Challenge Player Model v0.1

Updated: 09/17/2022

# Purpose

* Design metrics to quantify a Challenger’s skill at The Challenge.
* Use those metrics to design a prediction model for events in The Challenge.
  + Events are any of the games played: daily challenges, eliminations, finals, and total wins.

## Dataset Used

1. All challengers who have appeared on the show.
2. Only the flagship show was considered, can add the other data later.
3. Removed first 6 seasons and The Island.
   1. The first 6 seasons did not have eliminations and had inconsistent formats.
   2. The Island also had an odd format and few players, removed for simplicity.

This results in:

* 307 Challengers
* 29 Seasons
* 29 Final Challenges
* 384 Eliminations
* 422 Daily Challenges
* 91 Challengers Won
* 101 Challengers Placed in Final
* 104 Quitters

## Data Collection

Several sources are used, including main Wikipedia article, the challenge fan wikipedia, and previously compiled datasets. The data was broken into several csv files:

* master reference file containing total raw stats
* various detailed matrices of every
  + daily, who participated, and who won
  + elimination, who participated, who won, who lost, and any outliers (DQ, quit, injury, mercenaries, etc.)
  + final, who won, who placed and in what place, and any outliers (DQ, quit, injury, etc.)

# Approach and Background

Once the data was collected, We had to consider what stats were useful and normalize them across the Challengers.

The five major categories in the show were a good place to start:

|  |  |
| --- | --- |
| Category | Status |
| Daily Challenges | Complete |
| Eliminations | Complete |
| Finals | In Development |
| Political Game | Investigation |
| One-Off Modifiers | Investigation |

Initial principles for creating metrics:

1. Dailies
   1. Dailies are largely unimportant for raw data
   2. Really only important if you come in either first or last. And only in seasons where immunity is given to the winner and the loser is set directly to elimination.
2. Eliminations
   1. Eliminations are the most directly impactful, as the remove players from the Challenge.
   2. Winning/Losing eliminations can also impact being voted in, either being perceived as either a layup or strong in eliminations. Potential Intimidation score.
3. Finals
   1. Winning is best, of course, but even getting to a final should positively impact a challenger’s score.
   2. A point scale for placing in a file would work, with winning worth the most.
4. Politics
   1. Voting seems complex, but what we are really concerned with are if a player is voted into elimination as a binary.
   2. Getting voted in shows either poor political skill, being disliked and/or being perceived as a strong Challenger who needs to go home.
   3. Not getting voted in can be considered a neither good or bad.
5. One-Off Modifiers
   1. Small modifiers that can change the game is observable ways.
   2. Things like having best friends or romantic relationships can keep players in longer than their skill would allow on average.
   3. Violence and emotional breakdowns can remove players prematurely.
   4. Quitting and DQs are also considered.

# **Methods for each Metric**

## Daily Challenges

Resource:

* The master file for the results of all daily challenges was found attached to this youtube video: https://www.youtube.com/watch?v=9VDYjAvmxWI&ab\_channel=CaffeineConfessionals
* Initial statistics were described in this blog post, which helped me get started: https://thelavinyears.com/2021/02/23/challenge-domination-score-a-data-based-mtv-challenge-statistic/

Development:

Dailies were the most difficult to measure. It has the most variability across the three major game types, which include:

1. team size
2. number of teams
3. number of possible winners
4. chance of winning due to game type
5. Seasons where losers are thrown into elimination vs not
6. Type of game played (strength, intelligence, athletics) Though this can be hard to quantify.

From my research, the probability of winning a game was the best method for creating a Daily Score for each challenger. Generally, it is very hard to win dailies with a lot of players and small team sizes (compared to two teams competing).

A useful metric would account for how much better or worse a challenger is compared to a statistically average challenger. This is how we can calculate the **Wins Above Average.**

*eq. 1 Daily Score Calculation*

*Expected Wins (xW) = Winning Team Size / Total Players in Daily*

Wins Above Expected Wins (W/xW) = Total Wins(W) – Expected Wins (xW)

*Daily Score (DS) = (W/xW)/Total Dailies (tD)*

One note is the total number of challengers goes down as a season goes on, so each xW is slightly changed to reflect that.

Example:

CT

Total dailies = 164

Total wins = 54

xWins = 48.64

Daily Score: (54 – 48.64)/164 = 0.033

So, in a Daily challenge, CT is 3.3% more likely to win a challenge compared to random chance.

## Eliminations

Probably the easiest metric. Since the game is 1v1 (or 2v2 for pair games), on average, there is a 50/50 shot of being sent home between 2 average Challengers. DQs, injuries, and quitters are excluded from their score since we are only looking at what occurs in the actual event.

Unlike Dailies, Eliminations have the benefit of a standardized format, allowing us to use scoring methods developed for other 1v1 games like chess. Elo and Glicko have been used for rankings in chess, but a more recent tool has been developed for online multiplayer games, called TrueSkill.

It works for our purposes because

* it takes relatively few games to get a score for a challenger (n=3)
* tracks wins/losses over time and between challengers, giving more points for a lower scored challenger beating a higher scored one.
* Can be modified to work with 2v2 eliminations

The score is out of 50, with 25 being the average challenger score, and a variance value representing roughly the number of games played (confidence).

Example.

Jay Starrett has only 4 eliminations (2 win, 2 loss) but has a score of 28.99 and a variance of 4.68, being increased by beating CT (6 wins, 5 losses), whose score is similar 28.67 but has a lower variance 3.99. The lower variance is important to differentiate two challengers with similar scores to show one has played more and who’s score is less likely to change.

Another is Fessy (3 wins 0 losses), with a massive 40.5 score and 5.02 variance. Hes beat three very strong elimination challengers in Jordan, Kyle, and Nelson, but since hes only played 3 times, his variance is much higher.

Fig. 1 – TrueSkill comparison

Chart

Description automatically generated

* Mu equates to skill
* Sigma equates to variance of play

## Finals

With less data to work with, especially with wins, well need to get creative. We can start with some basics criteria:

1. Total wins over number of seasons
2. Total finals made by number of seasons (and place in that win)
3. DQs in final

We could get more granular with segmenting the finals and even seeing who was partnered up, but I don’t think that is necessary, because:

1. If you make it to a final and win, you probably are pretty good
2. If you make it and place second, you are also probably pretty good
3. If you DQ or come in last, you were probably a lay-up or not ready

Since final wins are binary, win or lose, it only helps your score if you have made it to one. Making it to a final is also a general stat boost, as you have made it to the end, which means you probably did a few things correctly.

## Voting/Political Game

Since you are safe if you don’t get eliminated, not being voted in is a huge invisible metric to getting to a final.

Generally, less votes overall means you were:

1. too intimidating to other Challengers.
2. have enough friends to make it through without going in.
3. played the game politically well enough to lower your chances of seeing an elimination.
4. Had friends willing to throw in on seasons where skulls are needed to qualify for the final. Ex. Lolo quit in Double agents because she believed she was unable to get enough votes to get sent down).

Other data we can consider are:

1. “Survival/Intimidation” Score: overall, how many times did the Challenger get voted-in to an elimination compared to the number of times they could have been.
   * 1. A low score would mean they either intimidate the other Challengers or are kept around as a lay-up.
2. “Beast” score: number of wins vs number of eliminations voted-in.
   * 1. A high score here would denote this person has a weak political game because they keep getting thrown in, but just demolish in elims.
     2. Could also be helpful for “stickiness” with rookies, who usually get thrown in until they lose.

## One-Of Modifiers

A list of other metrics we can use that are not directly quantified in the actual show. These can be small benefits or detriments that can help round out a player.

Here are a few I have come up with:

* Ride or Die:
  + having just one person who you know will never vote you can increase survival
* Romantic Relationship:
  + either just for a few episodes, or from outside the Challenge. Same as Ride or Die.
* Emotional Outbreaks:
  + non-violent or violent events where you go full Camilenator, but do not get removed from the game.
  + Changes how other players see you and can result in being voted-in and eliminated.
* Physical Violence (B.R.A.D Scale):
  + Belligerent Rowdy And Drunk.
  + Being removed from the game due to hitting someone.
* Injury Score:
  + Either by accident, getting sick, or getting too drunk, injuries are a death sentence in the Challenge.
  + Can result in a handicap going into events, or just being removed outright.
* Quitter:
  + Just quit the game. Can be a good reason (family emergency) or bad reason (just plain quit).
  + Ruling for this is based on TJ’s response, and his word is law.
* Rookie Status:
  + Usually, the first and second seasons Challengers are considered rookies.
  + survivability is low unless a rookie alliance can be established.
  + Making it to a final as a rookie could be considered a “layup” score.
* Legend Status:
  + Some players have a stigma to them, which can either hurt them (Ashley gets eliminated early in the later seasons) or help them (CT obviously).
  + Can be tracked by a combination of a lot of things, maybe an aggregate score?