Ron Richardson

CIS 598

Homework 5

# Introduction

Using machine learning techniques to predict the outcomes of sports events is nothing new. There has been lots of research in the major sports such as baseball, basketball and football, but the amount of research for golf has been sparse. What makes golf so hard to predict is the amount of randomness that can occur in a round of golf. A golfer might hole out a long shot where the probability of doing so is very high. A hole in one by a professional golfer is estimated at 3,000 to 1, whereas for an average golfer the probability jumps to 12,000 to 1 [1]. A golfer making a hole in one earns them roughly 2.1 strokes gained for the hole, where the mean strokes gained is usually between 0.5 and 1.

Determining the best feature set for predicting golf scores has been evaluated several times, starting with Davidson and Templin in 1986, where only three features explained 86% of a golfer’s scoring variance (greens in regulation (GIR), total putts, and driving proficiency) [2]. In 1992, Shmanske used three different, but similar, features for prediction [3]. This type of research continued, with similar variations, until Sen created a single metric to use for score prediction [4].

With the introduction of ShotLink by the PGA Tour in 2004, the amount of data has increased significantly [5]. ShotLink is a real-time system to collect every shot hit by a golfer on the PGA Tour. The data is collected by a team of volunteers at the tournament using a laser to pinpoint the starting and ending location of each shot while logging the type of turf the golfer hit from. Over the past 14 years, the amount of data that has come from this system is staggering. From 2010 to 2018 alone, over 10 million shots have been recorded and over 300 data features calculated.

Coming out of this trove of data was a new statistic called Strokes Gained. This statistic was first created by Mark Broadie in 2008 for putting and expanded in 2012 to include other aspects of the game [6]. Strokes gained provides a benchmark for comparing other golfers in certain skills and has been valuable in giving viewers and fans a better insight into how a golfer is performing.

Being able to predict scores for golf has implications not just within golf but also within the gambling community. Golf is already included in several daily fantasy sports services such as DraftKings and FanDuel, so being able to predict scores can help a gambler maximize his or her earnings. DraftKings and FanDuel are both valued at over one billion dollars, so the economic impact is huge, and with better models to predict golf scores, more users might be inclined to play, only boosting the revenue for these services.

Golf prediction can also be useful for coaching as well. By using the prediction model, a golf coach can teach a golfer how to avoid situations and shots that result in poor scores. This can play into the strategy aspect of the game, by teaching golfers to aim for certain areas of the hole to optimize their next shot, and ultimately their final score. For example, Zach Johnson is known as a fairly short driver of the ball. On longer holes, it might be advisable for Zach to lay up his second shot on a par 5 hole (avoiding any hazards near the green) and play a shorter and easier shot in for a better chance at making a birdie. This logic doesn’t work for a golf such as Brooks Koepka, who drives the ball much longer than Zach, as he will have a shorter distance for his second shot on a par 5, and thus have a better chance at avoiding hazards and having an eagle attempt. Knowing the shots that are optimal for a golfer is important, as that maximizes their chances for a lower score. Zach wouldn’t want to lay up to 100 yards away, when a predictive model shows he has better chances from 125 yards away.

Creating a predictive model on a shot by shot basis can also assist golf commentators. The model can provide a golf announcer instant information to relay to the viewer about the golfer’s chances. This gives the viewer a baseline to compare the actual result to, to really see how easy or difficult the shot was. Expanding on that idea, an automated commentating system could be developed for fans that just want pure information and no “fluff” that commentators sometimes fill in dead air with.

# Relevant Work

The “gold standard” in research for comparing performance of professional golfers comes from Mark Broadie of Columbia University. In “Assessing Golfer Performance on the PGA Tour” [6], a system called “strokes gained” was proposed to more accurately measure the performance of a golfer during a tournament. Certain golf statistics don’t show the whole picture of a golfer’s performance, but strokes gained can. One example is the total number of putts per round statistic. Golfer A might have only taken 29 putts during his round while golfer B took 31. Does this mean that golfer A is a better putter than golfer B? Not necessarily. Golfer A might have missed more greens in regulation (measured as being on the green is 2 shots or less than par for the hole). When missing a green in regulation, the golfer typically chips the ball towards the hole, and at least on the PGA Tour, can get the ball much closer to the hole than a shot from 100+ yards away. A shorter shot is considerably easier to get closer to the hole than a longer shot. If golfer A chips his or her ball to 2 feet away and makes that putt, he/she has only logged one putt. Golfer B hit the green in regulation but is 60 feet away. Golfer B’s first putt travels 58 feet, leaving only 2 feet. Golfer B then makes the 2-foot putt. Does this mean golfer B is worse of a putter than golfer A? Lagging a 60-foot putt to only 2 feet is considered a very good putt.

In his book “Every Shot Counts” [7], Broadie calculated the probabilities of one-putting from different distances, compared to the probability of three-putting from 2003 to 2012. From 2-feet, the probability of making it in one shot is 99%, with an average of 1.01 putts. From 60-feet, the one-putt probability is reduced to 2%, with an average of 2.21 putts. So, golfer B, having taken only 2 putts from 60 feet did 0.21 putts better than the average. Golfer A, however, did 0.01 putts worse than average from 2 feet away.

On average, a professional golfer on the PGA Tour only makes 50% of putts from 8 feet away [7]. From 40 feet away, the probability of taking 3 putts to get the ball in the hole is 10%, and only 4% to make it in 1 putt. Distance matters when comparing putting performance, as someone that makes a 40-foot putt in 1 shot did considerably better than a golfer that took 2 putts. One putt of 40 feet does not hold the same weight as one putt from 2 feet, and thus, should be ranked higher.

Using the strokes gained statistic can be helpful in predicting scores as it gives a good indication of the probability of making a certain shot. However, the difficulty of the shot isn’t completely included. In a paper submitted for the 2017 MIT Sloan Sports Analytics Conference, Levin went a step further to include certain characteristic of a shot in assessing the difficulty of it by including the type of turf the shot is taken from (fairway, bunker, rough, etc.), the distance from the hole, the angle of the approach, and characteristics of the hole, course, and day the shot was taken [8]. Typically, on a golf course, the rough is the tallest of the grass and borders the shorter fairway grass. Hitting a ball from the fairway is easier, since there is less grass in the way. When hitting from the rough, the grass can be long and thick, causing the club to twist or get caught, resulting in a less than perfect shot. Some courses historically have longer and thicker rough than others (such as a course hosting the U.S. Open), so a shot from that rough should be considered more difficult than at another course.

Wiseman [9] used five different regression models to predict the winning score of a golf tournament. The conclusion was made that Bayesian Linear Regression produced the highest R2 and the lowest error values, however most of the other models were quite close. The author appears to only use five features to predict the score (course par value, course yardage, average round 1 score, round 1 leading score, and total prize money), all of which are not at the player level but rather the tournament level. It is interesting to see that the course par value and yardage are relevant factors. K. C. Sen also did some research into what metrics make the most impact in a golfer’s performance, and a new metric was created which is a combination of three current statistics [4].

As mentioned in the introduction, several research papers have been produced that use a similar set of features. With the introduction of ShotLink, the amount of data available has increased. An update of this research to find if any of the new detailed statistics are significant in score prediction. Being able to break down the game of golf into finer details will only aid in prediction, as some statistics are overly generalized and can be misconstrued.

# Proposed Solution

One of the main reasons predicting scores in the game of golf is difficult is due to the amount of randomness that occurs. This random variability can have significant impacts on predictions. Kabir C. Sen attempts to combat this by eliminating any scores that deviate more than 1 stroke from par [4]. This appears to be a good method to eliminate the randomness as eagles, double bogeys and other scores more than 1 stroke away from par account for only 2.53% of all scores between 2010 and 2018 on the PGA Tour.

In addition to reducing the probability of those scores, feature reduction will be very important for score prediction. There are four categories of data that will be explored in this project:

* Strokes (43 features)
* Holes (55 features)
* Rounds (178 features)
* Courses (33 features)

A new feature set for each golfer will be calculated from the Rounds data. Averages and totals will be calculated and associated with each golfer from the years 2010 through 2018. Hole data will be aggregated by each hole on a course, and the same will be done for Course data. Another set of data will be explored, combining average Hole data with Course data. Without having explored this combination yet, I don’t know how useful it will be in predictions. I am hypothesizing that it will be useful for a general model that describes the course and hole difficulty, which should in turn help with overall score predictions. Combing the two should also help with performance during model training and testing.

Using the new four sets of aggregated data, running a principal component analysis (PCA) on each will help with dimensionality reduction. With 309 features available and over 10 million shots recorded, performance will be an issue with training the model, so any help in reducing that will be necessary. The goal is to separate out the features that account for 90-95% of the variance. Other techniques such as Incremental PCA and Randomized PCA will be explored to optimize performance and accuracy. Domain knowledge will also be important as several features are highly correlated and can most likely be eliminated to further reduce the number of features. This will be explored by investigating the Spearman correlations between several or all of the features.

Domain knowledge will also be required to add certain features, specifically with the Course and Hole data. Not included in the data available is the size of the green. If a green is small in size, it will likely be harder to hit with a shot, which has the chance to increase the golfer’s score. However, coordinate data is available for every shot on a hole, so calculating an estimated green size could prove useful. Drappi and Ting Keh [10] used the convex hull technique to calculate the green size, as well as bunker sizes using this coordinate data. An implementation of this technique will be included in the feature set to determine its usefulness in score prediction.

Once the features have been selected for each set of data, several models will be created. Determining the best training algorithm will be intensive, especially when considering past data to measure current skill levels. Drappi and Ting Keh [10] employed a stochastic search variable selection (SSVS) to help combat the high collinearity of features. Determining the optimal amount of past data to include will be important when predicting current performance. Wiseman employed a Bayesian Linear Regression model in his research, after trying out several other models, such as Neural Networks, Decision Forests, and a Boosted Decision Tree. These algorithms will be compared in this project, as well as investigating the use of model stacking and ensemble learning to further improve accuracy.

Overall, there will be six total models created and tested. One model will be created and for each of the four sets of data. Another model that combines all the significant features of each of the four individual models will be created. Finally, the last model will take only the most significant feature from the four sets and combine them into one model.

These models will be a combination of all golfers, but the goal of the project is to predict the score of one golfer at a time. This project will investigate if a golfer-specific model provides any significant increase in accuracy over a generalized model. To determine this, several golfers at random will be selected and specific models will be created. A potential of over-fitting might occur by being so specific to a golfer, and performance concerns are in play as well. However, having a golfer specific model can help a golfer train and learn better by running their game through several simulations. This can prove useful in the golf world, but my gut tells me that the results won’t be significant enough from an optimized and generalized model.

Another model that will be investigated will be a classification of golf holes to determine similarities. Using either a k-nearest neighbors or support vector machine algorithm, holes will be evaluated by their par values, and classified, creating three different models (one for par 3 holes, one for par 4 holes, and one for par 5 holes). This should be able to also reduce features and generalize the hole data for use in predictions.

Because of the number of models to generate and optimize is large, they will only be trained on one year of data (specifically 2018). Once these models are trained and optimized and the best one(s) are selected, they will be retrained on data from 2010 through 2018. To validate the model, the first few events of the 2019 season will be used. These events have already started at the time of this paper being written, so actual results will be available to compare against.

# Works Cited

[1] T.J. Auclair. June 29, 2018. Odds of a hole in one, albatross, condor and golf’s other unlikely shots Retrieved December 12, 2018 from <https://www.pga.com/news/golf-buzz/odds-hole-in-one-albatross-condor>

[2] J.D. Davidson and T.J. Templin. 1986. Determinates of Success Among Professional Golfers. *Research Quarterly for Exercise and Sport*, 57, 1 (1986), 60-67.

[3] S. Shmanske. 1992. Human Capital Formation in Professional Sports: Evidence from the PGA Tour. *Atlantic Economic Journal*, 20, 3, 66-80.

[4] K. C. Sen. 2012. Mapping statistics to success on the pga tour: Insights from the use of a single metric. *Sport, Business and Management: An International Journal,* 2, 1 (2012), 39-50.

[5] ShotLink Background. Retrieved December 12, 2018 from http://www.shotlink.com/about/background

[6] Mark Broadie. 2012. Assessing Golfer Performance on the PGA TOUR. *Interfaces,* 42, 2 (April 1 2012), 105-228. DOI: https://doi.org/10.1287/inte.1120.0626

[7] Mark Broadie. 2014. *Every Shot Counts.* Gotham, New York, NY.

[8] Adam Levin. 2017. Ranking the Skills of Golfers on the PGA TOUR using Gradient Boosting Machines and Network Analysis. In *MIT Sloan Sports Analytics Conference (March 3-4, 2017).* Boston, MA.

[9] Oisin Wiseman. 2016. *Using Machine Learning to Predict the Winning Score of Professional Golf Events on the PGA Tour.* Master's thesis. National College of Ireland, Dublin, Ireland.

[10] Christian Drappi and Lance Co Ting Keh. 2018. Predicting golf scores at the shot level. *Journal of Sports Analytics*, Pre-press, 2018, 1-9. DOI: 10.3233/JSA-170273