The Impact of Player Behavior Features on Game Performance Outcomes

Roni Roitbord, Saleh Hassan, Yagel Maimon
Technion – Israel Institute of Technology
CS Department

December 2024

Proposal

We want to identify and analyze the impact of player behavior features on game performance outcomes. Leveraging gaming datasets from Kaggle, this study aims to enhance decision-making processes in game design and player analytics by detecting key behavioral factors that influence player success. For example, checking the contribution of buying in-game purchases with cash for both the gameplay and the growth of a player, while the conclusion could be the adjustment of such packages and items for game developers to increase their profits.

1 Introduction

The gaming industry has grown into a significant sector, with multiplayer online games being a dominant force. Understanding the factors that influence player success is critical for designing balanced gameplay and fostering player engagement. This study aims to present the gap between raw data and actionable insights by identifying key behavioral features. Existing research highlights the importance of behavioral analytics in gaming. Techniques such as feature correlation, machine learning models, and dimensionality reduction are commonly used to understand player behaviors. For example, metrics like actions per minute (APM) and teamwork frequency have been correlated with higher win rates in eSports titles[1, 2].

2 Overview

In this project we analyze player behavior data to uncover which features have the most significant impact on game performance metrics such as win rate or ranking. By developing an algorithmic pipeline starting with a brute-force approach and optimizing it for efficiency, we will evaluate these insights using two or more gaming datasets. Deliverables include a detailed analysis of impactful features, computational performance metrics, and actionable recommendations for data scientists, gamers and game designers. The research is expected to provide valuable insights into player behavior and inform strategies for improving player satisfaction.

3 Research Question

Which player behavior features most significantly impact game performance outcomes?

4 Relevance/Impact

Identifying the factors that drive player success offers valuable opportunities for game developers to refine gameplay mechanics and enhance user engagement. Moreover, this research has broader social implications, including promoting diversity and fairness in competitive gaming and creating more enjoyable experiences for a wide range of players. The findings could inform the design of more effective matchmaking algorithms, fairer in-game economies, and gaming environments that support equitable participation for all player demographics which could make more profit for the game developers.

5 Methodology (Key steps)

In order to answer the research question, and find the most impactful features, we will follow these key steps:

1. Acquire datasets about player performance in games (from kaggle).
2. Clean the data and preprocess it by normalizing the data, handling missing values, and converting all features to a numerical format.
3. Explore the data to identify correlations between the features and performance metrics by running a causal DAG algorithm for instance.
4. Implement a ML based approach to predict the performance metrics based on the features with the highest correlation from the data exploration phase. And another brute-force approach which is based on all the features together as a baseline.
5. Test the model across datasets to ensure that it is generalizable.

6 Risk Analysis

6.1 Risk Of Selecting Bias

There are some potential biases that we want to handle so they won't affect our conclusions, those risks require having a fundamental knowledge base of how the game works:

1. Low winning rate due to old/cheap technology such as old phone technology (android) or low bandwidth.
2. Games per time frame (e.g. a player that plays 7 times a day and a player that plays once in 7 days will probably have different statistics)

3. Different timezones can present players with different skill sets and different requirements. for example, for a specific timezone, if the majority of players located in Asia, it can imply a harder difficulty. On the other hand, a geographic majority can also lead to different in-game purchases.

6.2 Risk Of Noisy Data

Since the data is collected automatically and didn't undergo checks by humans, some data may contain information that does not reflect reality, such as data of users that had some Network connection failures or other misleading information.

6.3 Risk Of Infeasible Calculations

Since there is thousands of user in clash royal The brute-force approach for feature analysis may result in high computational costs, especially given the size and dimensionality of the datasets.

7 Expected Outcome

1. Conclusions regarding the gameplay of a play, either to enhance the gameplay or the winning rate, show the contribution of features or hidden data on the player's gameplay

2. Provide financial advices for SuperCell which are the developers of the game, such advices that we found impactful on the data we investigated (e.g. timeframes and the exact time for special sales/deals, in-game purchase packs content and etc)

3. Similar to financial advices, we will try to provide strategic advices for the developing company and answer questions like what type should be the next developed character? and how to preserve the increase of new players while not decaying the income of the game to the company?

8 Research Plan

The research will proceed as follows:

1. Dataset Acquisition (Week 1-2): Identify and preprocess gaming datasets.

2. Baseline Algorithm Development (Weeks 3–4): Implement the brute-force approach.

3. Optimization and Testing (Weeks 5–8): Refine the chosen algorithms and validate our findings on multiple datasets.

4. Analysis and Reporting (Weeks 9–10): Interpret findings, generate visualizations, and prepare the final report.

9 Potential databases for research

Those are the databases we found regarding our research subject, please note that we won't use every database provided by any dataset:

9.1 Main Databases

The main data will be taken/sampled from the following datasets:

1. **Clash Royale S18 Ladder Datasets (37.9M matches)**:

   https://www.kaggle.com/datasets/bwandowando/clash-royale-season-18-dec-0320-dataset

2. **Clash Royale Battles (481M matches)**:

   https://www.kaggle.com/datasets/s1m0n38/clash-royale-games

Those databases will consist the primary data that we will investigate and conclude all of our outcomes based on those database/s. the main technical difference between those databases is that one has much more rows and less columns than the other one.

9.2 Side Databases

We might use the following databases to ease the work with the huge main databases:

1. **Clash Royale S18 Ladder Datasets for prediction**

   https://www.kaggle.com/datasets/tristanwassner/clash-royale-s18-ladder-datasets-for-prediction

9.3 Assisting Databases

We will use the following database to map and vectorize our tables:

1. **Clash Royal Dataset**

   https://www.kaggle.com/datasets/abhinavshaw09/clash-royal-dataset

References

[1] "The Analytics Behind Esports: Data-Driven Strategies in Competitive Gaming". In: (). URL: https://www.statology.org/the-analytics-behind-esports-data-driven-strategies-in-competitive-gaming/.

[2] "The structure of performance and training in esports". In: (). URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0237584.