PGA Tour Player Results vs Statistical Rankings

DAI Captsone I Project Proposal

Submitted by Richy Peterson

Objective

The goal of golf is simple, take as few strokes as possible to complete each hole. On the PGA Tour each tournament consists of four rounds of 18 holes. The tournament winner is the golfer with the fewest total strokes over the four rounds. While there are numerous types shots a player can make, there are four main aspects of golf that the shots fall into based on where the shot is taken from on the hole. Those aspects are driving, approach the green, around the green, and putting. The PGA Tour tracks statistics that represent how well a player is performing in each of these aspects. The objective of this capstone is to determine which aspect has the highest impact on a player's tournament results.

The Data Set

The project will use the "PGA Tour Data" data set found at the following link: https://www.kaggle.com/jmpark746/pga-tour-data-2010-2018. The set includes players' annual statistics from 2010-2018, as well as the number of wins and top 10 finishes each year. It is worth noting that the ideal data set would include the results for each tournament along with each players' statistical categories for that tournament. Unfortunately, player statistics for individual tournaments could not be gathered and built into a data set under the time constraints of this project.

The PGA Tour data set includes traditional statistics such as percent of fairways hit, greens in regulation (GIR), and number of puts; as well as the relatively new 'Strokes Gained' statistics that are aimed at isolating each aspect of the game and representing how a player is performing compared to others.

With the exception of the player names, all columns of interest are numerical. There are a two issues with NaN's in this data set. The first is that if a player did not have a win or top 10 finish it is recorded as NaN rather than zero. The second issue is that 634 of the 2312 rows have NaN for all the player statistics. While this is a significant portion of the data set, the statistics are only missing for 7.2% of the main rows of interest (players with wins or top 10 finishes).

Minimum Viable Product

At a minimum the project will present which aspect of golf (driving, approach the green, around the green, or putting) translates into better tournament results. If the analysis shows no higher correlation between tournament results and one aspect over the others the data supporting this will be presented.

Above the MVP, the golfers with the most wins and top 10 finishes over the time frame will be analyzed in more depth to track how their statistical rankings changed over time as well as their results. Perhaps this will give insight into the most consistent top golfers' attributes. There was also an outlier with one player having five wins in a single year. This player will be included in the analysis of top players, if he wasn't going to be based on the total number of wins and top 10s (maybe he just had one really good year).

Finally, if the above objectives are met the data set will be added upon to include years from 2004-2021.