Stat 143
Seth Billiau, Sarah Lucioni

Elo and Glicko-2 Rating Systems for 9-Ball Pool

May 18, 2021
Dasha Metropolitanksy

1 Background

Overview

9-Ball is a variant of pool in which two players attempt to pocket numbered balls in ascending order from 1 to 9. The winner of an individual game, also called a *rack*, is the player that pockets the 9 ball. A match involves two players racing to win a set number of racks.

Over the past thirty years, 9-Ball has emerged as the most popular game in the world of professional pool. Every year, there are around 10 to 20 major 9-Ball tournaments with lucrative prize pools. Some 9-Ball tournaments, like the mens and womens US Open Pool Championship, are open to professionals and amateurs alike. Similar to open tournaments in tennis and golf, however, the recipients of the tournament's $375,000 total payout are usually professionals. Other events are invitationals sanctioned by organizations like the World Pool-Billiard Association (WPA) and Matchroom Pool. In addition to stand-alone events, professional tour events like the Dynamic Billiard Euro Tour have also begun to emerge.

Nearly all tournaments end in single- or double-elimination bracket format. Round-robin play, group-stage matches, or pre-existing rankings typically determine bracket seeding. Once a bracket has been determined, players compete in matches in order to advance in the tournament.

Existing Rating Systems

The majority of existing rating systems are maintained by major pool organizations. Perhaps the most widely-accepted rating system in 9-Ball is maintained by the WPA, the international governing body for pocket billiards. WPA ratings have been used to determine qualification for major events like the World 9-Ball Championship. In this rating system, players earn points based on their performance in WPA-sanctioned events. The number of points earned depends on the event and the player's performance. Other large pool organizations like Matchroom Pool, Billiard Congress of America (BCA), and Euro Tour maintain similar ranking systems for their events.

These organization-backed rating systems are flawed for a number of reasons. First, organization-backed ratings only consider a player's performance in the events presented by that organization, meaning that many tournament results are inevitably ignored. Secondly, points awarded by organizations are not on the same scale, making it difficult to compare players across rating systems. For instance, public WPA ratings range from 200 to 25,263 while public Matchroom Pool ratings range from 14 to 151.

In addition, none of these point-based rating systems have probabilistic interpretations. Though rating differences may give some indication of relative strength, there is no principled way to quantify the probability that player A beats player B based on their ratings. These drawbacks indicate the need for an independent rating system with a common scale that has a probabilistic interpretation.

To address these concerns, Steve Ernst and Michael Page created FargoRate which rates pool players around the world on the same scale. At the time of writing, FargoRate maintains ratings for 235,574 professional and amateur players in 130 countries with information from 18,647,218 racks. FargoRate is an Elo-based rating system, so it has a probabilistic interpretation at the rack-level. FargoRate has been adopted by CueSports International (CSI) and is used for seeding at popular tournaments like the 2018 US Open Championship.

FargoRate is clearly an improvement upon organizational rating systems. However, there are drawbacks to using FargoRate to predict new tournament results at the professional level. FargoRate does not publicly release its player rating calculations or the data upon which the ratings are based, meaning that their ratings must be taken at face value. In addition, FargoRate's scope is extremely broad: it was created to compare all pool players in the world from Sunday bar league to the professional level. If the goal is to predict the results of high-level invitational tournaments, a system trained specifically for professional pool might do better. Lastly, FargoRate player ratings are calibrated to predict rack-level results, not match-level results, and FargoRate provides no measure of uncertainty for player ratings.

Project Goal

Given the drawbacks of existing rating systems, the goal of our project is to create Elo and Glicko-2 rating systems for professional 9-Ball pool based on publicly-available tournament results. These rating systems should be able to effectively predict the outcome of new invitational tournaments at the professional level.

2 Data

To create our own rating systems, we scraped data from a variety of publicly-available sources. In the absence of a Sports Reference page for pool, we relied on tournament- and organization-specific web pages to collect match-level results. We were able to collect data from 94 tournaments between 2007-2020 for a total of 5,085 matches (72,530 racks, 1,248 unique players) to use as training data. For each match, we recorded the number of racks each player won along with the names of each participating player, the match date, and the tournament.

We also scraped the results of the most recent major 9-Ball event, the 2021 Predator Championship League Pool (CLP) to be used as a validation dataset. The Predator CLP is a mixed-gender invitational tournament with 192 matches (1,418 racks) between 19 high-level, professional competitors.

The granularity of this data varies greatly depending on the tournament—some tournaments include a full record of matches played while others only include truncated results from a certain round and beyond. This likely introduces survivorship bias into our dataset: players that appear in more matches in our dataset have progressed to the latter stages of tournaments often, so we might expect them to be better players. Since our goal is to evaluate the best professional players, this bias should not create too much of a problem. For most professional 9-Ball players, a tournament really begins when the field of competition narrows, and they start to play other professionals. Since late stages of a tournament offer the most information about how professional players compare to other professionals, our training dataset is likely sufficient for the task of predicting a high-level invitational event like the Predator CLP.

3 Elo

Overview

The Elo rating system is a dynamically updating method used to assess the relative strengths of players competing in zero-sum games, such as chess. Players' ratings depend on their opponents' ratings and their game results. The difference in player ratings can then serve as a predictor for match outcome. Additionally, the ratings are only valid within the rating pool in which they were established, meaning that we cannot compare player A's rating to player B's rating unless they were constructed with the exact same system and data.

At a high level, Elo ratings are updated as follows. If a higher-rated player beats a lower-rated player, a few points are transferred from the lower-rated player to the higher-rated player. If a lower-rated player wins, then many points are transferred from the higher-rated player to the lower-rated player. If a game between a higher- and lower-rated player ends in a draw, then a few points are transferred from the higher-rated player to the lower-rated player because the higher-rated player was expected to win.

On a mathematical level, the two key equations in an Elo system are the expected scores calculation and the linear update rule. In our system, the expected score is equivalent to the probability that a player wins a match. The following logistic function specifies this probability. Given players A and B with ratings R_A, R_B, the expected score for player A (or, equivalently, the probability that player A beats player B) is:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

The 400 seen in the denominator is the Elo width which is generally accepted to be 400. This expected value is then used to update each player's. If player A truly scores S_A points, the new rating is:

$$R_A = R_A + K(S_A - E_A)$$

The update is proportional to the amount by which the player over or under performs inflated by the K-factor which is the maximum possible rating adjustment per game. To tune K, we find the value of K that maximizes the log-likelihood over the validation set. Recall that the validation set is a 2021 tournament consisting of 19 players. For each value of K, we fit ratings on the training data, then calculate the log-likelihood on the validation set. We select the K value with the largest log-likelihood.

Putting everything together, we train a baseline Elo system by initializing all 1,248 players with a rating of 1500. Then, each of the 5,085 matches are considered in chronological order. For each match, we determine the winner and loser. Then, using the expected score and update rule, we update both players' Elo ratings. The ratings are then returned and sorted to determine the best players.

Adjusted Elo Systems

We calibrate three additional Elo systems, since the baseline Elo method does not account for some of the additional information available in our data.

The first adjusted system accounts for score differential by implementing a Margin of Victory (MOV) multiplier inspired by FiveThirtyEight's NFL Elo system. The MOV multiplier inflates K by the logged score differential, and the system also discounts the MOV if an expected favorite wins (since a win is an unsurprising outcome). Given the score differential, SD, the winner's Elo rating, R_W, and the loser's Elo rating, R_L, the multiplier is:

$$MOV = \ln(SD + 0.5) \cdot \frac{2.2}{0.001 \cdot (R_W - R_L) + 2.2}$$

The second adjusted system treats each rack as an individual match in order to inflate the amount of data. However, we do not know the order in which racks were played, so a player's rating may be overinflated or deflated because we have to decide the order in which to update players' ratings.

The third adjusted system initializes ratings at different levels based on the number of games a player participates in. Since most of our data comes from tournaments, we hypothesize that players who appear in more games are likely better. 100 points are added to the initial 1500-point rating for the players in the top 90% based solely on the number of games. In our data, players in the top 90% appear in 17 or more games. On the other end, 100 points are subtracted from the initial 1500-point rating for the players in the bottom 10% which constitutes players with only 1 or 2 games. While this system unfortunately introduces data leakage since future data is used to initialize the model, we still want to test its efficacy.

Elo Results

For each of the four Elo systems (Baseline, MOV, Racks-as-Matches, Initial Ratings), we first calibrate K as described in the Overview section. The calibration plots are shown in Figure 1.

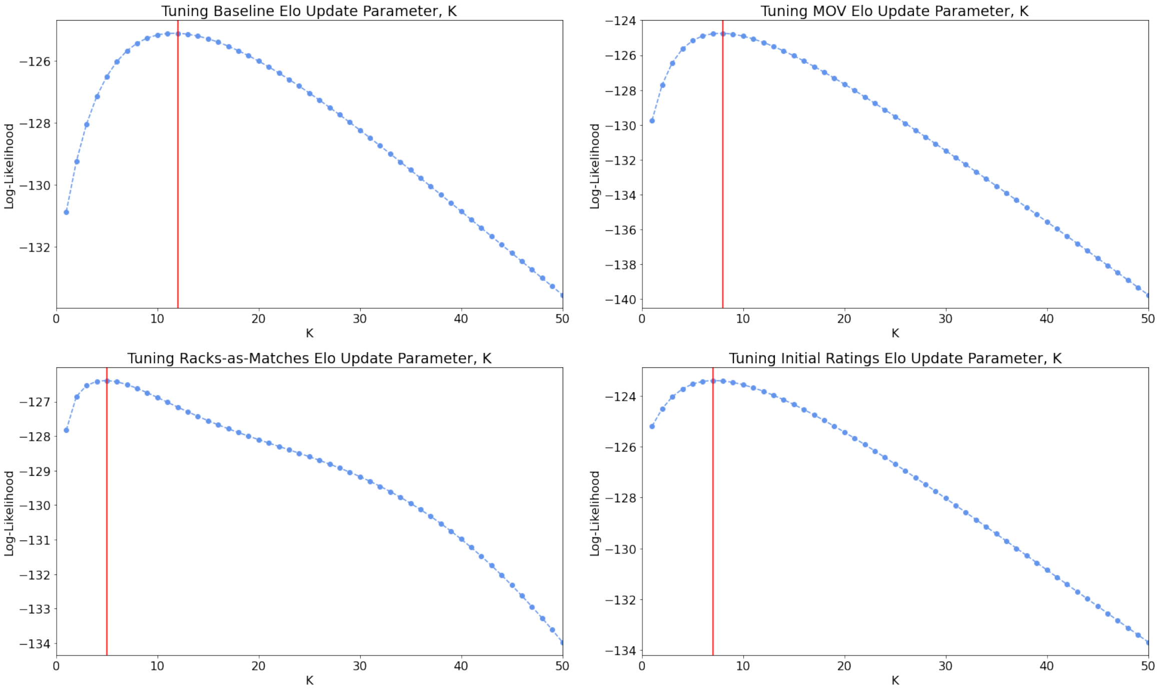


Figure 1: Tuning Update Parameter, K

Then, we re-calibrate each Elo system with its corresponding optimal K-factor and evaluate the log-likelihood on the validation set. Table 1 orders the systems based on best (i.e. maximized) log-likelihood.

Method	K	Log-Likelihood
Initial Ratings	7	-123.4
MOV	8	-124.7
Baseline	12	-125.1
Racks-as-Matches	5	-126.4

Table 1: Elo Standings

Overall, the Initial Ratings Elo system has the largest log-likelihood on the validation set. Both the Initial Ratings and MOV systems outperform the Baseline Elo method. However, inflating the data by taking racks as matches appears to worsen the model. In section 5, we will compare the Elo systems discussed here to the Glicko-2 systems presented in the subsequent section and the existing 9-Ball pool rating systems discussed previously.

4 Glicko-2

Overview

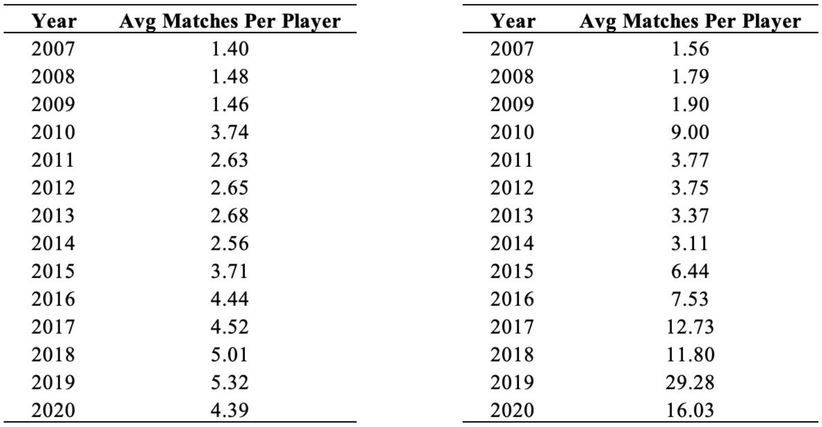
The Glicko system, developed by Mark Glickman, extends the Elo system by computing a ratings deviation (RD) which measures the uncertainty of a rating. A high RD corresponds to an unreliable rating, and it indicates that a player may not compete frequently or has only competed in a small number of games. The Glicko-2 system adds a measure of rating volatility. The volatility measure indicates the degree of expected fluctuation in a player's rating. The measure is large when a player has erratic performances and small when the player performs at a consistent level.

Applying a Glicko-2 rating system to 9-Ball pool is valuable because the number of matches per player varies significantly (it ranges from 1 to 171 in our data), so the RD measure is very relevant. Furthermore, we are interested in not only identifying the top players but also quantifying the consistency of their performances. This is captured by the volatility measure, a feature unique to Glicko-2.

In order to implement Glicko-2, we need to tune two parameters. The first is τ which constrains the change in volatility over time. According to Glickman, smaller values of τ prevent the volatility measures from changing by large amounts which in turn prevents enormous changes in ratings based on very improbable results. He suggests 0.3 - 1.2 as a reasonable range of values for τ.

The second parameter is the rating period which is a collection of games that are treated as having occurred simultaneously. Updated ratings and RD's are computed at the end of the rating period then used as the pre-period ratings and RD's for the next rating period. Glickman notes that the system works best when there are at least 10-15 games per player in a rating period, on average.

Figure 2a shows the average number of matches per player by year in our dataset. If we treat each year as a rating period, it is obvious that we are quite far from the suggested threshold of 10-15 games per player in a rating period.



Year	Avg Matches Per Player
2007	1.40
2008	1.48
2009	1.46
2010	3.74
2011	2.63
2012	2.65
2013	2.68
2014	2.56
2015	3.71
2016	4.44
2017	4.52
2018	5.01
2019	5.32
2020	4.39

(a) Players Unpooled

Year	Avg Matches Per Player
2007	1.56
2008	1.79
2009	1.90
2010	9.00
2011	3.77
2012	3.75
2013	3.37
2014	3.11
2015	6.44
2016	7.53
2017	12.73
2018	11.80
2019	29.28
2020	16.03

(b) Players Pooled

Figure 2: Average Matches Per Player by Year

In order to address this issue, we consider aggregating players who have fewer than x matches by number of matches played. For example, if $x = 20$, then all players who have played 1 game are treated as one player, all players who have played 2 games are treated as one player, and so on, for all players with less than 20 matches. The rating of an aggregated group hypothetically represents the average strength of players who have played the same number of matches. Figure 2b shows the average number of matches per player by year when we aggregate players with fewer than 20 games. We're now closer to achieving the 10-15 games per player threshold, particularly from 2016 onwards.

Intuitively, the number of matches played seems like a reasonable proxy for player strength, but only up until a certain point. For example, most people would probably agree that there's a huge difference between players with 50 professional matches under their belt vs. those with 5. It seems reasonable to assert that, on average, the ratings of players with 5 professional matches should be closer to each other than to the ratings of players with 50 professional matches. However, we could not confidently make the same claim about players with, say, 30 matches compared to players with 50 matches. Given this uncertainty, we test the value of player aggregation at different minimum game thresholds in the following section.

Finally, we note that another potential fix to the small number of matches per player in a rating period is to change the rating period itself. For instance, 2b above shows that the mean matches per player is much smaller in early years, so perhaps we should consider only including matches from 2010 or 2016 onwards. The trade-off here is that we throw away data, and intuitively, more data should mean more accurate ratings. We experiment with tuning the rating period in the next section.

Implementation

Our goal is to tune the Glicko-2 model such that we maximize log-likelihood on the validation set. In order to do so, we use the glicko2 function from the PlayerRatings package to calculate players' ratings and rating deviations. The resulting ratings and RD's are on the Glicko scale, so we convert to the Glicko-2 scale using the methodology provided by Glickman. Then, for each match in the validation set, we calculate the probability of player A beating player B (where player A was actually the winner) using the following formulas:

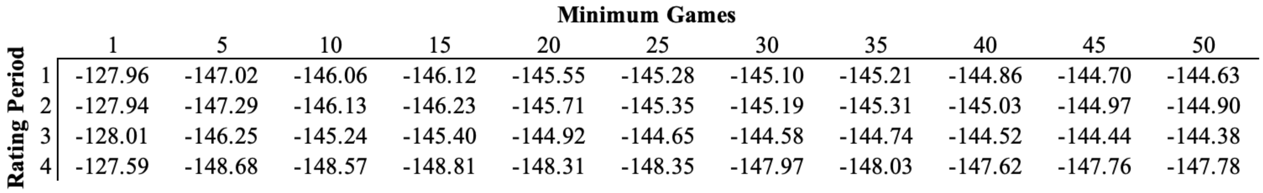
$$E[\mu, \mu_j, \phi_j] = \frac{1}{1 + \exp(-g(\phi_j)(\mu - \mu_j))} \quad \text{where} \quad g(\phi) = \frac{1}{\sqrt{1 + 3\phi^2/\pi^2}}$$

Note that μ is player A's rating, μ_j is player B's rating, and ϕ_j is player B's RD. We take the log of the calculated probability, then repeat and sum the results for all matches, giving us the log-likelihood.

We define the set of possible values for each parameter in our model as follows:

1. We use the suggested range of 0.3 - 1.2 for the change in volatility constraint, τ.
2. We consider four possible rating periods: (1) each year from 2007 to 2020 is a rating period; (2) each year from 2010 to 2020 is a rating period; (3) 2010 and the years from 2015-2020 are each distinct rating periods, and the years from 2011-2014 are treated as one rating period; (4) each year from 2016 to 2020 is a rating period.
3. For player aggregation, we consider minimum game thresholds ranging from 1 to 50 (where a threshold of 1 means there is no aggregation of players).

Our next step is to find the value of τ that maximizes the validation set log-likelihood for each combination of rating period and player aggregation. We find that the optimal value of τ is always 0.3. Finally, we compare the log-likelihoods produced by every rating period and player aggregation combination when τ is set to 0.3. Figure 3 shows the results. The model corresponding to the largest log-likelihood treats each year from 2016 to 2020 as a rating period and does not aggregate any players.

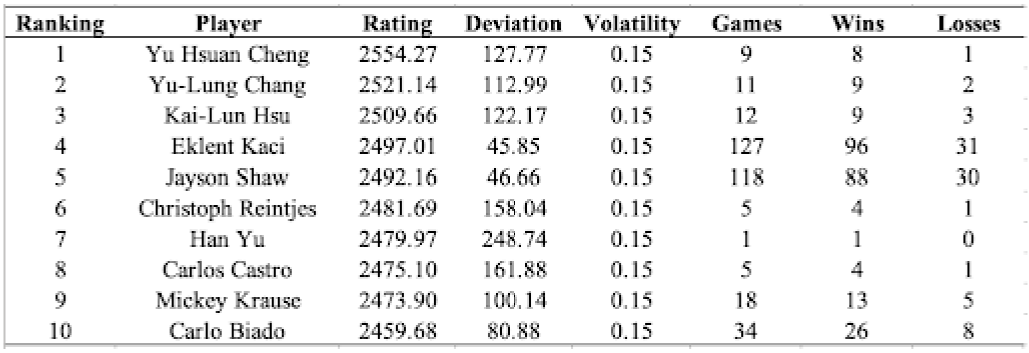


	Minimum Games										
Rating Period	1	5	10	15	20	25	30	35	40	45	50
1	-127.96	-147.02	-146.06	-146.12	-145.55	-145.28	-145.10	-145.21	-144.86	-144.70	-144.63
2	-127.94	-147.29	-146.13	-146.23	-145.71	-145.35	-145.19	-145.31	-145.03	-144.97	-144.90
3	-128.01	-146.25	-145.24	-145.40	-144.92	-144.65	-144.58	-144.74	-144.52	-144.44	-144.38
4	-127.59	-148.68	-148.57	-148.81	-148.31	-148.35	-147.97	-148.03	-147.62	-147.76	-147.78

Figure 3: Log-Likelihood of Models by Parameter Combination

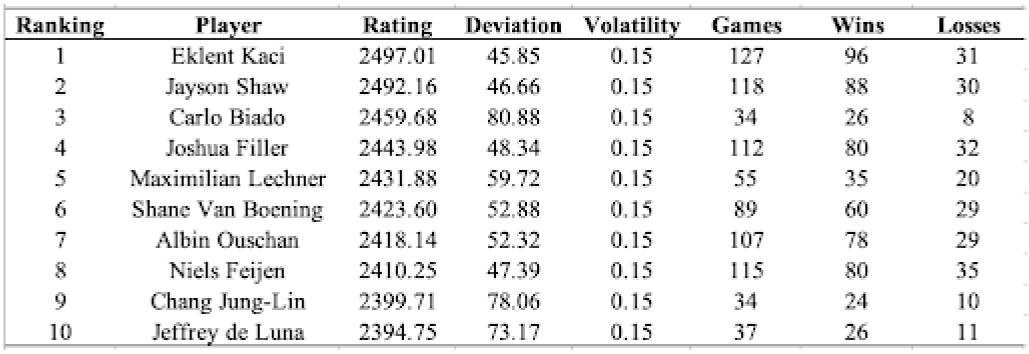
Figure 4 shows the players with the ten largest ratings according to our best Glicko-2 model. Note that the ratings and rating deviations are reported on the Glicko scale. We see that 7 of the top 10 players have fewer than 30 games - these players' deviations are much higher, as expected.

If we filter for players with at least 30 games, we see much more familiar names in the top 10 (Figure 5). Approximately 5% of players in our dataset have played at least 30 games, so if we again assume that number of matches is a proxy for strength, this group should represent the strongest players.



Ranking	Player	Rating	Deviation	Volatility	Games	Wins	Losses
1	Yu Hsuan Cheng	2554.27	127.77	0.15	9	8	1
2	Yu-Lung Chang	2521.14	112.99	0.15	11	9	2
3	Kai-Lun Hsu	2509.66	122.17	0.15	12	9	3
4	Eklent Kaci	2497.01	45.85	0.15	127	96	31
5	Jayson Shaw	2492.16	46.66	0.15	118	88	30
6	Christoph Reintjes	2481.69	158.04	0.15	5	4	1
7	Han Yu	2479.97	248.74	0.15	1	1	0
8	Carlos Castro	2475.10	161.88	0.15	5	4	1
9	Mickey Krause	2473.90	100.14	0.15	18	13	5
10	Carlo Biado	2459.68	80.88	0.15	34	26	8

Figure 4: No Game Threshhold Glicko-2 Ratings



Ranking	Player	Rating	Deviation	Volatility	Games	Wins	Losses
1	Eklent Kaci	2497.01	45.85	0.15	127	96	31
2	Jayson Shaw	2492.16	46.66	0.15	118	88	30
3	Carlo Biado	2459.68	80.88	0.15	34	26	8
4	Joshua Filler	2443.98	48.34	0.15	112	80	32
5	Maximilian Lechner	2431.88	59.72	0.15	55	35	20
6	Shane Van Boening	2423.60	52.88	0.15	89	60	29
7	Albin Ouschan	2418.14	52.32	0.15	107	78	29
8	Niels Feijen	2410.25	47.39	0.15	115	80	35
9	Chang Jung-Lin	2399.71	78.06	0.15	34	24	10
10	Jeffrey de Luna	2394.75	73.17	0.15	37	26	11

Figure 5: Thresholded Glicko-2 Ratings (30 Games)

Finally, we note that players have very similar volatilities. In fact, the difference between the smallest and largest volatility among all of the players in our dataset is only 0.0017. Even when we set τ, the change in volatility constraint, to the largest value in the suggested range, the difference between the most consistent and least consistent player is 0.0277. This implies that there aren't large differences in consistency of performance across professional 9-Ball pool players.

5 Results and Model Comparison

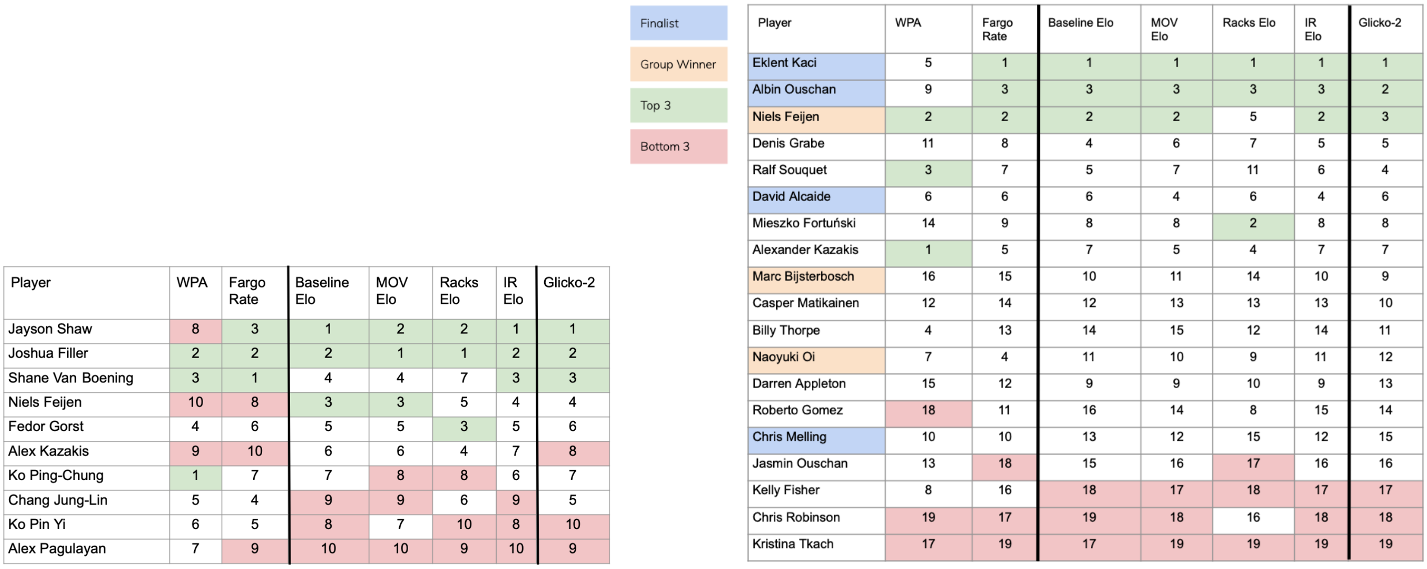
After training and tuning our models, we evaluate them via log-likelihood on the Predator CLP validation dataset (Figure 6b). We find that all of our models, both Elo and Glicko-2, outperform FargoRate in terms of log-loss on the validation set. Table 2 shows the systems, their parameters, and the log-likelihood ordered by maximized log-likelihood:

Method	Parameters	Log-Likelihood
Initial Ratings (Elo)	K = 7	-123.4
MOV (Elo)	K = 8	-124.7
Baseline (Elo)	K = 12	-125.1
Racks-as-Matches (Elo)	K = 5	-126.4
Glicko-2	$\tau = 0.3$	-127.6
FargoRate (Elo)	N/A	-128.2

Table 2: Model Comparison

As a sanity check, we also compare our ratings systems to those of the WPA (which rates players based on the WPA matches they participate in). Figure 6a below shows players' rankings in each rating system and highlights the top 3 players and bottom 3 players. For the most part, our Elo and

Glicko-2 systems agree on relative rankings. The WPA ratings are quite different from all of our rating systems which makes sense since their rankings only consider WPA games.



Player	WPA	Fargo Rate	Baseline Elo	MOV Elo	Racks Elo	IR Elo	Glicko-2
Jayson Shaw	8	3	1	2	2	1	1
Joshua Filler	2	2	2	1	1	2	2
Shane Van Boening	3	1	4	4	7	3	3
Niels Feijen	10	8	3	3	5	4	4
Fedor Gorst	4	6	5	5	3	5	6
Alex Kazakis	9	10	6	6	4	7	8
Ko Ping-Chung	1	7	7	8	8	6	7
Chang Jung-Lin	5	4	9	9	6	9	5
Ko Pin Yi	6	5	8	7	10	8	10
Alex Pagulayan	7	9	10	10	9	10	9

(a) WPA Rating Comparison

🟦 Finalist

🟧 Group Winner

🟩 Top 3

🟥 Bottom 3

Player	WPA	Fargo Rate	Baseline Elo	MOV Elo	Racks Elo	IR Elo	Glicko-2
Eklent Kaci	5	1	1	1	1	1	1
Albin Ouschan	9	3	3	3	3	3	2
Niels Feijen	2	2	2	2	5	2	3
Denis Grabe	11	8	4	6	7	5	5
Ralf Souquet	3	7	5	7	11	6	4
David Alcaide	6	6	6	4	6	4	6
Mieszko Fortuński	14	9	8	8	2	8	8
Alexander Kazakis	1	5	7	5	4	7	7
Marc Bijsterbosch	16	15	10	11	14	10	9
Casper Matikainen	12	14	12	13	13	13	10
Billy Thorpe	4	13	14	15	12	14	11
Naoyuki Oi	7	4	11	10	9	11	12
Darren Appleton	15	12	9	9	10	9	13
Roberto Gomez	18	11	16	14	8	15	14
Chris Melling	10	10	13	12	15	12	15
Jasmin Ouschan	13	18	15	16	17	16	16
Kelly Fisher	8	16	18	17	18	17	17
Chris Robinson	19	17	19	18	16	18	18
Kristina Tkach	17	19	17	19	19	19	19

(b) Validation Set Rating Comparison

Figure 6: Rating System Comparison

Finally, we apply our rating systems to the 19 players in the Predator CLP tournament and compare the players' rankings. Figure 6b highlights the true finalists and group winners as well as the top 3 and bottom 3 rated players in each system. Again, we see that our Elo and Glicko-2 systems mostly agree on rankings.

6 Drawbacks

Throughout our analysis, we have mentioned limitations such as data leakage in the Initial Ratings Elo model and survival bias in our training data. An additional limitation is that the tournaments in our dataset are primarily mixed-gender and men's due to differential access to data. As a result, the ratings for women in our dataset are likely less accurate on average than our ratings for men.

Furthermore, our dataset does not include some important match information. For example, the most important shot in 9-Ball is the first shot of the rack, known as the break. There is evidence to suggest that the player who takes the break shot has a moderate advantage in winning the rack. The effect size of this advantage increases significantly if the player pockets a ball on the break as they have the opportunity to stay at the table and continue playing. This is a potential explanation for our rack-level Elo system's poor performance relative to match-level models. Since the break shot tends to alternate during a match over the course of many racks, match-level models are less sensitive to the exclusion of this covariate than rack-level models. With the inclusion of this covariate, however, we hypothesize that the quality of rack-level rating systems would greatly improve.

Additionally, our dataset does not include any player- or tournament-level covariates. Player-level covariates like age, years of experience, and height (important for reaching certain shots), or exogenous tournament-level covariates like table manufacturer, pocket size, table size, and tournament rules could all be useful to consider. Including these covariates in our models might allow us to more accurately estimate the relative skill levels of our players and improve our prediction accuracy on new data.