METHODS OF SPORTS ANALYTICS IN GOLF

December 12, 2019

Haley Talton, Ted Henson, Pranav Ram, Jon Huml

Contents

1 Introduction

Golf is a sport that is enjoyed by people of all ages and experience levels. The game is played using various types of clubs to strike a ball with the intention of getting the ball from the tee box into the hole in as few strokes as possible. Golf is played on a course that consists of 18 holes. Each hole is classified by its par: the number of strokes that it should take a skilled golfer to get the ball into the hole. Pars range from 3 to 5. The score is how many strokes under or over par a player is for the amount of holes they have completed. Penalties in the form of strokes added to a player's score are incurred for lost balls and any rule violations. The player with the fewest amount of total strokes at the end of a round is declared the winner.

Professional and amateur golf tournaments are held year round, unlike other sports that have designated seasons. Golf is played all around the world, but based on the number of facilities, it is most popular in the United States. Japan is a distant second, followed by Canada and England [1]. According to the National Golf Foundation, golf is an $84 billion industry with a support level of around 33.5 million participants in the US alone [2]. A recent surge in the number of golf courses and rising demand for golf equipment worldwide leads to the assumption that the industry will only continue to grow in the years to come [3].

Golf is a multifaceted sport with high potential for data collection and analysis. Although the objective of the game is simple, golf has many layers and can prove to be quite complex. With the emergence of the ShotLink system in 2003, real-time shot by shot data became readily available, thus presenting a multitude of opportunities for statistical analysis in golf. Considering the complex nature of the game and emerging presence of data analysis in the sport, our team deemed golf to be an optimal choice for research purposes.

2 Literature Review

2.1 Performance

Prior to the utilization of the ShotLink system, useful golf statistics were virtually nonexistent. The ones that did exist, such as "Total Putts" or "Driving Distance" tended to be misleading and gave little insight into a player's performance. After the PGA Tour employed the ShotLink system, Mark Broadie, a business professor at Columbia University, was able to utilize the data from the system to revolutionize golf statistics with his "strokes gained" methodology.

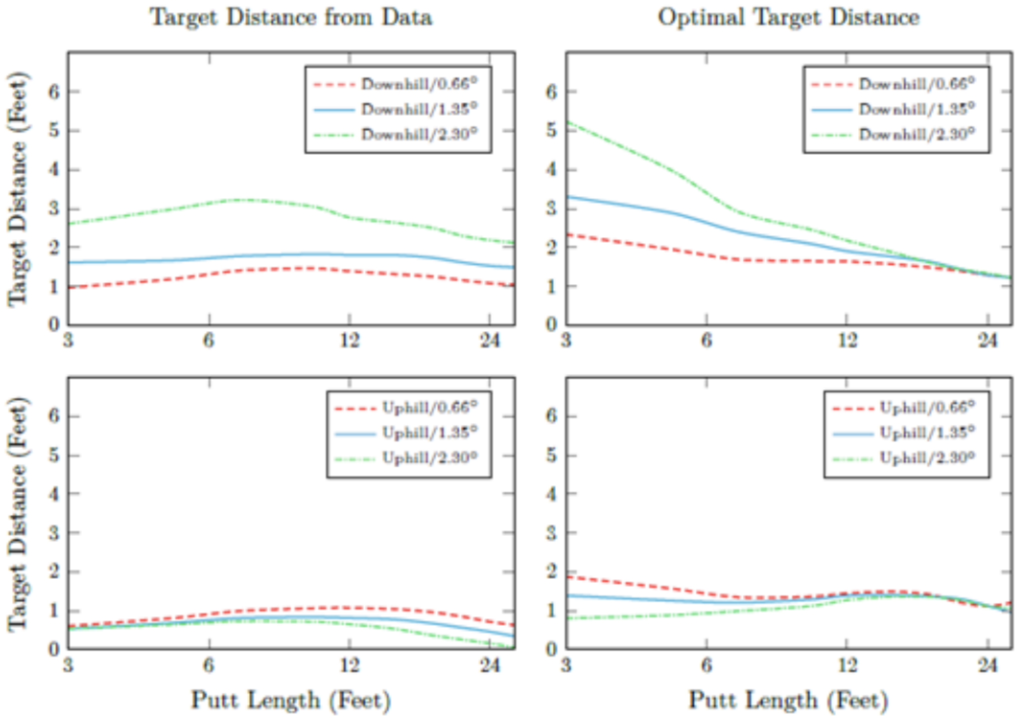
Pure score in golf is a mostly accurate yet suboptimal way to assess golfers given that equivalent scores are not the product of equivalent shots. There is a need to isolate the difficulty of specific shots that traditional statistics such as driving or putting accuracy, which are blind to distance or course-specific elements like grass or bunkers, cannot fully capture. Analogous to baseball's "wins above replacement" statistic, one method to contextualize performance is the "strokes gained" metric. Strokes gained is the difference of the average score from a given position before and after a shot. A negative number shows that the golfer saved some fraction of a stroke compared to the average player.

There are an infinite number of xyz-coordinates for the landing point of any given shot, so calculating highly accurate averages is difficult for specific spots on all fields. Rough averages are possible with the PGA Tour's sophisticated ShotLink ball-tracking system. Broadie has popularized this metric in the last decade with the rise of ShotLink, and strokes gained is now officially tracked by the PGA [4]. Tiger Woods is the known all-time leader in the metric, although pre-ShotLink golfers may have comparable scores; almost two-thirds of Woods' advantage in strokes gained is derived from the power and distance of his long game [5]. The recording of strokes gained has also shown that the importance of putting is far overstated in comparison to the long game, as putts are more sensitive to the random gradients or contours of a course's greens [6]. Course-specific models, rather than player-specific ones, have also been proposed to capture the geographical difficulty of a shot based on the terrain, hazards, and line-of-sight blockages for specific shots. The ISOPAR method is one mapping system that uses smoothing splines to describe the location and difficulty of shots in physical space [7].

Regarding course specific golf analytics, Todd Schneider applied Mark Broadie's strokes gained metric to show that despite the course at Augusta National being known for its "slick, undulating greens," all shots other than the putt are the most important [8]. Schneider notes that only three of the past thirteen Masters winners have been in the top 10 in putting strokes gained during the year they won the Masters tournament. Additionally, six of the 13 winners were below-average putters in terms of strokes gained and eight of the

13 winners ranked among the top 10 in strokes gained tee-to-green. All 13 winners were above-average tee-to-green players in the seasons they won. Although Schneider does not present anything new to the field of golf analytics, he applies Mark Broadie's strokes gained metric to refute the importance of putting at Augusta National in a concise and convincing manner.

In Mark Broadie's random putting model, he attempts to find areas in putting for better strategy [9]. He discovers that essentially across putting degrees and distances, players tend to target the ball shorter than the optimal targeting distance. See the graph below:



His explanation for this is that in the presence of green error (error in reading the distance and elevation) you should try to aim past the optimal distance to minimize the chance of not reaching the hole (as you have some leeway), but not too excessively as then your next putt is too far from the hole. He also runs an analysis that decomposes putting success into several factors and concludes that green reading ability is the most important.

Brian Leahy attempted to predict future performance of PGA tournament players using the PGA tour's proprietary Shotlink data [10]. Although some of his dissertation contains interesting results, much of it is marked by unnecessary detail, such as explaining the rules of golf and basic statistics for many pages. Despite this, he presented some interesting results and surveys prior research. He performed seven "experiments" as he calls them. In one experiment, he performed a two-sample t test on Adam Scott and Lee Westwood to see if there was a statistically significant difference in the player's performance (measured by the amount of money gained) before and after a caddy change. In the case of Adam Scott, Leahy claimed there may be evidence based on the t test between the two samples (before and after the caddy change). Notably, the 95% confidence levels in the two means do overlap. In the case of Lee Westwood, there was not significant evidence shown by the p value.

In another experiment, he used the Shotlink data to predict whether certain players would make the cut using k nearest neighbor, neural networks, and a logistic regression. His best model segmented his data into clusters, and then ran a logistic regression on the clusters to achieve an overall classification rate of 65% of making or missing the cut. Overall, Leahy produced some interesting models, ideas, and concepts, but the paper's unnecessary length (163 pages) hides some of its gems.

2.2 Gambling

Golf bets are becoming increasingly popular with the progress of sports gambling legislation in the United States. Odds ratios and lines maintain their meaning in golf, but there are far more degrees of freedom in betting since tournaments may be played by nearly 100 individuals. Bets exist for the overall tournament winner, round (a specific day of a tournament, usually Thursday through Sunday) winner, specific matchups between any two (or more) golfers, or an "each-way" bet, which pays out for both winning the tournament and specific placement [11]. Live bets allow certain gambles to be placed while the tournament is in progress.

3 Future Work

3.1 Business

Golf is one of the few sports where total purse winnings are distributed through a prize-based system (usually 18%, 10.8%, 6.8% for the top three players [12]). The function, we estimated, is roughly $f(x) = 24e^{-0.4x}$ for x (where x is finish placement) greater than or equal to one. Ties are paid by averaging over the number of places those players would have taken had they not tied [13]. For example, a second place tie for two players would pay the average of the second and third place prizes. While the tying prizes are easily-explained, the function we estimated is not. The PGA keeps the distribution constant but does not publish its methodology for this seemingly arbitrary function.

We propose a payout system that is based on quantile total strokes gained performance rather than a strict and unexplainable hierarchy. Since strokes gained can be measured on a daily basis for a tournament, highest payouts would always go to the top performer much like pure scoring would. However, payouts would be distributed based on performance relative to the field (the quantile score of the golfer). The prize gap would be smaller for a first place player who barely beats the second place player on the 18th hole in comparison to a first place finisher that demolishes the field by beating the next closest player by 10 holes. This would leave the current placement system in place, and the PGA Tour could still allot a certain percentage of winnings to the top three finishers if symbolic placement is important enough to the tradition of the sport; specific percentages within that allotment would simply vary on the magnitude of difference between the first, second, and third finishers. Regardless, increased transparency of the payout system is vital to adequate compensation of performance.

3.2 Performance

On the performance side of golf, much granular data has been applied to player metrics. One untapped analytical source, which is already being captured in raw form, is characteristic time (CT) of clubheads, especially drivers [14]. This is measured by bouncing a pendulum with a steel ball on the club's face. The elasticity of this bounce is measured as the "characteristic time" taken for the pendulum to bounce back to its original position. The statistic is captured only for compliance purposes, and according to the USGA, players are given binary "pass" or "fail" feedback. Open-sourcing the club data and publishing CT magnitude would allow for a new metric: club-adjusted power created (CAPC). CAPC would simply be the ratio of the yard length of a drive to its club characteristic time (its units would be yards per microsecond). This would allow players and researchers to quantify the effects of clubs. Any (or no) statistical significance of club effect on distance could modify the allowable CT, which is 257 microseconds as specified by the USGA. Since this data is already being collected, its practicality is only a matter of player agreement. A further extension of such a metric might include the length of clubs in a player's bag. Although this data is not collected by the USGA, measuring 14 clubs per player would also be practical to implement. Publicly available club length data would allow for highly accurate torque and angular momentum calculations for every PGA golfer. Golf clubs vary in size and composition per player more than any other hitting instrument in sports, and thus having more data about this equipment would unlock new performance-based analytical opportunities.

3.3 Strategy

A hole in analytical golf literature is the "golf as a strategy" approach. Much like football, where plays can be deliberated for a set amount of time before actually occurring, golfers can choose which club to use, how far to hit, and where to attempt to place the ball on the course to set up for the next hole. One new method that we propose is the risk-adjusted shot selection path. Given a hole's mapping, weather data, and a player's ShotLink data, there will exist an optimal path (or multiple paths) from the tee to the hole given the par constraint and the specific player's abilities. This path may or may not minimize distance to the hole and "exposure" or increased probability of hitting into a hazard. For example, Tiger Woods has the best known long game of all time. Such a player might choose to simply gain as much distance as possible on the driving shot, which will limit choices for the following shot. Another player with better accuracy but a worse long game may be better off aiming close to a bunker if this minimizes distance to a hole. The deviation between the optimal path or paths and the actual path taken would be captured by the risk-adjusted shot selection path. Higher deviation, captured by Euclidean distance from the optimal path, would imply higher risk.

Because individual golfers may only play a hole once in their career, the data over which we predict the optimal path could be quite sparse, leading to worse predictions and possibly unfairly penalizing players on certain holes. One variant of this methodology is the average path deviation, which aggregates information on the path that the average player chose to take from tee to hole. A player that wildly deviates from the average path with a lower score might have found a more creative (or possibly riskier) way to play a given hole.

3.4 Rules of the Game

Golf is often categorized as a slow sport. In the same spirit as Major League Baseball, which is now testing a 20-second pitch clock in spring training games [15], we propose a "windup time" metric that clocks how long players take to complete practice swings, stretch before swinging, and walk from hole to hole. The specific time limits would need to be learned from the official data, but any decrease in the average 4.5 hour [16] round would represent an improvement for the sport. Television ratings have declined by almost 50% for the PGA Championship, and 25% for the Master's since 1995 [17]. Some of this decline may be inherent to the nature of golf and the narratives surrounding the game. Tiger Woods' presence, especially his much-anticipated return to the US Master's in 2010, is highly correlated with viewership. A loss of young, household-name superstars following Woods may accelerate these declines in tandem with the slowness of the game. Nonetheless, the exponential rise of basketball suggests that fast-paced games are capturing the most attention in the sports landscape. The PGA will need to adapt, or else risk continued loss of viewership and fans.

4 Conclusion

Choosing a sport of interest outside of the five that we studied in class did not prove to be too difficult. Researching and studying golf significantly changed our perception of the sport, which seems simple on a surface level, and helped us to understand the analytics behind each stroke. Golf's worldwide professional presence and the emergence of the ShotLink system allowed us to evaluate its statistical methods with more rigor. We first analyzed the pure score, noting that not each score is the product of shots of equal difficulty. Furthermore, we looked at the use of course specific models in addition to player specific ones, to measure the difficulty of the shots in relation to the terrain and hazards of each course. Regarding the importance of putting, we looked at studies done by Schneider and Broadie which argued against the importance of putting and rather emphasizing the rest of the shots prior to the put. We also evaluated information presented by Leahy which used ShotLink data to attempt to predict the performance of PGA players with respect to certain variables.

Golf is gaining traction among bettors, with live betting driving interest. Monetary prizes are also relatively high for those who finish in the top three of tournaments. Future work can be done in this department by increasing transparency for tournament payouts. Most performance data is already being captured in some form, which is something that we as a group were surprised by when researching this topic.

Golf has not been a sport that is thought of as analytically sophisticated. However, its metrics are evolving. In addition to the change in payout transparency, our propositions for future work included quantifying the effect of club elasticity, and the introduction of an optimal path specific to certain players' strengths. The decrease in popularity of the sport in the last couple decades has been steady despite the introduction of analytical techniques, and our proposition to include a "windup time" similar to the MLB pitch-clock or the NBA shot clock is in hopes to recapture some of the sports viewership market. The texts that we studied gave us a new perception on the sport of golf, as we were unaware of the amount of high level methods used to gather and analyze data. The statistical techniques used by golf writers provided valuable insights and helped casual viewers understand some of the more advanced topics. Our hopes are that golf can reclaim some of the market that it once had and provide sports watchers with a product well designed for the movement towards faster paced sports.

References

[1] Heath, Elliot. ”How Many Golf Courses Are There In the World?” Golf Weekly. 25 June 2019.

[2] National Golf Foundation. *2019 Golf Industry Report Overview.* April 2019.

[3] Mordor Intelligence, LLP. *Golf Equipment Market Growth, Trends and Forecasts (2019 - 2024).* June 2019.

[4] PGA Tour. ”Strokes Gained: How It Works” Golf Weekly. 30 May 2016.

[5] Broadie, Mark. *Assessing Golfer Performance on the PGA TOUR.* 42(2). *Interfaces.* Feb. 2011.

[6] Minton, Roland. *ShotLink and Consistency in Golf.*

[7] Stöckl, Lamb, Lames. *Performance analysis in golf using the ISOPAR method.* 11(1). *International Journal of Performance Analysis in Sport.* 2011.

[8] Schneider, Todd. ”This Is How You Master The Masters” FiveThirtyEight. 05 April 2017.

[9] Broadie, Mark. *Golf Analytics : A Random Putting Model and its Applications to Optimal Targeting Strategy and Attribution Analysis.* 2015.

[10] Leahy, Brian. *Predicting Professional Golfer Performance Using Proprietary PGA Tour “Shotlink” Data.* 2014.

[11] Hennessey, Powers. ”How To Bet On Golf Legally” The Loop. 03 Oct. 2018.

[12] Scully, Gerald. *The Distribution of Performance and Earnings in a Prize Economy.* 3(08). *Journal of Sports Economics.* 2002.

[13] USGA. *The Rules of Golf for 2019.*

[14] USGA. *Procedure for Measuring the Flexibility of a Golf Clubhead.* 01 May 2008.

[15] Bogage, Jacob. ”MLB reportedly offers to postpone pitch clock until 2022” Washington Post. 27 Feb. 2019.

[16] Hoggard, Rex. ”Don’t single out players: Slow play a bigger problem” Golf Channel. 06 June 2018.

[17] Sports Media Watch. *Major Golf TV Ratings History.* December 2019.