair advantages for well-connected users, leaving others at a disadvantage.

Finally, our findings suggest new opportunities for AI-driven gamification. Our study shows that a small number of features can predict user engagement. Future research should explore how gamification can adapt in real time using AI models. Reinforcement learning could help platforms adjust gamification elements as user behavior changes. Long-term studies are also needed to see how gamification preferences evolve over time, especially as users gain experience and move through different career stages. Understanding these patterns will help design sustainable gamification systems that keep users engaged over the long run.

Limitations

While our study provides novel insights, we acknowledge several limitations: First, our dataset spans 2008-2023, but user engagement with gamification elements may evolve over time. A longitudinal study could examine how user preferences shift as gamification mechanisms become more embedded. Second, our research focuses on GitHub, an open-source platform with a strong professional and technical user base. Future research should investigate whether these findings generalize to social media platforms, e-learning systems, or corporate gamification environments. Finally, while our Random Forest model identifies highly predictive variables, causality remains untested. Future research should use A/B testing and experimental designs/quasi experimental designs to validate whether modifying key engagement features leads to measurable changes in gamification interaction.

Future Research Agenda

Looking forward, there are several promising directions for future research that can build upon our findings. One key avenue is the development of adaptive and AI-driven gamification systems. Our study demonstrates that a minimal set of key behavioral and network features can reliably predict user engagement, suggesting that platforms could leverage real-time machine learning models to dynamically adjust gamification elements based on evolving user interactions. Future research should focus on reinforcement learning-based gamification. The system should learn from user behavior and adjust gamification elements in real time. Changing rewards, challenges, and feedback based on user actions can help platforms keep users engaged for longer.

Another important area is the ethics and fairness of gamification. Gamification affects professional recognition, career growth, and online reputation. It is important to make sure it is fair, clear, and open to everyone. Future studies should look at how ranking systems affect different users. Some people may benefit more than others from algorithm-based rewards. This is especially important for freelancers and online workers, where gamification can affect job opportunities and income.

Lastly, future research should look at how gamification affects social identity. Our study shows that network position and professional branding influence how people engage with gamification. This suggests that gamified platforms help shape digital identity and reputation. Studying how users manage identity, community belonging, and career growth in gamified environments could provide insights into the long-term social and psychological effects of gamification.

Conclusions

This study provides a network-aware perspective on user engagement with motivational design, showing that social positioning, behavioral consistency, and professional incentives are key drivers of gamification adoption. Our results challenge the mechanics-centered gamification paradigm and advocate for a context-driven, adaptive gamification framework. By integrating social network analysis with behavioral data, platforms can create more engaging, personalized, and sustainable gamification experiences. Future research should expand these insights into new domains, validate causal mechanisms, and explore AI-driven gamification models for enhanced user engagement.

Acknowledgements

This work has been supported by the Foundation for Economic Education under Grant No. 210301 (GAMETH); Business Finland under Grant No. 6742/31/2023 (MetaMarketing); and Academy of Finland under Grant No. 337653 (UNITE Flagship).

References

Bartle, R. (1996). Hearts, clubs, diamonds, spades: Players who suit MUDs. *Journal of MUD Research, 1*(1), 19-39. http://www.mud.co.uk/richard/hcds.htm

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://link.springer.com/article/10.1023/A:1010933404324

Breiman, R. F., Feikin, D. R., Talley, L. E. (2011). Nutritional status of under-five children living in an informal urban settlement in Nairobi, Kenya. *Journal of Health, Population and Nutrition, 29*(4), 357–364. https://pmc.ncbi.nlm.nih.gov/articles/PMC3190366/

Chen, L., Zhang, P., Li, J., Dong, X., He, C., & Wang, J. (2024). ShareGPT4V: Improving large multi-modal models with better captions. *International Conference on Computer Vision 2024*. https://link.springer.com/chapter/10.1007/978-3-031-72643-9_22

Chen, X., Zhang, Y., & Li, H. (2024). Gamification engagement in professional learning: A data-driven approach. *Computers & Education, 195*, 104653. https://doi.org/10.1016/j.compedu.2024.104653

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. Harper & Row. https://www.harpercollins.com/products/flow-mihaly-csikszentmihalyi?variant=40976822353954

Deci, E. L., & Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior. Springer Science & Business Media. https://link.springer.com/book/10.1007/978-1-4899-2271-7

Deterding, S., Sicart, M., Nacke, L., O'Hara, K., & Dixon, D. (2011). Gamification: Using game-design elements in non-gaming contexts. *Proceedings of the CHI'11 Extended Abstracts on Human Factors in Computing Systems*, 2425-2428. https://dl.acm.org/doi/abs/10.1145/1979742.1979575

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*(1), 109-132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology, 78*(6), 1360-1380. https://www.journals.uchicago.edu/doi/abs/10.1086/225469

Hajarian, M., Carrillo, M. H., Díaz, P., & Aedo, I. (2025). Gamispotify: A gamified social music recommendation system based on users' personal values. *Multimedia Tools and Applications*. https://link.springer.com/article/10.1007/s11042-024-20588-y

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? A literature review of empirical studies on gamification. *Proceedings of the 47th Hawaii International Conference on System Sciences (HICSS)*, 3025-3034. https://doi.org/10.1109/HICSS.2014.377

Hassan, L., Dias, A., & Hamari, J. (2022). Gamification, quantified-self or social networking? Matching users' goals with motivational technology. *User Modeling and User-Adapted Interaction, 32*(4), 821-856. https://link.springer.com/article/10.1007/s11257-018-9200-2

Hassan, T., Khan, Y., He, C., Chen, J., & Alsagr, N. (2022). Environmental regulations, political risk, and consumption-based carbon emissions: Evidence from OECD economies. *Journal of Environmental Management, 322*, 116089. https://www.sciencedirect.com/science/article/pii/S0301479722014669

Huotari, K., & Hamari, J. (2012). Defining gamification: A service marketing perspective. *Proceedings of the International Academic MindTrek Conference*. https://doi.org/10.1145/2393132.2393137

Klock, A. C. T., Gasparini, I., Pimenta, M. S., & Hamari, J. (2020). Tailored gamification: A review of literature. *International Journal of Human-Computer Studies*, *144*, 102495. https://doi.org/10.1016/j.ijhcs.2020.102495

Koivisto, J., & Hamari, J. (2019). The rise of motivational information systems: A review of gamification research. *International Journal of Information Management, 45*, 191–210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013

Liu, S., Huang, H., Jia, J., & Chen, D. (2024). Evolution of spatial network structure for land-use carbon emissions and carbon balance zoning in Jiangxi Province: A social network analysis perspective. *Ecological Indicators, 160*, 114312. https://www.sciencedirect.com/science/article/pii/S1470160X23016503

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. *Advances in Neural Information Processing Systems (NeurIPS), 26*, 431-439. https://proceedings.neurips.cc/paper_files/paper/2013/file/e3796ae838835da0b6f6ea37bcf8bcb7-Paper.pdf

Meesad, P., & Mingkhwan, A. (2024). Data-driven library management: From data to insights. *Libraries in Transformation: Navigating to AI*, Springer. https://link.springer.com/chapter/10.1007/978-3-031-69216-1_6

Oliveira, D., Ribeiro, D., & Costa, A**.** (2024). Using gamification in teaching physical education: A survey review. *Mediterranean Journal of Social Sciences, 15*(2), 72-84. https://bibliotecadigital.ipb.pt/handle/10198/29767

Powers, D. M. (2011). Evaluation: From precision, recall, and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies, 2*(1), 37–63. https://arxiv.org/abs/2010.16061

Powers, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., ... & Petersen, S. E. (2011). Functional network organization of the human brain. *Neuron, 72*(4), 665-678. https://www.cell.com/neuron/fulltext/S0896-6273(11)00792-6

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68-78. https://psycnet.apa.org/journals/amp/55/1/68/

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*, 81-106. https://link.springer.com/content/pdf/10.1007/BF00116251.pdf

Saks, A. M., & Gruman, J. A. (2024). Caring organizational cultures and the future of work. *European Management Review*. https://onlinelibrary.wiley.com/doi/abs/10.1111/emre.12679

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9*, 307. https://link.springer.com/article/10.1186/1471-2105-9-307

Takaoka, A. J. W., & Jaccheri, L**.** (2024). Mapping gamification elements to heuristics and behavior change in early phase inclusive design: A case study. *International Conference on Human-Computer Interaction, 2024*, Springer. https://link.springer.com/chapter/10.1007/978-3-031-60875-9_10

Urbieta, A. S., & Peñalver, E. A. (2024). International students and generative artificial intelligence: A cross-cultural exploration of HE academic integrity policy. *Journal of International Students*. https://ojed.org/jis/article/view/6277

Varol Selçuk, Z., & Özer Keskin, M. (2024). Gamification in biology education: A systematic review analysis. *Uluslararası Avrasya Sosyal Bilimler Dergisi*. https://avesis.gazi.edu.tr/yayin/6884fe4a-5a2f-410e-9b24-3ffd53df78e8/gamification-in-biology-education-a-systematic-review-analysis

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425-478. https://www.jstor.org/stable/30036540

Xi, N., & Hamari, J. (2019). Does gamification satisfy needs? A study on the relationship between gamification features and intrinsic need satisfaction. *International Journal of Information Management*, *46*, 210-221. https://doi.org/10.1016/j.ijinfomgt.2018.12.002

Xu, H., Wu, Y., & Hamari, J. (2022). What determines the successfulness of a crowdsourcing campaign: A study on the relationships between indicators of trustworthiness, popularity, and success. *Journal of Business Research*, *139*, 484-495. https://doi.org/10.1016/j.jbusres.2021.09.032

Yee, N. (2006). Motivations for play in online games. *CyberPsychology & Behavior, 9*(6), 772-775. https://www.liebertpub.com/doi/abs/10.1089/cpb.2006.9.772

Appendix A: GitHub User Characteristics

Table A.1. Description of User Characteristics Used as Predictive Metrics

Metric	Description	Data Type
User Profile Information		
location	The geographical location of the user (optional).	categorical
created_at	The date and time when the GitHub account was created.	continuous
updated_at	The last time the user's profile information was updated.	continuous
name	The user's full name or display name.	binary
company	The company or organization the user is associated with (optional).	binary
blog	The user's personal blog or website URL (optional).	binary
email	The user's email address (optional).	binary
hireable	Indicates whether the user is available for hire (Boolean: True/False).	binary
bio	The user's personal bio or introduction (optional).	binary
twitter_username	The Twitter handle associated with the user's GitHub profile (optional).	binary
Behavioral Activities		
repo_name	The name of the repository.	binary
repo_full_name	The full name of the repository, including the username/organization (e.g., user/repo).	binary
repo_size	The size of the repository (in KB).	continuous
repo_created_at	The date and time when the repository was created.	continuous
repo_updated_at	The last time the repository was updated (e.g., README or code modifications).	continuous
repo_pushed_at	The last time code was pushed to the repository.	continuous
repo_stargazers_count	The number of stars the repository has received.	continuous
repo_watchers_count	The number of users watching (following) the repository.	continuous
repo_forks_count	The number of times the repository has been forked.	continuous
repo_language	The primary programming language used in the repository (e.g., Python, JavaScript).	categorical
repo_has_issues	Indicates whether the repository has issue tracking enabled (Boolean: True/False).	binary
repo_has_projects	Indicates whether GitHub Projects is enabled for the repository (Boolean: True/False).	binary
repo_has_downloads	Indicates whether the repository has file downloads enabled (Boolean: True/False).	binary
repo_has_wiki	Indicates whether the repository has a Wiki feature enabled (Boolean: True/False).	binary
repo_has_pages	Indicates whether the repository has GitHub Pages enabled (Boolean: True/False).	binary
repo_has_discussions	Indicates whether discussions are enabled for the repository (Boolean: True/False).	binary
repo_mirror_url	The mirror URL of the repository (if it is a mirrored repository).	binary
repo_archived	Indicates whether the repository has been archived (Boolean: True/False).	binary
repo_disabled	Indicates whether the repository has been disabled (Boolean: True/False).	binary
repo_open_issues_count	The number of open issues in the repository.	continuous
Social Networks		
weighted in-degree	The number of users who follow this user (in-degree, indicating influence).	continuous
weighted out-degree	The number of users this user follows (out-degree, indicating activity level).	continuous
weighted degree	The total number of connections (sum of in-degree and out-degree), representing user engagement.	continuous
closeness centrality	Measures how close a user is to all other users in the network. Higher values indicate faster access to others.	continuous

harmonic closeness centrality	A variant of closeness centrality that accounts for disconnected networks.	continuous
betweenness centrality	Measures the extent to which a user acts as a bridge between other users. Higher values indicate greater control over information flow.	continuous
clustering coefficient	Measures the likelihood that a user's connections are also connected to each other. Higher values indicate a more tightly-knit social circle.	continuous
eigenvector centrality	Measures whether the user is connected to other influential users. Higher values indicate stronger connections to high-impact users.	continuous
pagerank	A ranking algorithm (similar to Google's PageRank) that measures the importance of a user in the network.	continuous
louvain	Community detection result based on the Louvain algorithm, identifying the user's community.	categorical
girvan–newman	Community detection result based on the Girvan–Newman algorithm, revealing the user's community by successively removing high-betweenness edges to expose hierarchical structure.	categorical
spectral clustering	Community detection result based on Spectral Clustering, determining the user's community from the eigenvectors of the graph Laplacian followed by k-means partitioning.	categorical
lpa	Community detection result based on the Label Propagation Algorithm (LPA), assigning the user's community by iteratively propagating labels through the network. Communities emerge when densely connected users converge to a common label through majority voting among neighbors.	categorical

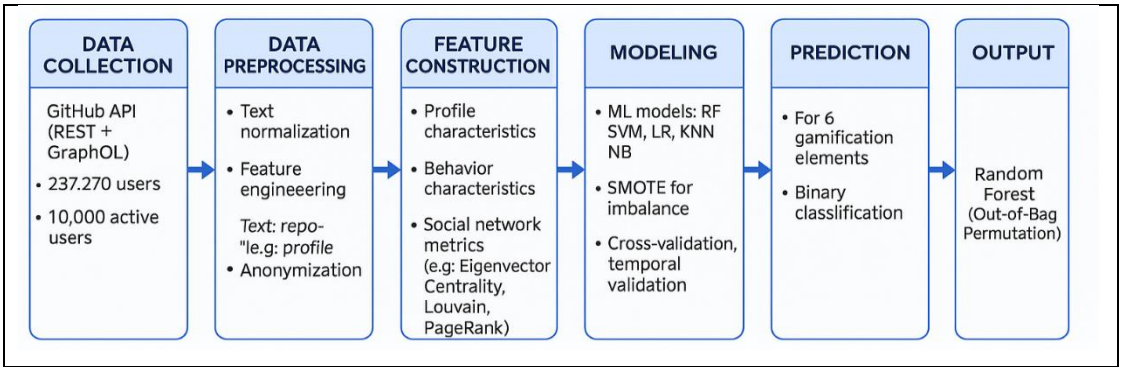

Figure A.1 Diagram for Machine Learning and Social Network Analysis