2022

Strokes Gained Analysis of Professional Golfers

Benjamin Allen
Syracuse University

Strokes Gained Analysis of Professional Golfers

A Thesis Submitted in Partial Fulfillment of the
Requirements of the Renée Crown University Honors Program at
Syracuse University

Benjamin Allen

Candidate for Bachelor of Science
and Renée Crown University Honors
May 2022

Honors Thesis in Sport Analytics

Thesis Advisor: _____
Dr. Rodney Paul

Thesis Reader: _____
Dr. Justin Ehrlich

Honors Director: _____
Dr. Danielle Smith, Director

Abstract

Round by round performance data from several international golf tours was collected to investigate a variety of analytical disciplines in professional golf, the central of which seeks to apply strokes gained methodology to golfers in our sample. Strokes gained is a statistical measure that aggregates a player's execution of a golf shot based on the player's lie and proximity to the hole after the shot. Exploratory data analysis included visualizing the relationship between each strokes gained skill category (off the tee, approach, around the green, and putting) and total strokes gained. This initial modeling of relative magnitude was applied to professional golfer Jon Rahm, which leads directly into a broader discussion about which types of golfers, as defined by strokes gained skill categories, are more likely have sustained success in professional golf.

To begin evaluating performance, a variety of logistic regressions were run in connection with the binary variable "won a golf tournament" to investigate these returns to skill. Improvements in strokes gained driving proved to have the largest connection with winning, followed by putting second, around the green third, and approach play last. Through the application of aggregated individual player performance data, k-means clustering was used to type each player into one of four clusters. Paired with the logistic regression interpretation, these results allow anyone interested in this field of study to investigate the career path of any golfer from 2017-2021, based on the criteria described.

Executive Summary

It is well documented that the ability to drive the ball far and straight is an essential part of having success throughout all levels of professional and amateur golf. Driving distance (average length of a tee shot) and fairway percentage (percent of tee shots that come to rest in the fairway) have been proven to be the most top heavy of all the skill categories in golf, meaning an elite driver of the golf ball is more likely to be a better overall player than, say, the best iron player or putter. This basic concept makes sense to the casual golf fan, as the game seems much easier if given the ability to hit shorter shots into the green. In this project, I will begin to test the legitimacy of this idea by evaluating the extent to which driving impacts a golfer's ability to win a tournament when compared to approach shots, around the green shots, and putting. The statistics used in this analysis will be the strokes gained equivalent of these skill categories, and it will be conducted using both logistic regression and k-means clustering, both of which are common machine learning techniques used in all applications of data science. In the end, I argue that strokes gained off the tee is the most influential skill category in golf, which in turn means that it has 1.) more predictive power over the other aspects of the game and 2.) more individual value. The mathematics behind the strokes gained formula and the subsequent modeling techniques used to examine skill category differences in professional golf will be examined in a latter portion of this essay. For now, I will focus on the broader scope of this project, as it relates to its placement within the field of sport, or more specifically golf analytics.

In researching the development of data analytics within the game of golf, far less literature exists in this context than that of the big four United States Sports (NFL, NBA, MLB, NHL). With that being said, this project gives me the unique opportunity to explore a relatively niche section of sport analytics by applying my passion of golf with my coursework as a

Syracuse University undergraduate. In addition, throughout this project, I have had the privilege of speaking to several PGA Tour player performance experts from prominent brands like Callaway and TaylorMade and all of them referenced the inevitable increase for talent of this nature in the job market. Overall, the conclusions I was able to draw from this thesis, not only provided an official academic setting to produce research I was interested in but paved a clear pathway toward future professional development in a field I find incredibly interesting.

Table of Contents

Acknowledgements

I would like to thank Dr. Rodney Paul, Dr. Jeremy Losak, Dr. Justin Ehrlich, and Professor Rick Burton for all they have done in advancing my professional development as an undergraduate at Syracuse University.

Introduction

Introduced to the world by Mark Broadie's groundbreaking paper *Assessing Golfer Performance on the PGA Tour* & subsequent 2014 book *Every Shot Counts*, strokes gained is widely accepted as the most principled and consistent way of measuring a golfer's ability in the modern era[1]. Despite having slightly different interpretations throughout every aspect of the game (driving, approach, short game, putting), the basic purpose is this: to quantify the number of strokes a player gains or loses from each shot they take, using four input values – the lie and distance from the hole prior to the shot & the lie and distance from the hole after the shot. For example, a shot from the fairway one hundred yards out that ends up ten feet from the hole is worth approximately +.23 strokes, while a shot from the same distance/lie that ends up 10 yards from the hole in a bunker is worth about – .61. On the PGA Tour, these values can change based on course difficulty and field performance, but the concept remains the same – each shot has an easily interpretable value that corresponds to the success of the shot. Over the course of a round, tournament, or season, these values are aggregated in the form of a telescoping sum, resulting in a cut and dry value that shows how many shots a player is gaining or losing relative to the field. Within this measure of total strokes gained, the four individual skill categories listed above (off the tee, around the green, approach, putting) can be evaluated separately, allowing players, coaches, and golf fans of any kind to the types of shots where a golfer is losing the most strokes. Despite the reliable baseline that comes with the strokes gained measure alone, there are certainly plenty of other factors that could be included in the strokes gained formula to potentially yield more accurate results. As such, this analysis will begin by exploring the

1) [1] Broadie, Mark (2011) Assessing Golfer Performance on the PGA Tour. Columbia University Graduate School of Business.

academic research related to just a few of these external factors alone, before getting into more comprehensive studies that incorporate a variety of visualization and modeling techniques. In any case, the simple form strokes gained equation is shown in figure 1, as a reminder of the relationship between individual skill categories and total strokes gained.

Figure 1:

$$\sum_{i=1}^{m} g_i = \sum_{i \in \mathcal{L}} g_i + \sum_{i \in \mathcal{S}} g_i + \sum_{i \in \mathcal{P}} g_i$$

In this equation, there are only three skill categories, as the sum of "L" (long game) is an aggregation of approach and driving. This factor will be evaluating separately in this paper

Literature Review

a) Excess Factors Influencing Golfer Performance

There are unlimited "excess" factors that could potentially impact golfers in any given scenario. The following ideas represent literature in this field of study.

As mentioned in the introduction, Mark Broadie is universally credited with the creation of strokes gained; however, he has also explored several other factors that can also be used to better understand the factors that lead to success on the course. In this line of study, his most prominent research has to do with the mental aspect of golf, also known as "the mental game." In this context, the mental game can be defined as the effect that positive and negative factors have on the decision making and subsequent performance of a golfer in any given scenario[2]. Oftentimes, this term is used in connection with a golfer's ability to handle pressure and rebound from bad shots. With this thought in mind, Broadie found that one's ability to control physiological functions like blood pressure and heart rate correspond heavily to increased performance under these sorts of conditions. As a result, many professionals today uses whoop bands to monitor these functions. This of course all relates back to potential alterations of strokes gained, as it's a perfect example of a time when strokes gained alone does not tell the entire story.

Digging deeper into the idea that certain shot conditions are naturally tougher than others, Will and Matt Courchene[3] investigate the possibility that golfer performance on the first tee of

---

[2] S. Chupaska (2020) *Q & A with Golf Analytics Expert Mark Broadie on the Future of Data in Sport*. Data Golf. Retrieved from https://www8.gsb.columbia.edu/articles/ideas-work/qa-golf-analytics-expert-mark-broadie-future-data-sports

[3] Courchene, W., & Courchene, M. (2016). *First Tee Jitters.* Data Golf. Retrieved from http://datagolfblogs.ca/first-tee-jitters/

any given round is significantly less than that of their performance on the same hole when playing it as their 10th hole of the day. This type of analysis makes sense given that generally half the field will play starting on the front nine, while the other half will start playing the back nine, essentially providing a 'natural experiment' to test this sort of disparity. In looking at this research question, they found that the PGA tour average dispersion from the middle of the fairway for any given player was about 8 percent greater when the tee ball was on a '1st tee' hole. Again, it's these small differences (in this case, the order in which the holes are played) that must be analyzed in order to retrieve the most accurate performance metrics possible.

A separate article by Will and Matt Courchene[4] simply looks at randomness as a means for better performance during a short period of time. By repeatedly sampling 50 round moving averages, the consensus is that there is "no persistence" in golf, which essentially means that it is not uncommon for players to get lucky or unlucky without improving their game. It is a bit of a catch 22 because logically, a player who consistently shoots lower scores should be considered "better" by definition, but it just so happens that the nature of golf allows this sort of randomness to occur. The general conclusion within this aspect of the game is that in the same sense that randomness can benefit a player, it can also hurt them undeservingly, leading unwarranted praise

[4] Courchene, W., & Courchene, M. (2019). *Golf is really, really Random!* Data Golf. Retrieved from https://datagolf.com/betting-blog-week5

or uncalled for slumps[5]. This is also one reason why coaching changes usually prompt better performance, as oftentimes a player is simply due for a few breaks to go his or her way.

Dating back to a more formal way of analyzing player skill variations in addition to strokes gained, we will now consider the *format* or setting in which the golf is being played. Although most shots will be taken in a normal stroke play tournament setting, seeing the potential differences that formats like sudden death playoffs or match play matches could certainly yield valuable insight. This exact question is answered in one study by the Journal of Human Sport and Exercise[6]. Specifically, the author walks through a full predictive modeling looking at the relative success/failure of players depending on the format they play (stroke play, match play, scramble). In conducting this study, researchers found that the following factors all increased when playing in formats that "most mimic a standard round of golf:" 1 putt rate, greens in regulation, proximity to the hole. What does this mean exactly? Well, some players are significantly worse off in settings like match play and scramble when compared to average PGA Tour play. As one might expect, this kind of information would certainly be valuable to say a Ryder Cup captain choosing his last few spots on the team or a sharp bettor analyzing odds for a match play event.

[5] Courchene, W., & Courchene, M. (2019). *Analyzing Coaching Changed on the PGA Tour.* Data Golf. Retrieved from https://datagolf.com/does-experience-matter-at-augusta

[6] Suzuki, T., Okuda, I., & Ichikawa, D. (2018). *Investigating factors that improve golf scores by comparing statistics of amateur golfers in repeat scramble strokes and one-ball conditions.* Journal of Human Sport and Exercise. Retrieved fromhttps://doi.org/10.14198/jhse.2021.164.09

b) Visualizing Shot Difficulty

Now that several excess factors affecting golfer performance have been introduced, I will now provide example studies that rely heavily on visualization the difficulty of any given shot.

This discussion leads into a highly touted series of projects produced by members of the International Journal of Golf Science[7]. One study in particular looks heavily into the idea of classifying the physical difficulty of certain golf shots beyond traditional "distance from the hole" criteria. The basic research question is this: While normal strokes gained measures do consider the lie of the golf ball (rough, fairway, sand) and the result of the shot based on distance from the hole, there must certainly be a way to take into account the external factors of a golf hole that make any given shot easier or harder than average. Accounting for a variety of hole tracking factors (slope, topography, etc.) through a tracking program called ISOPAR, researchers were able to configure several models and visualizations that re-measured and explained the success of players during the 2012 PGA tour season. While the exact methodology that the IJGS uses in this scenario may be beyond the scope of this analysis, the fact remains that utilizing detailed course mapping tools is essential in studies of this kind.

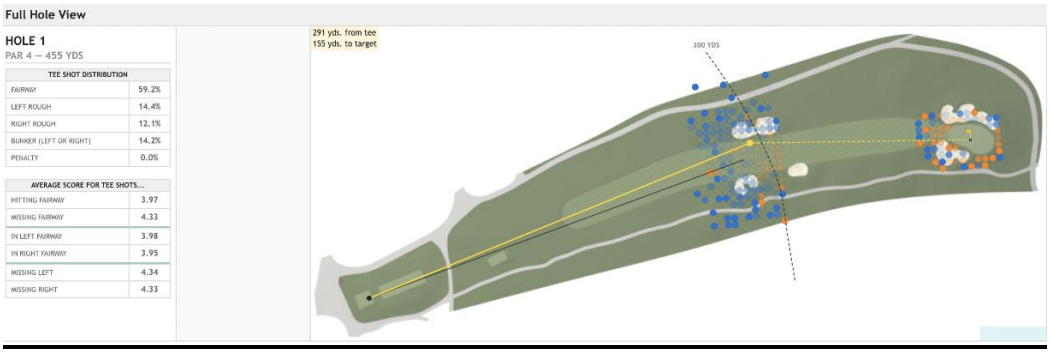
Another example would be the full course heat map of Torrey Pines South displayed in a separate data golf[8] article. In this piece, shot expectancy, shot count, and average shots to the

---

[7]Stockl, M., Lamb, P., & Lames, M. (2012). *A model for visualizing difficulty in golf and subsequent performance rankings on the PGA Tour.* International Journal of Golf Science. Retrieved from https://www.golfsciencejournal.org/api/v1/articles/4947-a-model-for-visualizing-difficulty-in-golf-and-subsequent-performance-rankings-on-the-pga-tour.pdf.

hole are the factors the user can see upon dragging the cursor to different parts of the map. This tool can be used to examine all eighteen holes with four years of Shot Link tracking data backing its legitimacy. (Figure 2 shows example shot). Although not a standard analytical study, this data golf article gives a very insightful look at the detail, effort, and data that goes into professional geocoding and aesthetic mapping. Overall, the ability to combine certain types of visualizations with traditional strokes gained analysis is sure to prove valuable if data of this nature is readily available.

Figure 2



[8] Courchene, W., & Courchene, M. (2021). *Hole Mapping.* Data Golf. Retrieved from https://datagolf.com/hole-heatmaps

c) SG Analysis – Performance Modeling

Upon visualizing some of the criteria listed in the previous sections, I will now move on to a more technical analysis of individual golfers. The following are examples of common practices of this kind as it relates to general performance modeling.

The first and most popular debate in the golf world today concerns the tradeoff between distance and accuracy of the tee. Traditionalists tend to believe that fairways should be valued more so than they currently are, while modern observers of the game understand the distinct statistical advantage that comes with driving the ball as far as possible. As such, analysts at data golf[9] dove headfirst into this idea by theorizing that there must be some value where chasing distance ends up hurting the player as a result of declining accuracy. To begin testing this theory, they modeled the strokes gained equivalent of 10 extra yards of distance off the tee versus that of 5% more fairways (average tradeoff for 10 yards of distance is 5%). They tested this idea in a variety of ways with several accompanying visualizations and came up with a few key takeaways. In recent years, 10 extra yards of distance proved to be much more beneficial for professional golfers by almost double. In addition, they found a strong correlation between increased distance and increased strokes gained approach performance, meaning the traditionalist attitude is likely outdated based on current literature of this kind.

The distance debate concept then leads into a separate player analysis article by the data golf team, which provided a full-scale analysis of Bryson DeChambeau's performance

[9] Courchene, W., & Courchene, M. (2020). *How important is driving distance on the PGA Tour?* Data Golf. Retrieved from https://datagolf.com/importance-of-driving-distance.

throughout the 2018-2019 season[10]. In looking at Bryson's tournament performance data through a variety of forms (interactive visuals with 50 round moving averages for the different strokes gained categories), Will and Matt Courchene were able to easily contextualize his performance historically speaking. Overall, they realized that his sustained excellence puts his career path close to that of Hideki Matsuyama and Dustin Johnson, a comparison that is looking quite accurate two years in the future. Although slightly different than a traditional wide-reaching study that utilizes data over much larger periods of time, this type of individual predictive analysis can be useful if paired with the right player. John Rahm's performance before and after his switch to Callaway immediately come to mind as an application of this concept to modern times.

Another common methodology that is commonly paired with the strokes gained statistic is none other than age curve modeling. As one might expect, the data golf team commonly uses age curve modeling to predict the career path of certain PGA Tour players. One study[11] from 2018 came to the realization that golfer age curves are fundamentally different than that of most other sports. In football or basketball, we can reasonably assume that a 24-year-old is entering the athletic prime of his career, thus his value is likely to increase in the coming years. In golf, oftentimes the opposite happens, as it is widely accepted that players can very easily struggle early in their career and find great consistency in their 30s and 40s. In any case, this model

[10] Courchene, W., & Courchene, M. (2019). *How Good Is Bryson Dechambeau?.* Data Golf. Retrieved from https://datagolf.com/betting-blog-week5

[11] Courchene, W., & Courchene, M. (2018). *Predicting the Career Trajectories.* Data Golf. Retrieved from https://datagolf.com/projecting-careers-blog/

predicted a few key downhill career arcs including the dissent of Rickie Fowler and the rise of Hideki Matsuyama.

In one article from 2017, researchers from MLKE[12] dive into the relationship between swing biomechanics and subsequent performance. To conduct this study, they analyzed factors like clubhead speed, flight pattern, dynamic loft, swing plane & face angle at impact, all of which were measured using 3D tracking and Doppler radar technology. With the help of random forests, decision trees, and other machine learning techniques, they found that the factors with the highest degree of predictive accuracy in relation with player skill were face angle, dynamic loft, and clubhead speed. The basic interpretation and actionable reasoning of this model is that Professionals hit down harder on the ball with higher speeds and unbelievable consistency, a concept that golf fans of all ages should be able to grasp. Like the excess factors list from the beginning section, there are basically unlimited ways to approach any given strokes gained analysis, and by testing several different techniques through the lens of performance modeling, a variety of meaningful insights are bound to take shape.

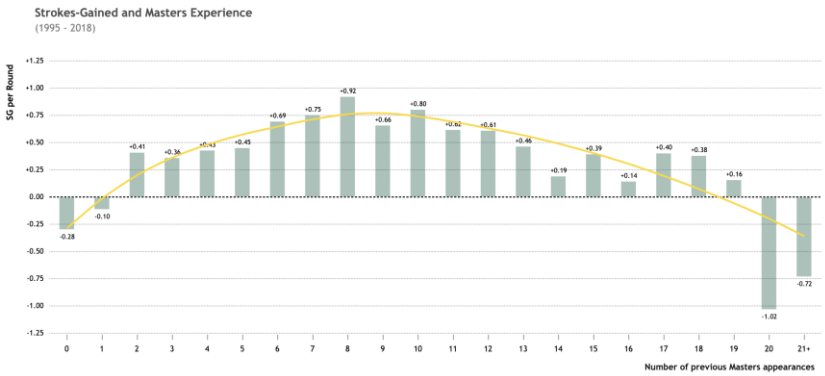
[12] König R., Johansson U., Riveiro M., Brattberg P. (2017) *Modeling Golf Player Skill Using Machine Learning.* In: Holzinger A., Kieseberg P., Tjoa A., Weippl E. (eds) Machine Learning and Knowledge Extraction. CD-MAKE 2017. Lecture Notes in Computer Science, vol 10410. Springer, Cham. https://doi.org/10.1007/978-3-319-66808-6_19

d) SG Analysis – Course Fit

Looking at strokes gained factors using course fit as the base is another common approach in evaluating the performance of professional golfers:

Golf commentators and pundits often tout that experience on certain courses is much more important than others. On a course like Augusta National for example, a certain base of knowledge is required to have great success in any given tournament, hence the fact that very few golfers win or contend during their first few times at Augusta National. In this article[13], Will and Matt Courchene analyze the extent to which these factors play a part at not only Augusta National, but around several top courses on the PGA Tour. In the end, the results suggest that for the most part, this idea about Augusta National holds true, as the actual vs expected strokes gained chart increases heavily as the 'years of experience' variable increases. I would imagine the trend shown in the image below would continue for a variety of courses throughout the world.

(Figure 3)



1. [13] Courchene, W., & Courchene, M. (2019). *How Important is Experience at Augusta National*. Data Golf. Retrieved from https://datagolf.com/does-experience-matter-at-augusta

In a similar sense, data golf also looked at the concept of "course-fit" through the lens of estimated fairway cost (i.e. How important is it to hit the fairway at any given course compared to average?). By running an algorithm with a sample that includes every PGA Tour event, the results yield a more improved way of measuring how a player will perform on any given course given his strengths or weaknesses. This article also introduces and attempts to account for the "randomness" of penalty shots off the tee, which in a four-day tournament heavily impacts a golfer's strokes gained measure in the short term. In addition, the visualizations and ranking system utilized in this piece are practical and easy to understand, which is important to keep in mind, given the complexity of the problem at hand.

e) Applications to Sports Betting

As it relates to brand value, betting odds are a good way to judge public perception. This section gives a brief overview of just a few outside the box ways to analyze golf betting.

One golf.com article details the intricacies of the data golf betting model, with the most valuable information being the potential predictors in addition to the types of bets that are most likely to yield winning results. For this model, they found that "top 20" bets have historically yielded better results as opposed to longshot winners or top 5 bets. Although this may vary by model, it is good to begin looking at the testing and application procedures that occur when going through this type of analysis.

Along the same lines, a separate data golf[14] article looks at different ways to quantify different factors related to golf betting and betting models overall. In doing this, they look at predicted probability charts, ROI analysis, & several other correlating measures, eventually gathering some interesting takeaways about betting sites like Pinnacle and DraftKings. These insights can be boiled down to the following statement: DraftKings odds move much less as the game time approaches, whereas Pinnacle and bet365 experience much more drastic shifts in odds and cashflow. This could signify a few different things, but most notably, that these sites have much different bookmaking strategies, which oftentimes correlates heavily to their use of strokes gained as a predictive measure of success. In professional golf, exploratory data analysis into individual players is common and most definitely valuable in its own right; however, it is arguably more important to understand the fundamental ways in which these data are used by

[14] Courchene, W., & Courchene, M. (2020). *How Sharp are Bookmakers.* Data Golf. Retrieved from https://datagolf.com/analyzing-betting-odds

betting companies at an infinitely larger scale. Although the identification of market inefficiencies through strokes gained is not a central portion of my analysis, there will always be a practical application to sports betting when analyzing any econometric principle such as this.

From general strokes gained measures to advanced modeling techniques to interactive course mapping visualizations, there are so many ways to look at the modern game of golf that simply didn't exist even 10 years ago. However, there are certainly a few constants that became clear throughout the course of this initial research. For one, course mapping query tools are widely regarded as the most efficient way to display strokes gained data in comparing shot impact/difficulty. In addition, at least 12-15 rounds must be included when looking at any small-scale performance change, as randomness is simply too large of a factor in golf compared to other sports. Lastly, at least some of the "excess" factors affecting performance must be included in conjunction with traditional strokes gained analysis if a proper career projection is expected to be achieved.

Methodology

a) Data Summary

The dataset used in this analysis was scraped from datagolf.com's historical strokes gained database, which consolidates the tracking capabilities of every available professional tour into one clean platform. As a result, this dataset includes round by round data from every fully tracked round throughout every professional tour from 2017-2021. By "fully tracked", I mean that each individual strokes gained category (putting, around the green, approach, driving) was included in addition to total strokes gained. This distinction warrants mentioning because many international and non-PGA Tour events had to be excluded from analysis because of these null values. In any case, my general reasoning for using strokes gained data to evaluate performance without the inclusion of other common stats like fairways, greens, putts, etc. is the vast array of evidence supporting the efficiency of strokes gained models, based on their low AIC values (Courchene 2018)[15]. Because understanding the fundamental nature of strokes gained is essential to understanding the practical application of my models, I've provided the following table that gives a brief performance analysis of the top 10 overall golfers in the dataset, based on mean strokes gained per round.

[15] Courchene, W., & Courchene, M. (2019). Datagolf Predictive Power Methodology, Retrieved from https://datagolf.com/predictive-model-methodology

Top 10 PGA Tour Golfers by Average Strokes Gained Per Round

Reminder: 'Total Strokes Gained' is the Sum of Each 'Strokes Gained' Skill Category

	Player Name	Total	Putting	Around the Green	Approach	Driving
1	Rahm, Jon	2.17	0.42	0.23	0.72	0.79
2	Fitzpatrick, Matthew	2.05	0.99	0.47	-0.09	0.66
3	Cantlay, Patrick	1.95	0.46	0.28	0.57	0.62
4	Oosthuizen, Louis	1.65	0.75	0.31	0.54	0.05
5	Casey, Paul	1.65	0.01	0.21	1.13	0.30
6	Pendrith, Taylor	1.64	0.50	-0.17	-0.56	1.88
7	DeChambeau, Bryson	1.56	0.45	-0.19	0.21	1.10
8	Migliozzi, Guido	1.53	0.46	0.14	0.75	0.18
9	Spieth, Jordan	1.45	0.46	0.38	0.60	-0.01
10	Ancer, Abraham	1.40	0.46	-0.06	0.64	0.36

**Some International Rounds are Excluded based on SG Tracking Capabilities*

(Table 1)

Based on the description of strokes gained described in the introduction, you'll remember that a general summary statistics table gives little to know value to the fitness and overall structure of the data, as each value is a telescoping sum that revolves around zero. Because of this concept, I've provided a bottom-ended sample that reiterates the application of the strokes gained metrics by displaying summary statistics for the bottom 100 rounds on the PGA Tour.

Summary Stats for the Bottom 100 Rounds on the PGA TOUR (2017–2021)

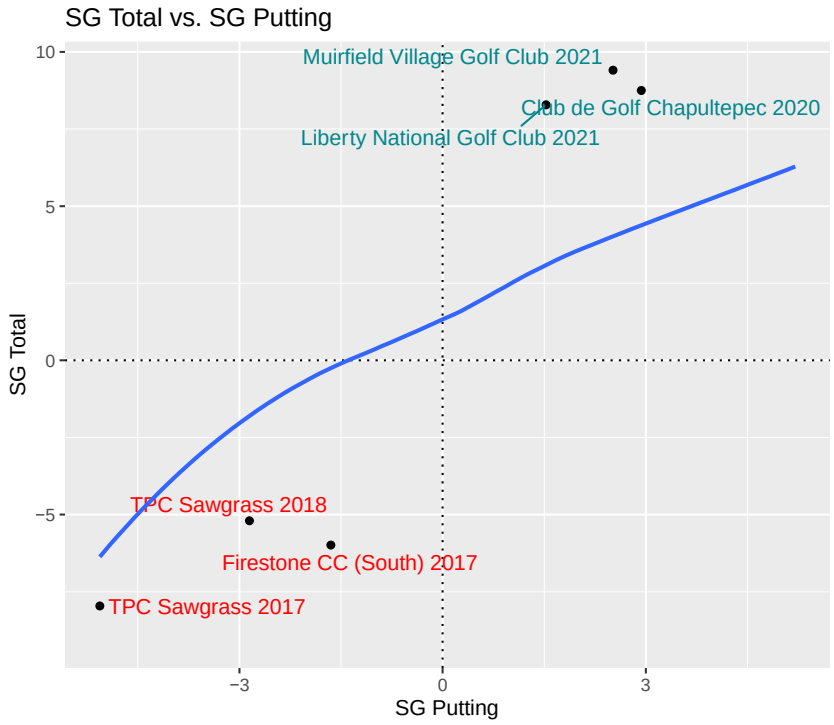
	vars	n	mean	sd	min	max	range	se
Round_Score	1	100	83.27	2.265	79	92	13	0.226
SG_Putt	2	100	–2.689	1.835	–8.884	1.927	10.811	0.183
SG_Around_the_Green	3	100	–2.18	1.875	–8.12	1.42	9.54	0.188
SG_Approach	4	100	–4.086	2.034	–9.859	1.527	11.386	0.203
SG_Driving	5	100	–2.891	2.423	–9.995	0.946	10.941	0.242
SG_Tee_To_Green	6	100	–9.151	2.371	–18.297	–1.884	16.413	0.237
SG_Total	7	100	–11.841	1.563	–19.172	–10.432	8.74	0.156

(Table 2). **In all contexts, "SG" stands for "Strokes Gained"*

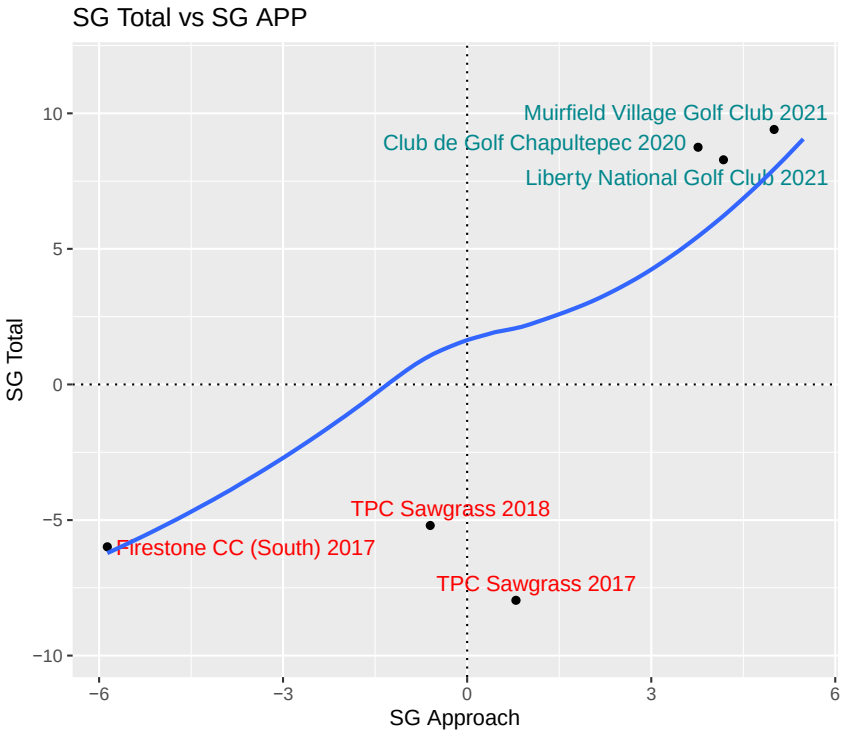
b.) Exploratory Data Analysis

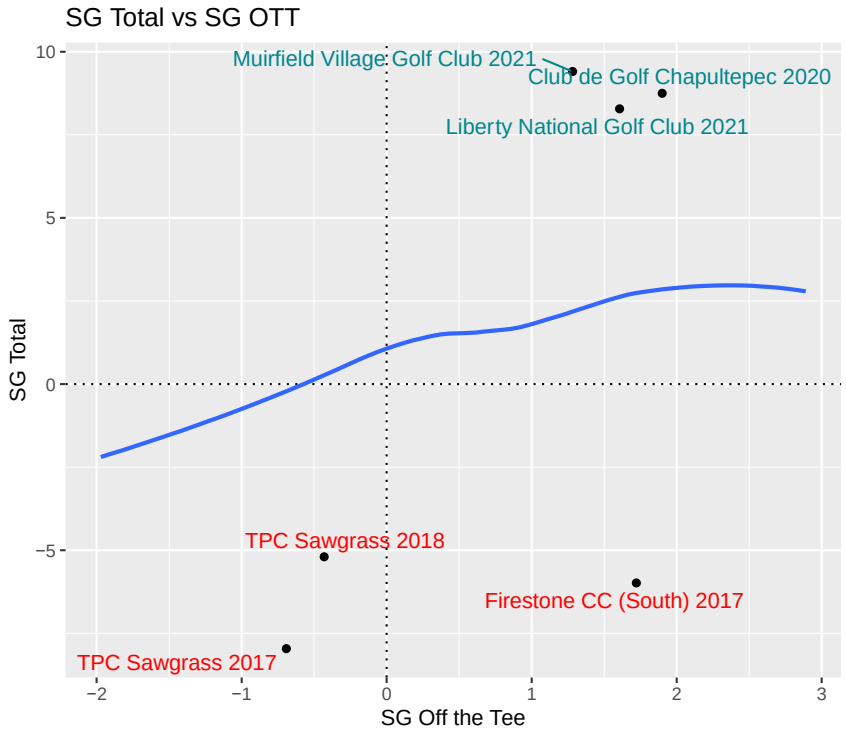
Now that I have described the data methodology, introduced the reasoning behind a stroke gained model, and provided summary statistics in a variety of contexts, I'll now provide a visual application of these figures to preface my modeling. The following set of visuals shed light onto the range and relative importance of each skill category compared to total strokes gained for world number two golfer, Jon Rahm. Each point in the graph represents a different round from Rahm's career & the resulting trend line from each skill category helps us understand where Rahm's strengths, weaknesses, and consistency lies relative to his overall performance.

(Figures 3 – 6)




SG Total vs. SG ARG
Muirfield Village Golf Club 2021
Club de Golf Chapultepec 2020
Liberty National Golf Club 2021
TPC Sawgrass 2018
Firestone CC (South) 2017
TPC Sawgrass 2017
SG Total
SG Around the Green
10
5
0
−5
−2
0
2


SG Total vs SG APP
Muirfield Village Golf Club 2021
Club de Golf Chapultepec 2020
Liberty National Golf Club 2021
TPC Sawgrass 2018
Firestone CC (South) 2017
TPC Sawgrass 2017
SG Total
SG Approach
10
5
0
−5
−10
−6
−3
0
3
6



In analyzing these graphics, it is apparent that Rahm is exceptionally consistent off the tee and around the green, while proving to be slightly erratic approaching the green and putting. As we know from Will Courchene's predictive power methodology, we expect putting to be the most variable skill in golf, so that insight is not too surprising (Courchene 2018)[16]. However, if we were to investigate the returns to skill of these categories, perhaps we would be able to value each player differently based on the way each individual player gains his or her strokes. Overall, it is important to realize that strokes gained visuals become more valuable when combined with the modeling results to come, which speaks to the practical application of this work in general.

[16] Courchene, W., & Courchene, M. (2019). Datagolf Predictive Power Methodology, Retrieved from https://datagolf.com/predictive-model-methodology

c.) Modeling Methodology

The first modeling technique used in this analysis was a logistic regression that seeks to identify the probability increase in winning probability that results from a 1 stroke increase in a golfer's mean strokes gained for a given tournament. This distinction on using mean strokes gained over aggregate strokes gained warrants some explanation. In this context, the mean strokes gained (MSG) from every player from every tournament was used to evaluate the probability significance in skill categories. In any case, the four skill categories (putting, around the green, approach, off the tee) were tested in comparison with our binary variable "Winning." As touched on in the exploratory analysis, the probability that a player overperforms expectation is largely reliant on his or her ability to improve on their lower-level skill categories, as opposed to their most dominant. In the case of Rahm, we were able to see that putting and approach had the most correlation to tournament success, even though he excels off the tee. The results of this model would build on this sort of insight as it sheds light onto the predictive nature of strokes gained as a whole. A more extensive modeling technique that could potentially yield more interesting results would be a partial proportional odds model that not only includes these strokes gained statistics, but other ordinal variables that provide a more complete picture of a golfer's specific strengths and weaknesses within their skill category. (For example, comparing proximity from 150-175, 176-200, 201-225 within the strokes gained approach skill category)

In part 2, I chose to cluster the athletes in my sample based on these same skills and subsequently evaluate player "type" from our sample. In this model, individual player data was concatenated to create a set of unique golfers for cluster analysis. Cluster analysis and logistic regression work quite well together in this instance, as the clustering results serve as infinite player sample support for the regression conclusions.

Results

a.) Logistic Regression

As mentioned in the method section, the first model is a logistic regression that sheds light onto the impact that each skill category has on a golfer's probability to win a tournament. For our model, **mean strokes gained** of each skill category was used in evaluating play throughout each tournament. This simply means that round-by-round data was compressed into tournament-by-tournament data in order to ensure accurate probabilities. The output of the logit model is as follows:

(Table 3)

Strokes Gained Returns to Skill Model

	Dependent variable:
	Won
avg_putt	2.044***
	(0.123)
avg_arg	1.762***
	(0.163)
avg_app	1.954***
	(0.118)
avg_ott	2.532***
	(0.185)
Constant	-8.963***
	(0.323)
Observations	22,864
Log Likelihood	-488.932
Akaike Inf. Crit.	987.864

To help interpret these coefficients, we can transform the log odds from the output into true odds in the table below:

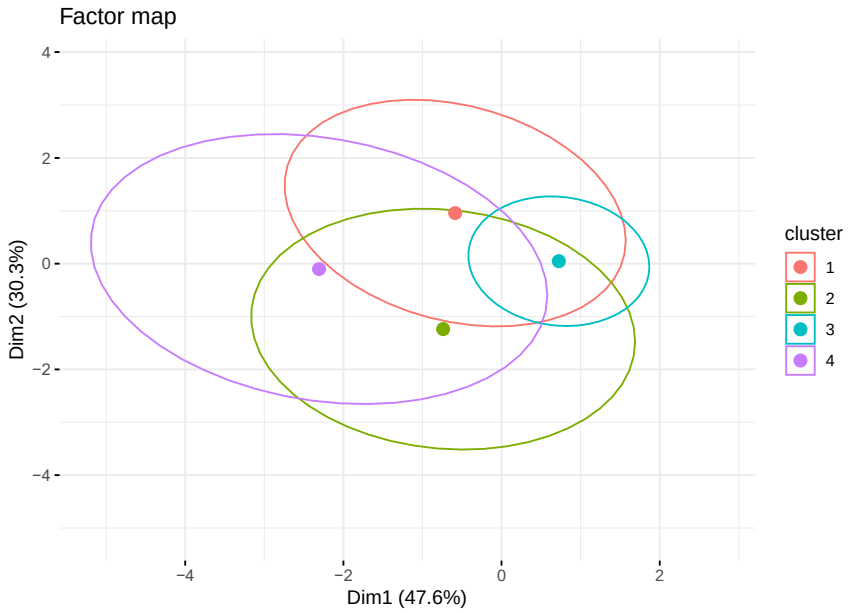
	Transformed Log Odds
1 avg_putt	7.724
2 avg_arg	5.823
3 avg_app	7.055
4 avg_ott	12.580
5 Intercept	0.0001

(Table 4)

As both tables above depict, the value of an additional mean stroke **off the tee** is significantly higher than any other skill category. One interpretation of these results is that high level players who comparatively gain more strokes from driving are more likely to contend in any given golf tournament. This information is useful in a variety of contexts, but especially when looking at everyday changes a golfer makes to his game. In prepping for any given tournament, season, etc., a player may emphasize putting or short game at the potential expense of his ball striking. Because the strokes gained tradeoff supports driving as the more significant factor, the results of this model advise against that sort of thinking.

b.) Clustering

The next model is a k-means clustering algorithm, where each input point represents a different golfer with varying skillsets in terms of how they gain their strokes. Because we determined certain skill categories increase the probability of winning more than others, a logical next step would be to cluster and evaluate golfers that possess similar proficiencies. Upon running the model, we get a plot that depicts our four clusters along with the associated central points.



(Figure 7)

Each colored section represents a different type of golfer based on their most prominent strokes gained skill category. Players like Bryson DeChambeau are in cluster 4 based on their prodigious strength off the tee, while players like Dean Burmester are in cluster 1, based on their putting prowess.

Conclusion

The overarching conclusion to this analysis is that the relationship between a player's "type" and their probability of contending in a golf tournament is closely intertwined. To be specific, all strokes gained in a golf round are not weighted equally, as strokes gained off the tee are significantly more valuable and predictive of performance when compared to all three other skill categories (around the green, putting, and approach) In terms of this study's practical application, it can be said that brands, coaches, and players can more easily predict one's ability to underperform or overperform depending on the means by which they gain their strokes.

Lastly, the long-term value of this project comes not only in the results of my models, but the reader's ability to replicate the modeling and visualization techniques on any golfer in the time frame. For example, as an extension of this analysis, interactive dashboards displaying any given player's cluster and subsequent win probability over time could certainly be refined and developed. Overall, it is my hope to expand this research in ways such as this in order to improve my understanding of strokes gained performance analysis, as I enter the field of sport analytics.

Works Cited

1) Broadie, Mark (2011) Assessing Golfer Performance on the PGA Tour. Columbia University Graduate School of Business.
2) S. Chupaska (2020) *Q & A with Golf Analytics Expert Mark Broadie on the Future of Data in Sport*. Data Golf. Retrieved from https://www8.gsb.columbia.edu/articles/ideas-work/qa-golf-analytics-expert-mark-broadie-future-data-sports
3) Courchene, W., & Courchene, M. (2016). *First Tee Jitters.* Data Golf. Retrieved from http://datagolfblogs.ca/first-tee-jitters/
4) Courchene, W., & Courchene, M. (2019). *Golf is really, really Random!* Data Golf. Retrieved from https://datagolf.com/betting-blog-week5
5) Courchene, W., & Courchene, M. (2019). *Analyzing Coaching Changed on the PGA Tour.* Data Golf. Retrieved from https://datagolf.com/does-experience-matter-at-augusta
6) Suzuki, T., Okuda, I., & Ichikawa, D. (2018). *Investigating factors that improve golf scores by comparing statistics of amateur golfers in repeat scramble strokes and one-ball conditions.* Journal of Human Sport and Exercise. Retrieved fromhttps://doi.org/10.14198/jhse.2021.164.09
7) Stockl, M., Lamb, P., & Lames, M. (2012). *A model for visualizing difficulty in golf and subsequent performance rankings on the PGA Tour.* International Journal of Golf Science. Retrieved from https://www.golfsciencejournal.org/api/v1/articles/4947-a-model-for-visualizing-difficulty-in-golf-and-subsequent-performance-rankings-on-the-pga-tour.pdf.
8) Courchene, W., & Courchene, M. (2021). *Hole Mapping.* Data Golf. Retrieved from https://datagolf.com/holeheatmaps
9) Courchene, W., & Courchene, M. (2020). *How important is driving distance on the PGA Tour?* Data Golf. Retrieved from https://datagolf.com/importance-of-driving-distance.
10) Courchene, W., & Courchene, M. (2019). *How Good Is Bryson Dechambeau?.* Data Golf. Retrieved from https://datagolf.com/betting-blog-week5
11) Courchene, W., & Courchene, M. (2018). *Predicting the Career Trajectories.* Data Golf. Retrieved from https://datagolf.com/projecting-careers-blog/
12) König R., Johansson U., Riveiro M., Brattberg P. (2017) *Modeling Golf Player Skill Using Machine Learning.* In: Holzinger A., Kieseberg P., Tjoa A., Weippl E. (eds) Machine Learning and Knowledge Extraction. CD-MAKE 2017. Lecture Notes in Computer Science, vol 10410. Springer, Cham. https://doi.org/10.1007/978-3-319-66808-6_19
13) Courchene, W., & Courchene, M. (2019). *How Important is Experience at Augusta National.* Data Golf. Retrieved from https://datagolf.com/does-experience-matter-at-augusta
14) Courchene, W., & Courchene, M. (2020). *How Sharp are Bookmakers.* Data Golf. Retrieved from https://datagolf.com/analyzing-betting-odds
15) Courchene, W., & Courchene, M. (2019). Datagolf Predictive Power Methodology, Retrieved from https://datagolf.com/predictive-model-methodology