Elo and Glicko in the National Hockey League

Caleb Ren

April 30, 2021

Introduction

The Elo and Glicko systems are both ratings systems to determine the relative strengths of teams or players within a sport. Rating systems assign a relative strength to each team to provide a way of predicting hypothetical match outcomes. The Elo system was originally designed for zero-sum games like chess by Arpad Elo. The US Chess Federation adopted the Elo rating system in 1960 as a way to systematically assign self-adjusting strength ratings to each player. A player with a higher Elo rating is presumed to be stronger than a player with a lower Elo rating and the probability that the higher-rated player wins is larger than the alternative. After each match, Elo points are redistributed from player to player. If the stronger player wins, a small amount of points might be transferred from loser to victor such that any Elo points forfeited by the loser are transferred to the winner. However, should an upset occur, a large amount of points might be transferred. The Elo system, though originally designed for chess, has been implemented in a wide variety of other games such as basketball[1] and even esports like Dota[2] and League of Legends.[3]

The Glicko system, designed by Mark Glickman in 1995 with a significant refresh in 2012, is an improvement over the Elo system that relaxes several assumptions. First, Glicko assigns a *ratings deviation* in addition to the point estimate Glicko rating that measures a 95% confidence interval the Glicko rating in order to capture uncertainty in ratings.[4] In addition, Glicko outcomes are not balanced in a zero-sum fashion like in Elo: a loss of x Glicko points by the loser does not imply a gain of x Glicko points by the victor. Glicko ratings are also calculated not match-by-match but in larger batches of about 10-15 games per player. Although more difficult to calculate, Glicko has been implemented in nearly as many situations as Elo due to its assumption that ratings deviations grow larger over time without active play—as players sit out of tournaments, their most recent Glicko rating becomes ever more unreliable.[5]

The sports analytics and election prediction site FiveThirtyEight assigns Elo ratings to the National Hockey League's annual season. However, there has not yet been an implementation of Glicko for the National Hockey League in any major sports publication or outlet. My aims through this research project are to:

1) Calibrate Elo and Glicko rating systems for the National Hockey League by finding tuning parameters that maximize log-likelihoods;
2) Compare the ranking of NHL teams under the Elo and Glicko rating systems;
3) Visualize the Elo and Glicko ratings for the NHL from 2009 to 2019.

Methodology

The dataset employed is the NHL 2009 to 2019 dataset distributed in class. This dataset covers 10 NHL seasons from the 2009-2010 to the 2018-2019 season. The data was separated into training and

[1] https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/
[2] https://www.datdota.com/ratings/top?type=elo64
[3] https://leagueoflegends.fandom.com/wiki/Elo_rating_system
[4] Note Elo also assumed a Normal distribution for each player's Elo rating, but the confidence interval for Elo ratings is almost never reported in practice because it is not involved in Elo rating updates.
[5] https://www.bayesesports.com/files/blogpictures/rating_algorithms.pdf

validation sets (2009-2017 seasons and 2017-2019 seasons, respectively). In order to maintain consistency, both the Elo and Glicko systems were trained on the 2009-2019 NHL data and tuning parameters were optimized according to a log-likelihood function.

Elo and Glicko ratings were not reset between seasons. Though there is an argument that ratings should be reset prior to each season, the choice to maintain previous ratings (including rating deviations and volatility under Glicko) was to capture long-term team momenta and trends rather than examining ratings season-by-season.[6]

There are several unique data cleaning considerations. Of note is that 2 NHL teams changed either name or location since 2009. The Phoenix Coyotes rebranded to the Arizona Coyotes for the 2014-2015 season and the Atlanta Thrashers moved to Winnipeg becoming the Jets in 2011. To deal with these issues, I renamed any data point in the dataset to the most recent name (e.g. "Atlanta Thrashers" to "Winnipeg Jets") since these franchises more or less remained the same though they went by different names. In addition, the NHL added the Vegas Golden Knights franchise for the 2017-2018 season as the 31st team in the league, meaning no previous data exists on the Golden Knights prior to 2017. Therefore, I initialized a new set of Elo and Glicko ratings starting in 2017 for the Golden Knights (1500 for both, 350 for Glicko rating deviation) and used the 2017 through 2019 data points as a validation dataset.

Another data limitation is the 2012-2013 NHL lockout.[7] Due to a labor dispute between the NHL players and the owners, the regular 82-game season was reduced to just 48 games. Although this would not be a large concern for the Elo rating system, the Glicko rating system tends to become unstable when there are few games played within a rating period. Therefore, each rating period is allocated such that each team plays about 10 matches per rating period, so the 2012-2013 season only contains 4 rating periods compared to a typical season's 8 rating periods.

Simulation of Elo and Glicko systems was conducted using Python. The Elo algorithm was adapted from Laurie Shaw's Elo algorithm while the Glicko algorithm was coded from scratch using specification from Mark Glickman's corrected Glicko-2 whitepaper.[8] Data visualization was conducted in R as opposed to Python in order to take advantage of `ggplot`'s more advanced visualization capabilities. An interactive version of the Python notebook to run Elo and Glicko with user input for the model parameters is available here at Google Research Colab.[9]

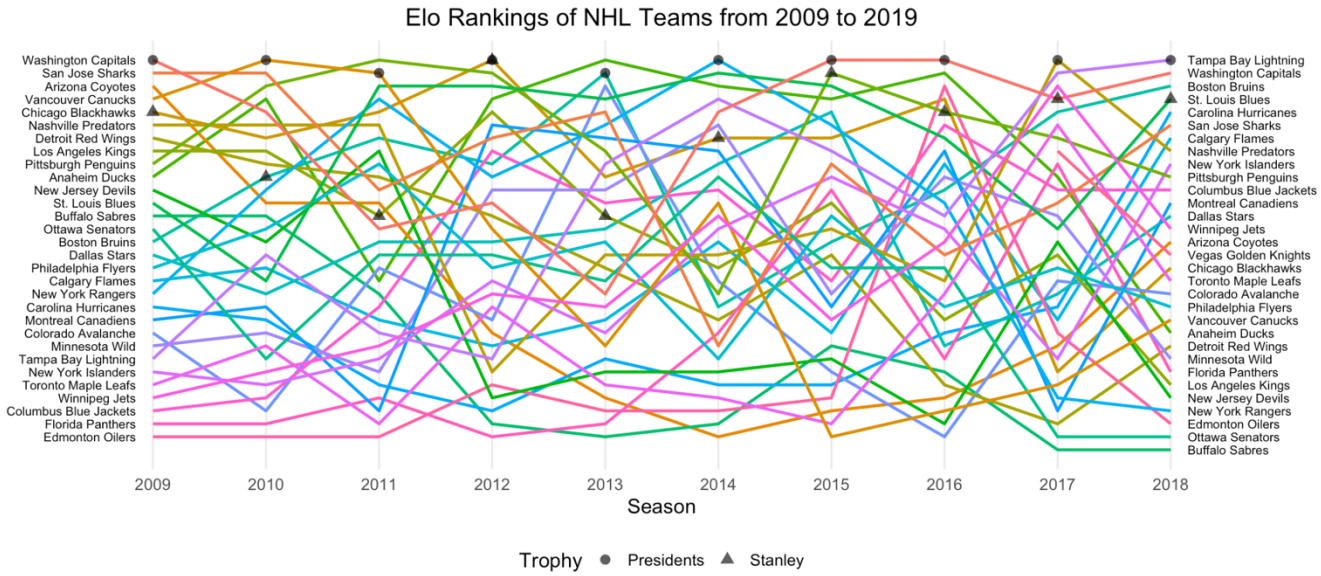
Discussion

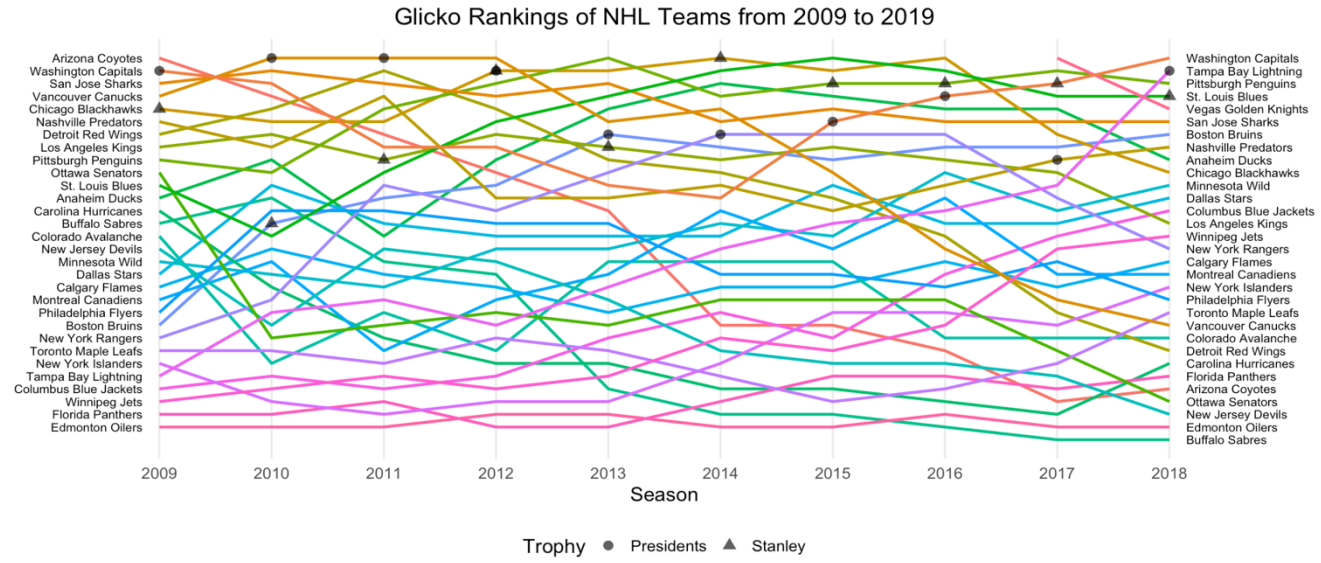
After calibration, the optimal value of the Glicko system parameter τ is 0.2 and the optimal value of the initial volatility σ is 0.02. According to the specification of the Glicko-2 algorithm, smaller values of τ suggest highly improbable collections of game outcomes, which indicates that the NHL produces more upsets than comparable sports.[10] One possible source of this is hockey's low-scoring nature, combined with a high degree of randomness. Ties are frequent and broken by overtime and shootouts. In fact, about a quarter (23.6%) of games in the 2009-2019 dataset end in overtime (12.2%) or shootout

[6] Some sources do allow the option to toggle between "seasonal" and "continuous" ratings: https://hockeyeloratings.com/teams/elo
[7] https://en.wikipedia.org/wiki/2012%E2%80%9313_NHL_lockout
[8] http://www.glicko.net/glicko/glicko2.pdf
[9] https://colab.research.google.com/drive/1KINxylJcseNSsSULLO-9_uA6iENHynai?usp=sharing
[10] This was suggested in the "Luck vs. Skill" presentation given by Patrick Wang on April 23, 2021.

(11.4%). Thus, a significant proportion of hockey games hinge on a single crucial sudden-mouth moment in overtime or shootout, increasing the chance that upset victories may occur.

The optimal value of the Elo *k*-factor was *k*=13 after validation using log-likelihood. The year-over-year Elo rankings show huge swings in Elo rankings every year for each team, most likely due to the Elo rating system's lack of sensitivity to ratings deviation that the Glicko-2 system incorporates into the algorithm. As seen below, the rank-flow diagram for Elo rankings shows much more chaotic motion every year compared to the Glicko-2 rankings. This suggests that using a well-calibrated Glicko-2 system offers more stability and less drastic shifts in year-over-year rankings than a similarly-calibrated Elo system. Thus, as a recommendation, the Glicko-2 system seems to be better suited for a sport like the hockey as implemented in the National Hockey League.





I examined the Glicko-2 rankings year-over-year to see how the Glicko rankings for each team changed. The rank-flow chart above shows the change in Glicko rankings from the end of the 2009-2010 season

to the end of the 2018-2019 season. All teams started with the same Glicko rating of 1500, a rating deviation of 350, and a volatility of 0.02. Teams that won the Presidents' Trophy that season are additionally marked with a circle while teams that won the Stanley Cup playoffs are marked with a triangle. In general, the Glicko rankings were quite good at predicting successful teams for each season. The top team in Glicko rankings in 2010 and 2011 (Vancouver Canucks) took home the Presidents' Trophy for the best regulation season point record. In every season except the 2013-2014 season, at least one of the Presidents' Trophy or Stanley Cup winner fell within the top 5 Glicko ranking finishers. There are 2 major outliers: the 2010-2011 season when the 14th-ranked Boston Bruins pulled off a major upset to take home the Stanley Cup, ending a 39-year championship title drought; and the 2013-2014 season, the first season after the 2012-2013 NHL lockout.

The trajectory of several teams over the years have been of note. The Arizona Coyotes (then the Phoenix Coyotes) led the rankings for the 2009-2010 season, where they achieved a decent 50-25-7 record in a competitive Western Division and scored 107 points. Over the next decade, they slipped further down the league, ending in 28th place for the 2018-2019 season. Though the high early rankings for Arizona may have been the Glicko algorithm calibrating itself and rating deviation stabilizing, it seems that Arizona did get worse: they only racked up 86 points in 2018. This is almost entirely reverse the trajectory of the Tampa Bay Lightning, which started in 28th position and climbed up to 2nd in the 2018-2019 season, the year they won the Presidents' Trophy.

The Buffalo Sabres started 2009 in the middle of the pack and ended last in the league in 2018-2019. Their performance flagged during the lockout and never quite recovered, as the Sabres went from a 39-32-11 record prior to the lockout to a 21-21-6 record during the lockout season. The following season, Buffalo kept a constant number of wins but more than doubled their losses for a 21-51-10 record, leading to the precipitous 10-spot drop in Glicko rankings from 18th to 28th.

Future Directions

Future extensions of this project include calibrating Elo and Glicko systems for players in a two-way play game (offense and defense),[11] calibrating Elo and Glicko systems using shot opportunities rather than game-level statistics, and further calibrating Chessmetrics[12] or Universal[13] rating systems to the NHL and comparing against Elo and Glicko.

In addition, a limitation to the way that I have chosen to conduct my Glicko ratings is to treat the *number of games* per rating period as fixed, rather than fixing the time interval (every 4 weeks, for instance). I made the conscious choice to do so since otherwise there would be a large increase in rating deviation every year during the off-season from June until October. However, an argument for fixing the rating period to time interval would be to investigate the effect of busier or sparser weekly schedules on team ratings, as well as creating a more sensitive rating system to the disruptive effects of the 2012-2013 NHL lockout on Glicko ratings.

[11] http://www.hockeystatsrevolution.com/
[12] http://chessmetrics.com/cm/CM2/Formulas.asp
[13] https://en.wikipedia.org/wiki/Universal_Rating_System