

An Analysis of Driving on the PGA Tour

Keagan Anderson and Joshua Yap
CSE 163 Final Project
Winter 2020

March 18, 2020

Abstract

Does an increase in driving distance come with a decrease in accuracy? How valuable is accuracy as compared to distance? These are a couple of key questions addressed in the paper. We analyze 2019 PGA Tour driving data to answer them and provide insights into when different driving strategies can work most effectively.

1 Summary of Research Questions and Results

1.1 How does the average driving distance of a player correlate with their driving accuracy percentage?

On average, there is a negative correlation between driving distance and driving accuracy.

1.2 How do different course difficulties (in terms of driving) impact players that fall into different driving profiles?

As course difficulty increases, accuracy becomes more valuable than distance. Short, accurate drivers tend to drive it worse on easy driving courses than long, inaccurate drivers, but the disparity in performance decreases with increasing course difficulty.

1.3 How does knowing the relative driving performance of a certain profile inform a player's strategy going into a tournament?

We can figure out the difficulty of an upcoming course and use the data to determine if an aggressive driving strategy (long, inaccurate) would work better than a conservative one (short, accurate).

2 Motivation and Background

Knowing how accommodating a golf course is to different types of players is interesting because there are so many factors that affect scoring and one cannot tell just by looking at the course which factors will contribute

most. A typical measure of success in golf is score (ie. how many total shots a golfer takes in a round of golf), but this does not tell the player anything about how many shots they took in each area of the game (driving, approach, putting, chipping). As such, a ‘strokes gained’ metric was introduced in 2011. For an explanation of strokes gained see Appendix A.

Our dataset has strokes gained data, so we can use this to determine how well a player drove the ball that week. We are particularly interested in driving because it is an aspect that can be easily strategized: a golfer going into a round knows exactly where he will hit all 18 tee shots from. This cannot be said for the second, third, etc. shots on each hole. One of our team members is an avid golfer and is intrigued by how data can inform golf strategy.

Notes:

- (I) SG: OTT is Strokes Gained: Off-the-Tee and will be the metric used to measure driving performance in this project.
- (II) ‘Performance’ in this paper only measures driving performance, not overall performance of the other parts of the golfer’s game.

3 Dataset

The dataset was pulled from https://www.kaggle.com/bradklassen/pga-tour-20102018-data#2019_data.csv as a csv file. It provides data scraped from the PGA Tour’s website and has 5 columns: Name, Date, Statistic, Variable, Value. The Statistic and Variable columns encode many more columns. For example, under Statistic, there is Driving Distance, Driving Accuracy, Average Putting, etc. The Variable column is an observable of the Statistic, such as Average, Maximum, Total, etc. In total, there are 1449 players, 378 statistics and 1496 variables. It is noteworthy to mention that the value of each variable in the data is cumulative through the season up to that date, so we needed to further process it to get course-level information.

4 Methodology

- A. Make a plot of Driving accuracy against Driving distance.
- B. Define driving profile parameters.
 - (i) A driving profile is a player’s combination of driving accuracy and distance (eg. long_inaccurate, short_accurate, etc).

- (ii) The parameters are threshold values we define to separate players into these profiles.
- C. See how these players perform against the field at different courses.
- (i) For all players in a certain profile, compare their driving performance to the rest of the field at easy courses and hard courses using Strokes Gained: Off-the-Tee (SG: OTT) as the metric.
 - (ii) Compare driving performances by computing a field average of SG: OTT and compare it with the SG: OTT of players in the profile.
 - (iii) ‘Easy’ courses are defined to be ones that have a higher field average SG: OTT, ‘hard’ courses are ones that have lower SG: OTT.
 - (iv) Make a plot of SG: OTT against course difficulty for each driving profile.
 - (v) Combine the plots onto a single axis to compare differences in driving profiles as course difficulty changes. If, for example, long and accurate drivers have a higher SG: OTT average on difficult courses than short and accurate drivers, then we can conclude that long and accurate drivers drive the ball better on hard courses than short and accurate drivers.
- D. Given a course that players are about to play in 2020, determine its difficulty from our 2019 dataset and predict a certain driving profile that is going to produce better relative driving performance.

5 Results

This section goes through our research questions sequentially and answers them with data we have processed.

5.1 How does the average driving distance of a player correlate with their driving accuracy percentage?

Our first research question is answered by the plot in figure 1. There is a negative relationship between the two variables, which is what we expect. One sacrifices accuracy for distance (on average).

5.2 How do different course difficulties (in terms of driving) impact players that fall into different driving profiles?

To answer this question, we first filtered the dataset by players with a specified profile. The possible profiles are ‘Short Accurate’, ‘Short Inaccurate’, ‘Long Accurate’ and ‘Long Inaccurate’. We also defined an easier driving course as one that has a more positive SG: OTT average for the field that week and vice versa. Once



Figure 1: Data of driving accuracy against distance, cumulative through the 2019 season until 08-25-2019. Linear regression shows a negative correlation between the two variables.

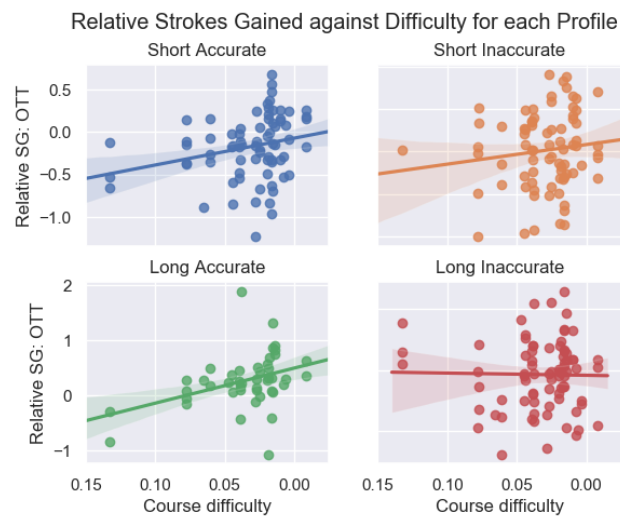


Figure 2: SG: OTT against course difficulty for each profile.

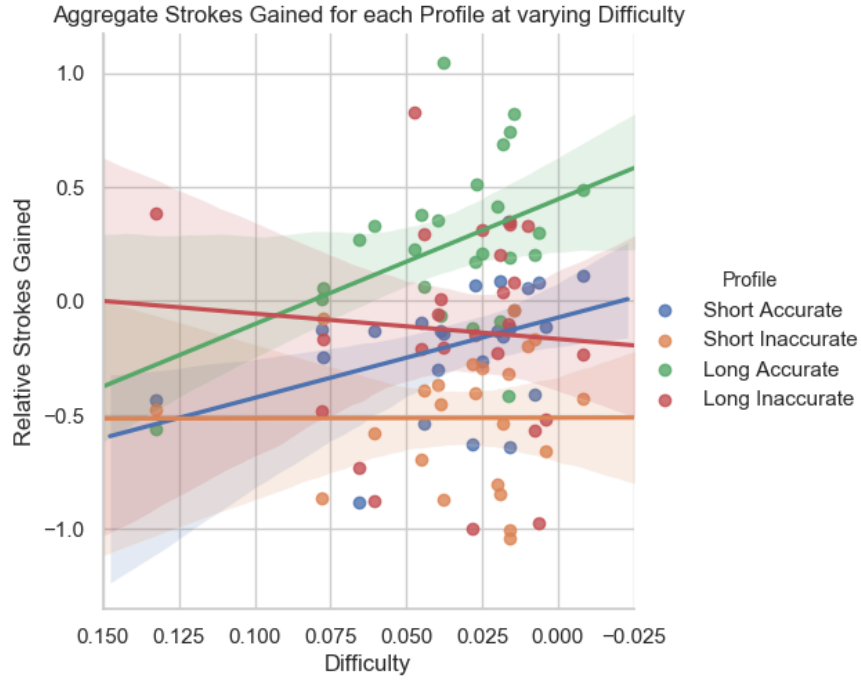


Figure 3: SG: OTT against course difficulty for all profiles overlayed on the same axes. Lines show trends of each profile.

we calculated the difficulty of each course, we made the set of plots in figure 2, which shows the relationship between relative SG: OTT and course difficulty.

At this point, clarifications need to be made. ‘Relative SG: OTT’ is how the players in our profile performed relative to the field average that day. For example, if, for a given course, the field averaged an SG: OTT of 0.1 and our player had an SG: OTT of 0.15, his relative SG: OTT is +0.05. This avoids the problem of producing biased plots since we would expect a player to average a higher SG: OTT at a given course. In our plots, the Difficulty scale goes from positive to negative values. The values are the average SG: OTT for the field, so the left side of the axis (larger values) corresponds to an easier course and the right side (smaller values) corresponds to a harder course.

We can make out trends from the plots, but the data is largely clumped together. Long Accurate drivers have the trend with the least variance: their driving performance improves relative to the field as courses get harder. We then wanted to compare the profiles on the same axes but there were too many data points to make out a trend. We therefore took the average of all players in a profile that played a certain course and plotted it as a single point. This reduced the number of data points while keeping as much information as possible. The resultant plot in figure 3 was produced. The data points look like they are in columns because each course is assigned a difficulty and every profile plays the same courses. The real point of interest here is comparing Long Inaccurate drivers to Short Accurate drivers. We want to see if it is worth giving up

accuracy for distance, and what courses this strategy would be most effective on. For easier courses, the disparity in performance between the two profiles is clear: Long Inaccurate drivers do better. As the difficulty of courses increases, the disparity in SG: OTT decreases which indicates that having more accuracy becomes more valuable, even at the expense of distance. At the highest difficulty of courses in our data, the Short Accurate drivers perform better than Long Inaccurate drivers, but the variance at this point is large enough that we cannot be confident of this result. Nonetheless, the trend of accuracy becoming more valuable for more difficult courses is a reasonable interpretation of the data. To back up that statement, Long Accurate drivers also see an increase in performance against the field at more difficult courses.

5.3 How does knowing the relative driving performance of a certain profile inform a player's strategy going into a tournament?

If they had been playing the PGA Tour over the last weekend, it would have been The Players' Championship at TPC Sawgrass. Based on last year's data, the course difficulty is 0.0157. Our plot suggests that Long Inaccurate drivers would drive the ball just as well as Short Accurate drivers. However, this is also a region where there is a marked difference between Long Accurate drivers and the others. Their driving performance averages about half a stroke better per round than Short Accurate and Long Inaccurate drivers. This is huge in a 4 round tournament.

6 Reproducing Results

In order to produce this statistical analysis on Driving Accuracy and Driving Distance we started with the 2019 PGA Kaggle Dataset. From there we needed to sort the data set into a row of driving distance and a row of driving accuracy. We then merged this data by player and date to get an effective dataset to graph driving accuracy against driving distance. We made a Player class in order to graph individuals driving accuracy against their driving distance. This gets an estimate on types of players we want to analyze moving forward.

From here, we needed a way to calculate course difficulty so we could see how different types of golfers perform on different difficulty courses. We decided it was most effective to judge a course's difficulty by taking the mean Strokes Gained Off the Tee of everyone on the given day, subtract the previous tournament's mean SG: OTT (because the date in the data set is cumulative and we want Strokes Gained only over that specific course/day) and divide that number by 4 because there are normally 4 rounds in a tournament. This results in the mean Strokes Gained per round on a specific course. We then created thresholds to sort golfers by driving profile to compare over the different difficulty courses. The thresholds were decided to spread

the golfers out among 4 categories. Golfers less than 55% accurate and hitting more than 300 yards were long-distance inaccurate. Hitting more than 300 yards with above 65% accuracy sorted golfers into the long-distance accurate category. Less than 60% accuracy and less than 290 yards were short-distanced and inaccurate. And finally golfers who hit it under 270 yards with above 70% accuracy were short-distanced and accurate.

With the ability to score courses on their difficulty and golfers' on their strategy, we filtered the data set so when given a player type, the data set would only contain those profiles. Furthermore it calculated the average relative Strokes Gained by profile (which was calculated by subtracting the player type's average from the overall average on that difficulty course). From there we decided to graph the relative Strokes Gained against each difficulty course, for each profile. We were able to see the patterns and tendencies of certain player types on certain courses such as how long distance-accurate types were above all in easy courses but trended toward the median on more difficult ones, or how long distance inaccurate players performed close to the long distance accurate ones. We also graphed the total Strokes Gained for every profile on every difficulty course to give us an effective performance profile of every type.

7 Work Plan Evaluation

Overall we kept to the work plan loosely. Other assignments came up that we needed to complete, but there was constant work done on the project over time. Additional complexity was added by classes that had format of homework or finals changed due to COVID-19.

8 Testing

We wrote a test file and made 2 smaller datasets to test a few of the functions. Other functions were tested by plotting the data to see if it looked right, and by cross-referencing our output with the PGA Tour's statistics website https://www.pgatour.com/stats/categories.ROTT_INQ.html.

9 Collaboration

Keagan Anderson and Joshua Yap collaborated on this assignment.


Appendices

Appendix A

The strokes gained formula explained

Rickie Fowler made birdie on TPC Sawgrass' 18th hole in the final round of the 2015 PLAYERS Championship before winning the tournament in a playoff. Fowler hit his tee shot 330 yards on the 446-yard, par-4 before sticking his 116-yard approach shot 16 feet, 11 inches from the hole. He then made the birdie putt.

We'll use Fowler's birdie to explain how strokes gained statistics are calculated.

tpc-18-rickiefowler-strokesgained

Tee shot: TPC Sawgrass' 18th hole is a 446-yard, par-4. The PGA TOUR's scoring average, or baseline, on a par-4 of that length is 4.100. Fowler hit his tee shot on No. 18 in the fairway, 116 yards from the hole. The TOUR scoring average from the fairway, 116 yards from the hole, is 2.825. He gained 0.275 strokes on his tee shot. Here's how:

Baseline for tee - Baseline for second shot - 1 = strokes gained: off-the-tee
 $4.100 - 2.825 = 1.275 - 1 = +0.275$

One is subtracted from the difference between the two baselines to account for the shot that Fowler hit.

Approach shot: As noted above, the baseline from 116 yards in the fairway is 2.825. This is the average number of shots it takes a TOUR player to hole out from this distance. Fowler lost 0.001 strokes by hitting his approach shot to 16 feet, 11 inches. The baseline from 16 feet, 11 inches is 1.826. Here's how Fowler's strokes gained: approach-the-green were calculated for the shot.

Baseline for approach shot - baseline for putt - 1 = strokes gained: approach-the-green
 $2.825 - 1.826 = 0.999 - 1 = -0.001$

Putt: Fowler gained 0.826 strokes by making his 16-foot, 11-inch birdie putt at 18. The TOUR baseline for a birdie putt of that length is 1.826, while the baseline for a holed shot is, of course, 0.

Baseline for putt - baseline for putt - 1 = strokes gained: putting
 $1.826 - 0 = 1.826 - 1 = +0.826$

A player's strokes gained statistics for the round are the sum of his strokes gained and lost on all 18 holes. Adjustments are made to account for a course's difficulty.

Shot	Location	Baseline from location	Next location	Baseline from next location	Strokes gained
1	446 yards (tee box)	4.100	116 yards (fairway)	2.825	$(4.100 - 2.825) - 1 = +0.275$
2	116 yards (fairway)	2.825	16' 11" (green)	1.826	$(2.825 - 1.826) - 1 = -0.001$
3	16' 11" (green)	1.826	Hole	0	$(1.826 - 0) - 1 = +0.826$
Total	446 yards (tee box)	4.100	Hole	0 (3 shots)	$4.100 - 3 = +1.100$
Strokes gained: total -- $0.275 + (-0.001) + 0.826 = 1.100$					

Figure 4: Example of Strokes Gained metric from <https://www.pgatour.com/news/2016/05/31/strokes-gained-defined.html>