Ranking the Skills of Golfers on the PGA TOUR using Gradient Boosting Machines and Network Analysis

Paper Track: Other Sports
Paper ID: 1685

ADAM LEVIN
Data Scientist and Golf Fan
adamlevin44@gmail.com

Abstract:

Until recently, ranking the skills of golfers on the PGA TOUR was best accomplished by using imprecise summary statistics such as Driving Accuracy and Average Putts Per Round. Since 2003, The PGA TOUR, through their ShotLink Intelligence™ program, has collected detailed shot-level data, which provides coordinates of the locations of shots along with other information. Through the analysis of this data, more fine-grained and precise estimates of the skills of golfers on tour are now possible.

The problem of estimating the skill of golfers in different aspects of the game given data from competitions is not simple. This work recognizes a wide array of statistical challenges associated with this problem, which a number of previous approaches to the problem have failed to adequately acknowledge.

A brand new approach to the problem is presented which invokes comparisons of the quality of shots taken on the same hole during the same round. The comparisons are utilized in a Network Analysis technique, which is generalized to suit the needs of the problem. This approach is supported with empirical evidence of stronger correlations with the future success of the golfers than the system currently used by the PGA TOUR.

1. Introduction

Golf is a game that requires a variety of skills – driving off the tee, hitting approach shots from the fairway or rough, and putting, to name a few. Precise estimation of the skill of professional players in the various aspects of the game is useful for a variety of purposes. With accurate estimations of how players' skill sets compare, players and coaches can create data-driven training plans and fans watching the game can gain a greater understanding of the strengths and weaknesses of their favorite players.

Competitive play is not setup in a way that makes the estimation of skill levels simple. It is far from a scientific experiment where players are told to take multiple attempts from precise locations under controlled circumstances. In golf, players take around 72 shots per round but every shot is unique. A slight change of angle can make a shot entirely different. The quality of the lie can make two shots taken very close to one another very different. Weather conditions can vary from the morning to afternoon.

This paper improves what is currently being done to estimate the skill levels of the players on the PGA TOUR. The principle followed in this paper is that measures of skill are as valuable as the accuracy with which they predict future performance of the players. Previous attempts at estimating the skill levels of golfers on the PGA TOUR have neglected a few subtleties that make this problem challenging. In this paper, these subtleties are detailed and a novel approach to the problem is given. The skill estimates that result from this approach are used to predict the future performance of golfers. Superior predictive performance is demonstrated using the new approach compared to a reproduction of the current standard. Theoretical reasoning is supplied to justify the new method.

This work would not have been possible without detailed shot-level data that the PGA TOUR started collecting in 2003 using their ShotLink™ system. The availability of these data has opened up the possibility of understanding the professional game in greater depth statistically. This work also owes its foundation to the work done by Mark Broadie of Columbia University. His work in developing the Strokes Gained concept (explained in Section 3) has advanced our understanding of the game by being the first to really quantify individual skill sets of the players on the PGA TOUR. His contributions and the work of others in this area are summarized in Section 4 of this paper.

2. Dataset

The dataset used in this paper was provided by the PGA TOUR through their ShotLink Intelligence™ program. Volunteers equipped with special equipment collect the data. At the shot level, the data contain locations of all shots of the players on the PGA TOUR since 2003. Data from the round level – number of strokes taken in a round – is also used in this paper. Data used begins at the start of the 2003 season and goes through the 2016 TOUR Championship. Some summary statistics from the raw data are provided in Tables 1 and 2.

Query	Result
Number of Shots	16,469,637
Number of Players	2,054
Number of Courses	107
Number of Tournaments	561
Number of Rounds	2,244
Number of Holes*	40,392

Table 1: Summary Statistics. *Number of Holes means number of unique hole-day combinations.

Turf	Percentage
Green	40.7%
Tee Box	25.3%
Fairway	16.6%
Primary Rough	8.0%
Intermediate Rough	2.4%
Green Side Bunker	2.3%
Fringe	1.8%
Unknown	1.2%
Fairway Bunker	1.2%
Native Area	0.4%
Other	0.2%
Water	<0.1%
Grass Bunker	<0.1%

Table 2: Percentage of shots taken from different turfs in raw data.

Preprocessing steps were carried out prior to any analysis. These steps were taken to ensure the data's quality and cohesiveness. More details on the preprocessing steps are given in Appendix B.

3. The Strokes Gained Statistic

Before the availability of detailed shot-level data, multiple statistics were used to quantify specific skills in golf including Driving Distance, Fairways Hit, and Greens In Regulation (GIR), among others. To illustrate the lack of precision that results from these statistics, take GIR as an example. GIR is the count of the number of holes on which a golfer reaches the green in two strokes less than the par value of the hole or fewer. GIR

attempts to quantify a golfer's skill with his or her approach shots. However, if two players start from the same position in the fairway and one hits it on the green 80 feet away and the other hits it to the fringe 18 feet away, the player who hit it on the green will be credited with a GIR while the other player will not, despite having left his ball in (arguably) a less desirable position.

This example illustrates the need to quantify the "desirability" of a particular location on a particular course on a particular day, or equivalently the difficulty of playing a shot from a particular location. It also motivates quantifying the quality of a particular shot by taking the difficulty of the starting location and subtracting the difficulty of the finishing location. This is the idea developed by Mark Broadie and is named the Strokes Gained Statistic.

To continue with the previous example, if the two golfers started from the fairway where it takes an average golfer 3.3 strokes (which tends to correspond to about 225 yards on tour), and we know how difficult it is for the average golfer from the locations where the two golfers' balls ended up, we can quantify the quality of both players' shots. From 80 feet away on the green the average PGA TOUR golfer takes about 2.3 strokes to get the ball in the hole on average, while from 18 feet away on the fringe the average tour player takes about 1.9 strokes to get it in on average. Following the convention established in Broadie (2008), the Strokes Gained Statistic is then calculated using the following equation:

$$shotquality = Difficulty_{start} - Difficulty_{finish} - 1.$$

To conclude the example, the player whose ball ended up on the green had a shot quality of 0 (3.3 – 2.3 – 1), while the player whose ball ended up on the fringe had a shot quality of 0.4 (3.3 – 1.9 – 1). A positive shot quality corresponds with a shot that was better than the average player would have done and a negative shot quality corresponds with a shot that was worse than the average player would have done.

3.1 Assumptions of Strokes Gained System

Before continuing towards making a model of how difficult a given shot is, it is useful to think about the assumptions of the Strokes Gained framework. The first assumption is that we can estimate with reasonable accuracy how difficult a shot is. This

is actually quite a challenge and there are potential pitfalls in doing this, which will be discussed shortly.

The second assumption is more fundamental. What does it mean to quantify the difficulty of a given shot? In Broadie (2008) this is defined as the average number of strokes taken from a given location by an *average* player. There is a subtle assumption in this method – that the desirability of a given location is the same for all players. This is generally a safe assumption because it's mostly true; the desirability of different locations is very similar for all players. However, it's useful to acknowledge that this method is a simplification of how the game is actually played. A consequence of this simplification is that the possibility that a player acts strategically is ignored. For example, a player could be faced with an option to lay-up on a par five, or try to hit it on the green, which is surrounded by bunkers. If this player is an excellent bunker player, this will certainly factor into his decision about whether or not to go for it. However, post-hoc evaluation of the quality of this players' shot will take into account the desirability of the location he ends up at as measured by the theoretical performance of an average golfer from that location and thus will not correctly account for the strategic thinking that was involved in playing the shot.

This work will focus on coming to terms with the first assumption. The second assumption is more complex and will be left for another contributor.

4. Modeling Difficulty

Modeling the difficulty of a shot is challenging for a few reasons. First, the difficulty of a shot can vary with conditions that can be very specific to the situation: the hole setup, the weather, the lie, and the angle of approach. These data do not contain direct information about the location of the hole relative to the edge of the green (hole setup), the weather, or the lie. The extent to which these factors have an effect on the difficulty of a shot must be inferred from the data.[1]

There are also potential pitfalls in modeling difficulty without thinking about the consequences of how a model is learning. When fitting a model for difficulty that contains

[1] The approach given in this paper infers the effect of hole setup and weather but not the quality of the lie. The 'lie' variable in the data was not found to be useful.

information that distinguishes between different courses, there is a potential for misinterpretation of the results because the players who played on one course might be of a higher caliber than the players who played on another course. This has been pointed out in Fearing et al. (2010).

Attempts to use spatial clustering, nearest-neighbor algorithms, kriging, or spline interpolation to infer difficulty of a shot directly from the observations on a particular hole runs into a subtle bias – players who end up playing a shot close to one another might have general skill levels that are correlated with one another. For example, a favorable location to play from – an area containing approach shots following well-placed drives, for example – might attract the balls of players who are already playing well and thus be more likely to succeed on the following shot.

Additionally, through the use of a complex model that has potential to fit the data very closely, there is an unintended consequence of using a model that has been trained on the same observations that it is assessing the difficulty of. The unintended consequence is quite subtle and involved. Interested readers are directed to Appendix C.

Producing absolute measurements of difficulty of a shot is therefore very challenging. In the rest of this section, previous attempts at this task will be outlined. Then, a new approach with a subtle change in intention will be introduced. This approach will sacrifice a universal baseline estimate of difficulty of any single shot for the sake of allowing fair comparisons of two shots taken on the same hole and the same day. The comparisons will then be utilized using a network analysis technique to determine the relative skill levels of all the players on the PGA TOUR.

4.1 Previous models for difficulty of a shot

Broadie (2012) models difficulty of a shot separately for 5 categories of shots – tee, fairway, green, sand, and rough. Distance is used as the primary predictor of difficulty and piecewise polynomials are fit to model the relationship between distance and difficulty for all shots except putts. For putts, a physical model of probability of one-putting combined with a physical model of probability of three-putting is used. Neither elevation change nor angle of approach was considered as predictors.

In Broadie (2012), a model for distinguishing between course-round difficulty and player skill was done at a global level – estimating the overall difficulty of a course and skill levels of the field without allowing for the possibility that particular types of shots might be more or less difficult at certain courses or certain fields more or less competent at certain types of shots. Additionally, this model assumed players' skills were static, not changing through time. According to comments made by Broadie subsequently[2], strokes gained statistics currently used on tour are adjusted by the average strokes gained performance of the field for each category of shot during each round to produce *Strokes Gained to the Field*. The problem with this method is that it neglects the possibility of the quality of field varying at different tournaments. This method of evaluating performance will be compared to the novel method in the results section of this paper.

Fearing, Acimovic, and Graves (2010) modeled difficulty of putts using generalized linear models for probability of holing out and distance to go. The challenge of estimating the intertwined quality of field and course difficulty was acknowledged and a model was fit with player and hole-specific effects. The authors' approach allows for situational putting performance predictions. This approach is admirable. However, similarly to Broadie (2012), it assumes players' skills are static, not changing through time. The authors focused mostly on putting; they fit a similar model for off-green performance but do not distinguish between different potential off-green skills (short-game versus long-game, for example).

Söckl et al. (2013) introduces the ISOPAR method. This involves interpolating a smoothing spline to infer difficulty of a shot based on the observations on a particular hole during a particular round. Unfortunately, in using these values to measure performance, the authors do not recognize the biases involved with this approach that were discussed above – the varying quality of a field, the bias for desirable locations to more frequently contain the shots of more capable players, and the unintended consequence of using of a model that has been fit to the same observations that it is predicting.

Finally, Yousefi and Swartz (2013) take a Bayesian approach to estimating the difficulty of putts by allowing the possibility for difficulty to vary from different portions of the green, which they divide into eight quadrants. This approach is similar to Söckl et

[2] http://fyre.it/oC8arn2k.4

al. in that it ignores the aforementioned biases – there is no mention of varying quality of the field, nor any mention of the possibility that the observations in a particular quadrant might be biased according to the general ability of the players whose balls end up there.

4.3 A Change in Intention

The goal is to model difficulty of a shot given characteristics of the shot – turf the shot is taken from (fairway, bunker, etc.), distance from the hole, angle of approach, and particular characteristics of the hole, course, or day on which the shot was taken. Difficulty of a shot has been defined by Broadie (2010) to be the number of strokes a player of average caliber would take from the given location. To actually estimate the difficulty earnestly, one must simultaneously infer the difficulty from a particular location and the skill-level of the player taking the shot. If one wishes to incorporate features that identify the varying difficulty of shots on different courses, on different days, on different holes, for different types of shots, with player-skills that differ for different types of shots and that change through time, the number of parameters to estimate becomes immense.

Instead of attempting this type of model, this paper acknowledges that the comparison of the qualities of shots that are taken on different holes or different days is not an apples-to-apples comparison. Instead of computing shot-quality using performance relative to an absolute baseline, individual shots taken on the same day and on the same hole will be taken as observations used to compare golfers' skills relative to one another. Relative skill-levels of the golfers are then computed using an analytical technique on the network of all PGA TOUR players. This network analysis method will be described in more detail in Section 6 of this paper.

This methodology provides more freedom in creating a model for difficulty of a shot because all that is necessary is that the model produces estimates that allow fair comparison between the quality of two shots taken on the same day and on the same hole.

5. Building the Model

Without the requirement that a model be useful in comparing the difficulty of shots taken on different days or on different holes, different approaches to making a model are

possible. For example, one could fit models using only the data from each hole and day, or for each tournament. Modeling using limited data has the potential advantage of smaller bias, but comes at the cost of greater variance.

The model chosen here will use all of the data simultaneously but allow for the possibility of changing difficulty across different days, courses and holes through the use of indicator variables and their interactions with the feature space. The ideal balance in the tradeoff between bias and variance will be found using a cross-validation strategy (to be described later in this section).

The rest of this section will describe the feature space, the model specification, and the model selection process. The variable to be predicted is the number of shots taken from a location and no data to identify the golfer taking the shot is used.

5.1 Feature Space

Distance is the most important feature for predicting difficulty of shot. The relationship between distance and difficulty of a shot is highly non-linear and is different for different turfs. By fitting various regression models, it is possible to visualize this relationship. One that fits the problem reasonably well is isotonic regression. Isotonic regression is a non-parametric regression that fits a step function to model a bivariate, monotonic relationship. The relationship between distance and difficulty is not, in fact, monotonic for all golf shots. For example, for off-the-green shots from certain angles of approach, specifically when there is not much green between the player and the hole, shots of slightly longer distance can be considered easier. However, it is a useful method to visualize the data. In Figures 1 and 2, isotonic regression models are shown as a means of comparing the relationship between distance and difficulty for different turfs.

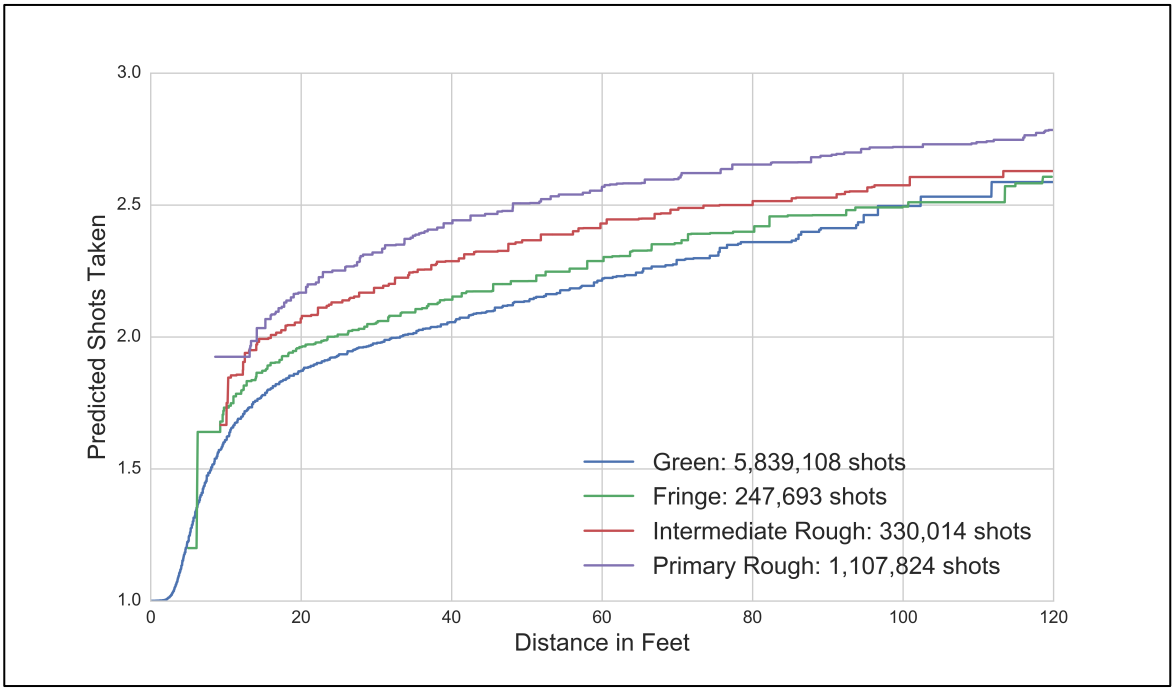


Figure 1: Isotonic Regression models for shots taken from short distances. In general shots from the green are slightly easier than shots from fringe, which are slightly easier than shots from Intermediate Rough, which are slightly easier than shots from Primary Rough.

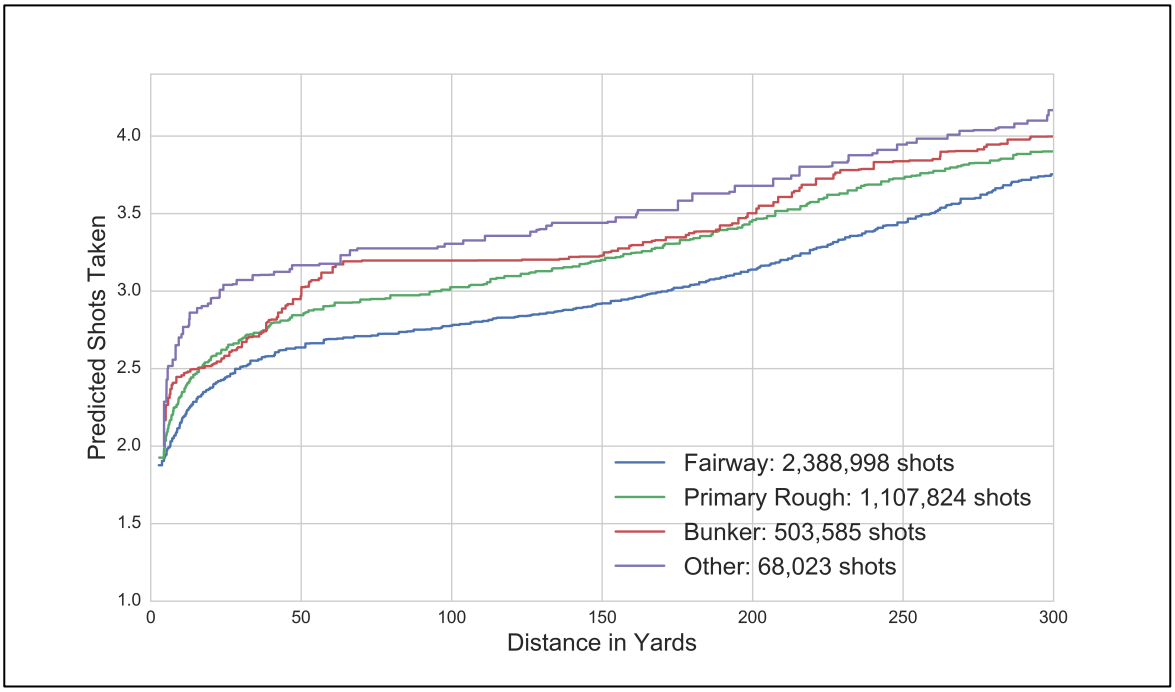


Figure 2: Isotonic Regression models for shots taken from longer distances. Close inspection reveals that shots from the Primary Rough are more difficult than shots from the Bunker for distances between 20 and 40 yards, while less difficult for other distances.

These plots provide justification for distinguishing between different turfs in modeling difficulty of a shot. In this paper, separate models are fit for each of seven turfs – Green, Fringe, Fairway, Intermediate Rough, Primary Rough, Bunker, and Other.[3] The 'Other' category includes all shots not in any of the other six categories; it contains shots recorded as from 'Unknown', 'Native Area', 'Other', 'Water', and 'Grass Bunker'. Strong arguments could be made that 'Water' should be its own category given a potential difference in difficulty resulting from a penalty stroke and that 'Grass Bunker' should be included in 'Primary Rough'. Grouping these shots into one 'Other' category is an approximation that could perhaps be improved on.

Without data on wind, temperature, or condition of a lie, the amount of general (not course, round or hole identifying indicator) features to predict difficulty of a shot is limited. Elevation change, which is in the data, is a statistically significant predictor of difficulty but it does not help explain very much of the variance compared with distance. A new general feature can be derived to encapsulate the difficulties of different angles of approach for off-green shots. This feature is called 'Green to Work With', which is golf jargon for how much green is between a location and the hole. Because the location of the edge of the green is not given in the data, this measure must be approximated from the data. Figure 3 explains this feature visually.

The algorithm used to produce this feature is presented in Algorithm 1. Similarly as with slope, Green to Work With is a statistically significant predictor but does not help explain very much of the variance. It is most statistically significant for shots from the Primary Rough, which comports well with common golf sense – it is more critical to have plenty of green to work with when one is in the rough since it is more difficult to apply spin to the ball and control the run out. Table 3 displays the added benefit of the features elevation change and Green To Work With for each category of shots.

[3] No model was fit for the first shots on each hole because the difficulty – the number of shots taken on average from a location – is taken to be the average score on the hole for the round.

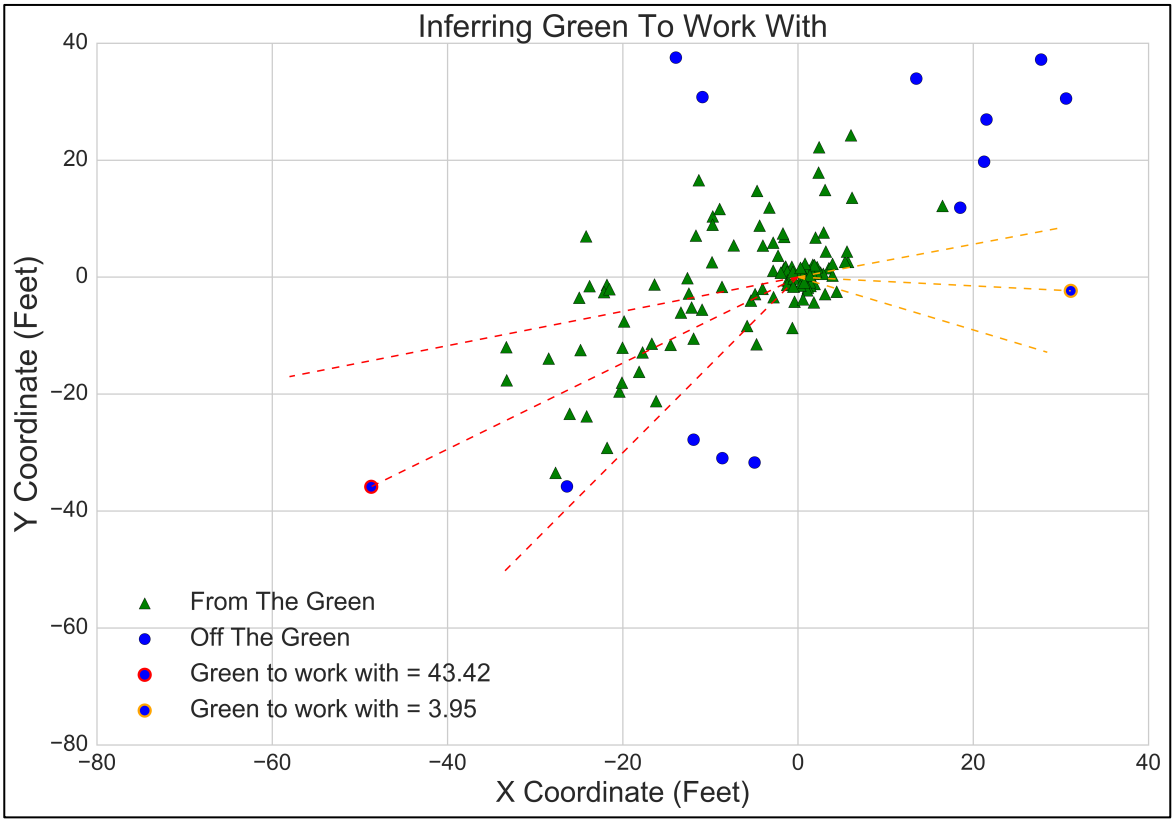


Figure 3: The red point has an inferred Green To Work With of about 43 feet, while the orange point has an inferred Green To Work With of only 4 feet. This corresponds with a more difficult angle for the shot from the orange location.

As mentioned earlier, the feature space also contains indicator variables that indicate the course, the interaction between year and course, the interaction between hole and course, and the interaction between round, year and course. The rationale behind the inclusion of these variables is to allow a model to determine if certain shots were more

Algorithm 1: Produce Green To work with

Given: *angle_of_shot, shots_on_green, slack, slack_increment, max_slack*
Returns: *green_to_work_with*

Subset = *shots_on_green*[angle – *slack* < *angle_of_shot* < angle + *slack*]
While len(Subset)==0:
    slack += *slack_increment*
    If *slack* > *max_slack:*
        Break
    Subset = *shots_on_green*[angle – *slack* < *angle_of_shot* < angle + *slack*]
If len(Subset)==0: *green_to_work_with* = null
Else: *green_to_work_with* = max(Subset.Distance_from_hole)

difficult on any particular course, or on any particular holes, or during any particular rounds.

5.2 Model Selection and Fitting

The model used is a Gradient Boosting Machine. This algorithm was chosen because of the ease with which it models both non-linear relationships and interactions between features. Another attractive feature of this model is that it produces very accurate predictions because of the many levers available to help balance the tradeoff between bias and variance.[4]

Special attention was paid to the strategy used to produce the estimates of difficulty to be used in the subsequent skill estimation process. Since this model is very complex – with all of the indicator variables there can be as many as 40,000 features – the unintended consequence discussed in Appendix C must be mitigated by not using the same observations that the model is evaluating the difficulty of when fitting of the model.

The grouped cross-validation prediction strategy that addresses the concern of overfitting the data in evaluating the difficulty of shots naturally leads to the choice of grouped cross-validation for tuning the parameters of the Gradient Boosting Algorithm.[5] This process was handled in an automated fashion utilizing a Bayesian Optimization library.[6] For each category of shot, models were tuned using different subsets of features. Results for models fit with and without the features Elevation Change and Green to Work With are shown in Table 3. Results for models tuned with varying subsets of the indicator variables are shown graphically in Figure 4. In Appendix A, there are plots that show the model's predictions of difficulty for various situations of interest.

Despite not using the observed true number of shots taken to predict the difficulty of a given shot, this model is more accurate in predicting number of shots taken than the baseline currently used by the PGA TOUR. This model has a mean-squared-error of 0.4965 – with 95% confidence interval (0.496, 0.497) - on the 12.3 million shots that have the currently used baseline estimate available, while the currently used baseline has a

[4] The XGBoost library was chosen because of the ease with which it handles large datasets.
[5] See Appendix C for a description of this strategy and the motivation for it.
[6] https://github.com/fmfn/BayesianOptimization

mean-squared-error of 0.5025 – with 95% confidence interval (0.502, 0.503). Confidence intervals were computed using the bootstrap.

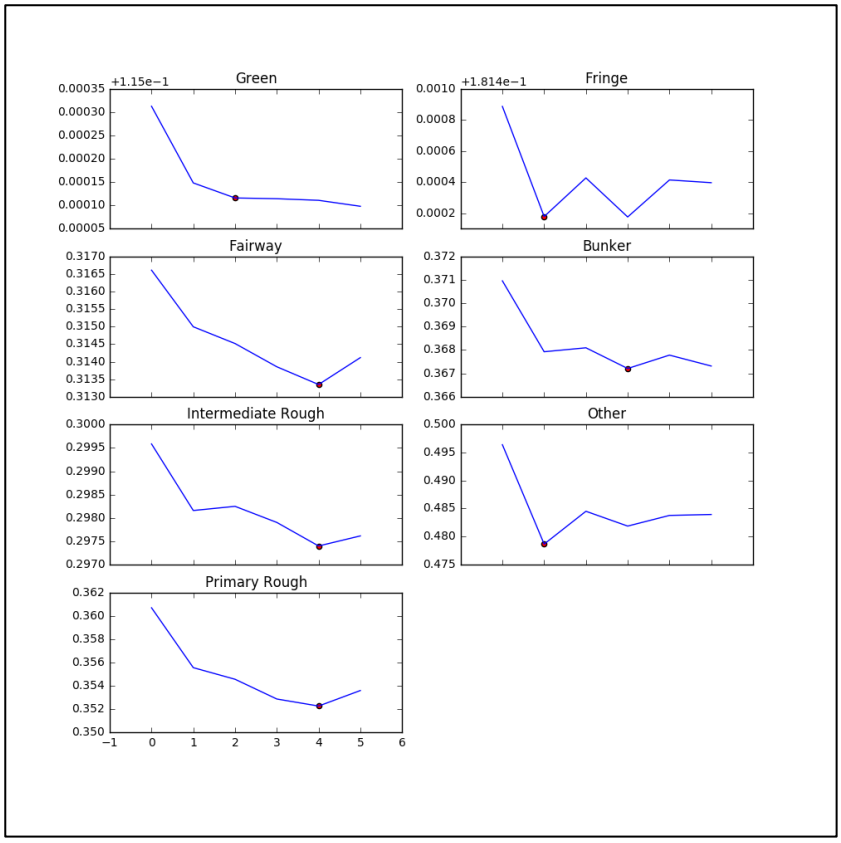


Figure 4: Cross-Validated Mean Square Error of models of varying complexity. Features-spaces from left to right: no indicators, with course, with year-course, with hole-course, with round-year-course, and with hole-year-course indicators respectively. This was done in a greedy fashion. Complexity choices are indicated with points.

Turf	Just Distance	Including Elevation Change	Including Green to Work With
Green	0.5898 ± 0.0003	0.5904 ± 0.0003	X
Fairway	0.2945 ± 0.0011	0.2962 ± 0.0011	0.2969 ± 0.0011
Intermediate Rough	0.4356 ± 0.0013	0.4375 ± 0.0013	0.4386 ± 0.0014
Primary Rough	0.3900 ± 0.0008	0.3927 ± 0.0008	0.3966 ± 0.0009
Fringe	0.1236 ± 0.0011	0.1255 ± 0.0010	0.1266 ± 0.0010
Bunker	0.3677 ± 0.0025	0.3704 ± 0.0025	0.3719 ± 0.0025
Other	0.2041 ± 0.0039	0.2102 ± 0.0038	0.2120 ± 0.0037

Table 3: 15-fold Cross-Validated R-squareds and standard errors for varying features spaces. Green to Work With is not relevant for shots from the Green.

6. Network Ranking System

Up until this point, the focus has been on producing estimates of difficulty of shots that allow for fair comparison between two shots taken on the same turf, the same hole and the same day. In this section, the system that makes use of these comparisons to rank the players' skills against one another's is presented.

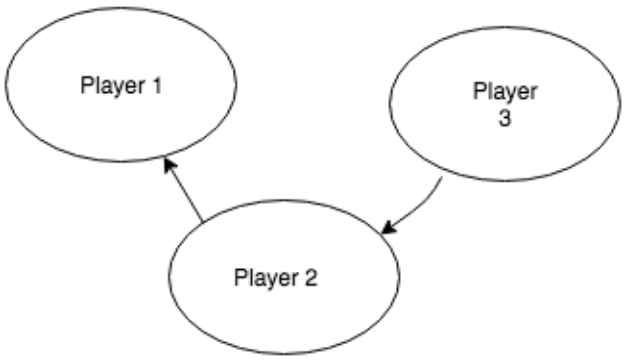
Park and Newman (2005) present a system for ranking college football teams given game outcomes using a network analysis. This work may be applied to golf by taking the 'games' to be comparisons between two players who have taken shots from the same turf on the same hole and on the same day. The approach starts by assembling an Adjacency Matrix that contains the data of the observed comparisons. Take, for example, the observed data to be the following:

Player Number	Shot Quality	Hole	Round	Course	Tournament	Turf
1	-.1	7	1	45	1	Green
2	.1	7	1	45	1	Green
2	-.1	18	3	64	2	Green
3	.4	18	3	64	2	Green

If we can make a fair comparison between the skills of players if they take a shot from the same turf on the same hole during the same round, how can we use these comparisons to estimate how good Player 1 is compared to Player 3 at putting, for example? In Park and Newman (2005), the idea is to compose an adjacency matrix that records the number of 'wins' one 'team' has over the other in the corresponding cell, like this:

0	0	0
1	0	0
0	1	0

The 1st row and column correspond with the results of player 1. A win for team *i* over team *j*, which in this context means a higher shot quality, corresponds to a 1 in the i,j^{th} cell of the matrix. This is a network with directed edges, which can also be represented by this diagram of nodes and edges:



The motivation for the ranking system is very intuitive. If two teams have a result against one another, the winning team (if there is one) receives a *direct win*. By traversing the network along the edges, teams accumulate *indirect wins*. In this example, since Player 3 has a win versus Player 2 and Player 2 has a win versus Player 1, Player 3 has an indirect win against Player 1. This is considered an indirect win of distance 2, since it required two jumps along the edges to arrive at Player 1 from Player 3. Indirect wins are a means of comparing the strengths of teams who do not have direct comparisons on record. Indirect wins, for both intuitive and mathematical reasons, must be given less weight than direct wins.

This simple network does not contain indirect wins of distance greater than 2, but networks representing larger data might have indirect wins of very large distances. For most networks that represent real data, as one counts the indirect wins at greater and greater distances, the amount of indirect wins becomes very large. In Park and Newman (2005), wins are down-weighted by a factor α^{d-1}, where alpha is a user-specified parameter that is less than 1 and d is the distance of the win. A player's 'win-score' is defined to be the sum of all his or her direct and indirect wins. Taking A to be the adjacency matrix, Player i's win score can be calculated as follows:

$$w_i = \sum_j A_{ij} + \alpha \sum_{j,k} A_{ij}A_{jk} + \alpha^2 \sum_{j,k,h} A_{ij}A_{jk}A_{kh} + \ldots$$

$$= \sum_j A_{ij}\left[\mathrm{I} + \alpha \sum_k A_{jk} + \alpha^2 \sum_{kh} A_{jk}A_{kh} + \ldots\right]$$

$$= \sum_j A_{ij}\left[\mathrm{I} + \alpha w_j\right] = \sum_j A_{ij} + \alpha \sum_j A_{ij}w_j$$

The vector of all win-scores can now be expressed compactly and solved for:

$$w = rowsums(A) + \alpha Aw$$
$$w - \alpha Aw = rowsums(A)$$
$$(\mathrm{I} - \alpha A)w = rowsums(A)$$
$$w = (\mathrm{I} - \alpha A)^{-1} \bullet rowsums(A)$$

For the infinite series to converge and the solution to be meaningful, alpha must be less than the reciprocal of the largest eigenvalue of A. This system inherently accounts for 'strength of schedule' since a direct win against a stronger opponent will result in more indirect wins than a direct win against a weaker opponent. Strength of schedule applies in golf as well. For example, Tiger Woods and other top golfers are known to only compete in high caliber tournaments; thus the fact that they have comparisons against other high caliber players rightfully earns them higher ratings. The relative importance of indirect wins – and thus the importance of strength of schedule – can be manipulated by setting alpha. Alpha should be fit to data so that the rankings that result are most predictive.

In Park and Newman (2005), the comparisons are recorded as either 0 or 1 – corresponding with losses or wins. The different numbers of opportunities that each team (or player in this context) has is dealt with by also computing a 'loss score', which involves the same computation on a network where edges represent losses. The overall team strength is then computed by subtracting a team's loss-score from its win-score. This is quite a powerful framework that can be applied to all sorts of situations involving recorded comparison between entities in a network with the goal being to estimate the strength of the entities involved.

This approach can be generalized to handle measures of magnitude of a win or a loss. In football this could be the score of the game. In golf it could be the magnitude of the difference in shot quality of shots taken on the same turf, the same hole, and the same day. In the earlier example, Player 3's 'victory' over Player 2 should be considered more impressive than Player 2's victory over Player 1 because of the magnitude of the differences in shot quality. The network now has a weight assigned to edges or flow. This weight must be non-negative for the network to retain it's meaning. To ensure non-

negative values, the difference in shot-qualities can be transformed using the sigmoid function thusly:

$$ComparisonScore_{ij} = \frac{1}{1+e^{-ShotQualityDifference_{ij}}} + .5,$$

with $ShotQualityDifference_{ij}$ being the difference between the shotquality of a shot by player i and the shotquality of a shot by player j (shotquality was defined on page 4). The .5 is used for no reason other than it is more aesthetically pleasing for the scores to be centered around 1 rather than .5. Filling the Adjacency Matrix with these scores, the example matrix becomes:

0	.95	0
1.05	0	.88
0	1.12	0

The same Park, Newman equations can be used to solve for player ratings as before, with an important change in the method of accounting for the different numbers of opportunities that each 'team' has had. Instead of computing both a win-score and loss-score as Park and Newman propose, one computes the strength-score using this new matrix plus an additional score – the 'everyone ties' score. The 'everyone ties' score, just like it sounds, is the score that would result if all the comparisons had been tied. To compute the 'everyone ties' score, one comprises a normalizing matrix, G (for games), which has the number of recorded comparisons between team i and team j in the i,jth cell. With the example data this matrix is:

0	1	0
1	0	1
0	1	0

The measure of strength of each team is then the computed as follows:

$$StrengthScore = \frac{w_A}{w_G} = \frac{(\mathrm{I} - \alpha A)^{-1} \cdot rowsums(A)}{(\mathrm{I} - \alpha G)^{-1} \cdot rowsums(G)}.$$

The denominator normalizes each team's score for the number of 'opportunities' it has to accumulate points while completing the infinite walk along the edges of the network.

Being able to account for different magnitudes of 'wins' or 'losses' is useful in a variety of settings. Also useful, for both rating golfers or other competitors in a network, is to allow players ratings to change through time. It is well known that golfers' general abilities and specific skills fluctuate over time. To allow for this, instead of representing all of a golfer's observations throughout time with one node in the network, one can represent a player's skill for each time period with a node in the network. A player might have one node representing his or her performance for a few tournaments, and other nodes representing his or her performance for other time periods. The player's rating at any point in time is the Strength Score of the node representing that player's performance during the time period. A player's node in a new time period should not start as a blank-slate but instead it should 'inherit' some fraction of the observations from the player's previous time periods.

For the example data, taking each tournament to be a time period, this new matrix will have six nodes – one for each player-time period. Taking beta to be the fraction of an observation that is inherited from one time period to the next for the same golfer, this new matrix looks like this:

0	.95	0	0	0	0
1.05	0	0	0	0	.88β
0	0	0	0	1.12β	0
0	.95β	0	0	0	0
1.05β	0	0	0	0	.88
0	0	0	0	1.12	0

The matrix has been shaded to highlight the block structure. The diagonal blocks contain unweighted observations from the comparisons that occurred in the corresponding time interval. The off-diagonal blocks are down-weighted because they contain observations that are inherited across time intervals. To calculate the rating of Player 2 during Tournament 2 one would consider row 5, for example. The inheriting of observations is bi-directional. The rationale behind this choice is that upon receiving comparisons between player A and player B in time period t, this information can be used to better estimate the strength of both player A and B in time period t-1. Observations can be passed over more than one time period. Weights of observations passed between time periods can be computed using any sort of function - *h(t)* where t is the number of time periods - that makes sense. In general, the score and normalizing matrix now look like this:

$$A = \begin{bmatrix} A_1 & h(1)A_2 & h(2)A_3 & \dots & h(N-1)A_N \\ h(1)A_1 & A_2 & h(1)A_3 & \dots & h(N-2)A_N \\ h(2)A_1 & h(1)A_2 & A_3 & \dots & h(N-3)A_N \\ \dots & \dots & \dots & \dots & \dots \\ h(N-1)A_1 & h(N-2)A_2 & h(N-3)A_3 & \dots & A_N \end{bmatrix}$$

$$G = \begin{bmatrix} G_1 & h(1)G_2 & h(2)G_3 & \dots & h(N-1)G_N \\ h(1)G_1 & G_2 & h(1)G_3 & \dots & h(N-2)G_N \\ h(2)G_1 & h(1)G_2 & G_3 & \dots & h(N-3)G_N \\ \dots & \dots & \dots & \dots & \dots \\ h(N-1)G_1 & h(N-2)G_2 & h(N-3)G_3 & \dots & G_N \end{bmatrix}$$

Table 4 contains the Park and Newman ‘win-score’, ‘loss-score’ and strength along with the generalized rating, ‘everyone ties’ rating, and Strength Score described here and the further generalized Strength Score in which players’ ratings are allowed to vary through time for the sample data given earlier.

One need not stop there in incorporating relevant information towards ranking the relative skills of the players using this system. There is no reason that the values in each normalizing matrix must all be 1. If one believes certain comparisons are fairer than others, one could multiply an observation by a weight corresponding with its fairness in *each* of the described matrices. For example, if two shots taken on the same turf, the same hole and the same day are taken very close to one another spatially, a comparison between

these shots may be considered fairer than one between two shots that were taken from further apart. The rationale is that the model, which assesses the difficulty of a given location, is not perfect. Its misassessments are likely to be spatially correlated, so that shots taken from locations close to one another likely allow for fairer comparisons. Likewise, shots taken close two each other temporally are more likely to allow for fairer comparisons than shots that were taken a long time apart. Weather conditions (wind and rain), and course conditions (speed or bumpiness of the greens) are potential contributors to difficulty of a shot that the model is unaware of and that change with time. Thus, errors of the model are likely correlated temporally as well. Weighting the observations inversely with distance between shots and time between shots allows for the fairer comparisons to be emphasized more.

	PN Win-Score	PN Loss-Score	PN Strength Score	Gener. score	Num. Opportun.	Strength Score	Gener. Score w/ time	Num Opportun. w/ time	Strength Score w/ time
Player 1, Time 1	0	1.64	-1.64	11.2	12.4	.897	10.9	12.1	.899
Player 2, Time 1	1	1	0	16.8	17.9	.940	15.5	16.5	**.942**
Player 3, Time 1	1.64	0	1.64	13.2	12.4	1.06	10.2	9.67	1.06
Player 1, Time 2	0	1.64	-1.64	11.2	12.4	.897	8.69	9.67	.899
Player 2, Time 2	1	1	0	16.8	17.9	.940	15.4	16.5	**.939**
Player 3, Time 2	1.64	0	1.64	13.2	12.4	1.06	12.8	12.1	1.06

Table 4: Park-Newman Strength Score, Generalized Strength Score, and Generalize Strength Score with Time computed with alpha=90% of max., beta=.8 for the sample data. With time period considered, player 2's rating decreases from tournament 1 to tournament 2. The value of beta or the particular function *h(t)* controls the 'reactivity' of the rating.

7. Results

For a complex system to be worth its salt, it must produce significantly better results than a simpler system. In this section the *Strokes Gained to the Field* system for

measuring golfers' skills is compared to the Network system proposed here. The Network system produces skill measures that are significantly more correlated with future performance in eleven of the fourteen categories of skill.[7] In all categories except the 'other' category, the Network method produces skill measures that are more correlated with future success.

For this experiment, the data available up until time t is used to compute measures of player's skills in different categories, which are then used to predict future performance of the golfers in time t+1. The categories were chose a priori using knowledge of the game of golf; they are shown in Table 5.

Category	**Category Code**	**Percent of Shots**
Green or fringe, less than 5 feet	green0	20.0%
Green or fringe, between 5 and 10 feet	green5	6.2%
Green or fringe, between 10 and 20 feet	green10	7.4%
Green or fringe, more than 20 feet	green20	9.6%
Tee Shots on par 3s	tee3	5.8%
Tee Shots on par 4s and 5s	tee45	19.7%
Intermediate Rough or Primary Rough, less than 30 yards	rough0	3.4%
Intermediate Rough or Primary Rough, between 30 and 125 yards	rough30	1.9%
Intermediate Rough or Primary Rough, greater than 125 yards	rough125	4.9%
Fairway, less than 100 yards	fairway0	4.3%
Fairway, between 100 and 180 yards	fairway100	8.1%
Fairway, more than 180 yards	fairway180	4.5%
Bunker	bunker	3.6%
Other	other	0.5%

Table 5: Categories for which player skills will be computed. The 'rough' categories include both intermediate and primary rough and the 'green' categories include both fringe and green. The category code in reference to the categories is used for the rest of the paper.

[7] Significant at the alpha = .05 level.

Shotquality (or Strokes Gained) baseline measurements currently used on the PGA TOUR are supplied in the data. Following the convention established in the *Strokes Gained to the Field* system, the average strokes gained of the field in each category during each particular round is subtracted from the shotquality measurement supplied. These measurements are taken as observations of a player's skill and a weighted-average, with more weight given to more recent observations, is used as a predictor of future results. In using comparisons between shotquality and the field's average shotquality for a particular type of shot during a particular round, the *Strokes Gained to the Field* system fails to take into account the quality of the field during the round or potential differences in difficulty of certain shots between holes within a round. The network system proposed here takes both of these subtleties into consideration.

For the sake of computational feasibility, each time period is specified to contain a group of four tournaments. After each group of tournaments has occurred, ratings of each player's skill are computed using the data up to the following time period. These ratings are then used to predict the players' results in the following tournament group. Observations for all tournaments within each time period contribute equally to ratings computed by both methods. Observations dating back 28 tournament groups (about 3 years) are used in computing each rating.

There are 128 tournament groups in the data for which the Strokes Gained Baseline measurements (those currently used by the PGA TOUR) are available. The first 28 tournament groups are not used since only the Network system would have a full 28 tournament groups worth of data available prior to tournament group number 29. Of the 100 tournaments groups remaining, 50 are allocated as training data – for fitting the parameters associated with each method – and 50 are set aside as testing data. There are over 24,000 player-tournament observations in both the training and testing sets.

The weights for the *Strokes Gained to the Field* system are determined by a half-normal density, normalized so that at distance t=0, the weight is 1. Figure 5 shows this function graphically at various values of beta, a shape parameter. The beta that provides the strongest correlation with future results is chosen for each skill category.

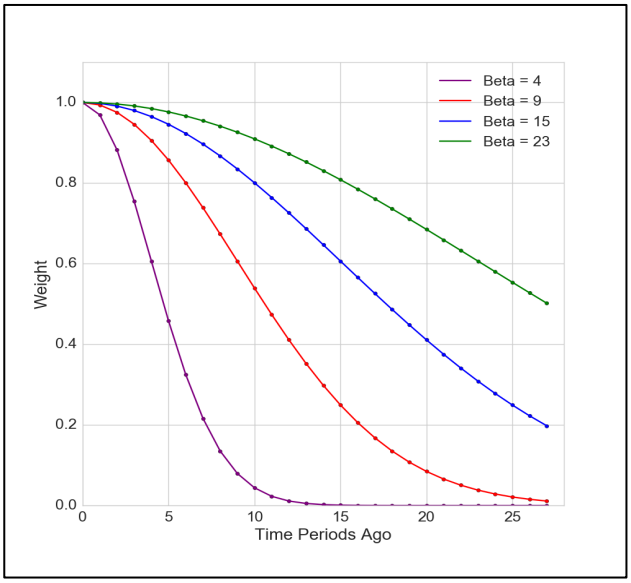


Figure 5: Half normal function used to specify the relative importance of more recent observations versus earlier observations for both the Strokes Gained to the Field method and the Network method.

On the same data, the Network system presented here is used to compute the ratings of each player in each skill category. Despite combining the shots from the primary rough and intermediate rough into one 'rough' category and the shots from the green and fringe into one 'green' category, only shots taken from the same turf are considered to be valid comparisons. More experimentation is needed to determine if this is the best approach. Using the training data, the best combination of parameters is selected. All in all, there are six parameters that were used to calculate the rankings. They are summarized in Table 6.

The dependent variable, the finishing position of a player in a tournament, is computed by ranking all the players by their stroke totals in each tournament (giving preference to those who played more rounds) and taking a percentile. The competitors in each tournament are ranked according to the ratings computed by both methods in each of the fourteen skill categories. These rankings constitute the independent variables. The correlations of each of the independent variables with the dependent variable on the test data are presented in Table 7. In determining whether the differences in correlations are significant between the two methods, two tests are conducted in addition to a bootstrap

Name of Parameter	Description
Epsilon	upper limit on how far shots can be away from one another and be considered an observation (proportional to the distances of each shot from the hole)
E_d	negative exponent to raise the distance between the shots to in computing the weight of an observation
E_t	negative exponent to raise the difference in time between the two shots to
W_d	weight for relative importance of the distance weight versus the time weight
β	shape parameter of a half-normal curve that determines the proportions of observations inherited across time periods
α	Park, Newman parameter that determines the relative value of indirect wins

Table 6: Descriptions of the parameters used in the Network method.

confidence interval of the difference of the correlations. Since both sets of independent variables are being correlated with the same dependent variable, the correlation coefficients are not independent of one another. This significance test has been studied by Steiger (1980) and more recently by Zou (2007). The Steiger test of significance results in a z-score and a corresponding p-value. The Zou test results in an interval which indicated significance if it does not contain 0. One-sided test results are used to test the hypothesis that the Network raking is more correlated with the dependent variable than the Strokes Gained to the Field ranking.

The network method is better in every shot category except the 'other' category, in which it is actually significantly worse. This is because the network system relies on having multiple shots from the same turf on the same hole and the same day. The 'other' shots are so rare that this strategy does not perform well since there are very few observations available. For this reason, the *Strokes Gained to the Field* method of

computing ratings will by used for the 'other' category in the regression to follow in the next section.

Category Code	Correlation Strokes Gained to the Field Method	Correlation Network Method	95 % C.I. of difference of correlations (bootstrap)	Z-Score using Steiger's test	P-Value using Steiger's test	Confidence Interval using Zou's test, 95% confidence
green0	0.037	0.046	(0.005, 0.012)	4.84	< .001	(0.005, 0.012)
green5	0.017	0.025	(0.003, 0.013)	3.25	.001	(0.003, 0.013)
green10	0.033	0.045	(0.006, 0.017)	3.92	< .001	(0.006, 0.017)
green20	0.039	0.044	(-0.001, 0.011)	1.65	.050	(-0.001, 0.011)
tee3	0.088	0.098	(0.006, 0.014)	4.54	< .001	(0.006, 0.014)
tee45	0.106	0.117	(0.006, 0.015)	4.30	< .001	(0.006, 0.015)
rough0	0.035	0.045	(0.001, 0.018)	2.24	.013	(0.001, 0.018)
rough30	0.047	0.051	(-0.007, 0.014)	0.74	.228	(-0.006, 0.014)
rough125	0.073	0.087	(0.006, 0.023)	3.31	< .001	(0.006, 0.023)
fairway0	0.069	0.072	(-0.004, 0.010)	0.79	.215	(-0.004, 0.010)
fairway100	0.072	0.079	(0.002, 0.011)	2.69	.004	(0.002, 0.011)
fairway180	0.083	0.090	(0.002, 0.012)	2.71	.003	(0.002, 0.012)
bunker	0.038	0.045	(0.000, 0.013)	1.91	.028	(0.000, 0.012)
other	0.034	0.014	(-0.041, -0.001)	-2.06	.020	(-0.040, -0.001)

Table 7: Correlations between the two skill measures and future tournament finishing position (on the test data) and tests to determine significance of the difference between the correlations.

7.1 Predicting Future Performance using Network Rakings

Since there are many different strategies one can take in computing ratings of players' skills using these data, it is important to apply the ratings to predicting the future to evaluate the efficacy of different methods.

Predicting the results of players on the PGA TOUR in individual tournaments is surprisingly difficult. Running a regression on finishing position using the rankings of players' skills produces a cross-validated R-squared of only 3.5%. Using the *Strokes Gained to the Field* method yields 3.4%. An average of the rankings produced by the two methods yields 3.6%. The results of the regression using the network method on the data

from the 50 tournament groups of the test data are shown in Table 8. Standard linear regression is used – no amount of regularization was helpful in terms of predictive accuracy.

'Course Profiles' were calculated by utilizing data from past tournaments to estimate the importance of each skill at each course in determining the finishing position of the competitors. This was done in an Empirical Bayesian style, using rank correlation of players' performances within each of the skill categories during a round with the round scores of the players as a proxy for importance. Although there is strong evidence that there is a difference in importance of the different skills in determining finishing position at different courses, including these variables in the regression as additional features interacted with the existing ranking in the corresponding skill categories did not improve the results. Perhaps in the future, with more observations of tournaments at the same courses, these 'course profiles' might prove useful in predicting the outcome of tournaments.

Variable Name	Coefficient	Standard Error	T-Statistic	P-Value
constant	0.26	0.01	28.7	< .001
tee45	0.12	0.01	15.0	< .001
green0	0.05	0.01	7.28	< .001
tee3	0.05	0.01	6.09	< .001
fariway0	0.04	0.01	5.31	< .001
fairway100	0.03	0.01	4.81	< .001
rough125	0.03	0.01	4.67	< .001
rough30	0.03	0.01	4.32	< .001
green10	0.03	0.01	4.03	< .001
green20	0.03	0.01	3.91	< .001
rough0	0.02	0.01	2.97	.003
fairway180	0.02	0.01	2.29	.022
other	0.01	0.01	1.52	.128
green5	0.01	0.01	1.19	.234

Table 8: Ordinary Least Squares regression with finishing position predicted from ranking in each of the skill categories. Both the dependent and independent variables are percentiles so the coefficients may be interpreted accordingly. For example, all else held constant the 0.05 coefficient on green0 predicts that a player's finishing position would improve by 5 percentage points upon improving from worst in the field to best in the field in this category.

The regression in Table 8 produces a cross-validated mean absolute deviation of 0.244. Since the dependent variable is a percentile, it can be interpreted that this average error of prediction of the model is equivalent to predicting a player might finishing a tournament in 25th place out of 100 competitors, yet his real outcome is 49th place (or 1st place). To further put this number in perspective, a prediction of the mean outcome - 0.504 – for every player in every tournament produces a mean absolute deviation of 0.250. Hence it is found that there is a large amount of unpredictability in the results of any single tournament, which is indicative of parity in the competition on the PGA TOUR.

7.2 Practical Use of Network Rankings

The methods utilized here to estimate the skills of the players normalize for the number of comparisons in each skill category. As a result, when computing top 10 ratings in a given skill category one will often observe that players who have very few observations are ahead of players with greater numbers of observations. This result is not surprising and can be explained by the familiar concept of regression to the mean. For practical use, requiring the number of observations of a golfer to be in the top 200 among golfers on tour is a reasonable choice in computing top 10 lists. In the appendix, this approach is taken to produce tables of the top 10 golfers in a few skill categories after the 2016 season and after the 2008 season.

Also included in the appendix are time series of the rankings of some notable golfers. One chart has both the *Strokes Gained to the Field* method and the Network method for a given skill category in order to show to degree of similarity in the rankings produced by each method. A web app that allows users access to these time series would aid both players and fans in gaining practical insights into the strengths and weaknesses of the players on the PGA TOUR.

7.3 Relative Importance of Skills on the PGA TOUR

The topic of explanatory value of the different skills on the PGA TOUR is taken up in Broadie (2012). Broadie utilizes a variance decomposition to conclude that the long game explains about 72% of the variability of golfers' overall skills. Broadie

acknowledges that variability does not equate with importance so the question of which skills are more important in determining results on the PGA TOUR has remained open.

With the results of the correlations of the skill ratings of golfers with future results and the results of the regression on future results presented earlier in this section, one might be tempted to come to the conclusion that the Tee Shots on par 4s and par 5s are the most important of the fourteen categories of skill. However, this conclusion makes an assumption about just what is meant by 'important'. It turns out that two reasonable but different meanings of importance lead to two different orderings of the skill categories, which provide insight into, and conform with common wisdom about, competition on the PGA TOUR.

One definition of relative importance is the predictive value of each of the categories of skills. With this definition in mind, the relevant measure of importance is the correlation between the rankings of the players' skill measures in a category with players' results in future tournaments. These correlations were shown in Table 7.

An equally valid definition of importance has to do with explanatory power. According to this definition, the correlation between performance in a category during a given time period and overall performance during the same time period is the measure of importance. One time period of interest on which to base this measure is a round of golf. The correlations between within round performance in a category of shots and overall round performance (field average round score minus player's round score) are displayed in Table 9 along with the correlations of skill measures and future performance that were previously displayed in Table 7.

Using the explanatory power definition of importance and considering a time period to be one round results in a starkly different ordering of the skill categories in which mid-range putting seems to be the most important skill. This seeming contradiction can be explained using a bit of golf acumen. When asked about the prospect of having a good performance prior to a big round, professionals often speak about the importance of 'getting a few putts to fall'. This truism is born out in the numbers.

Category Code	Explanatory Correlation	Explanatory Rank	Predictive Correlation	Predictive Rank
green10	0.29	1	0.04	9
green5	0.29	2	0.02	13
fairway100	0.27	3	0.08	5
tee45	0.25	4	0.12	1
tee3	0.25	5	0.10	2
green20	0.25	6	0.04	12
green0	0.23	7	0.05	8
rough125	0.19	8	0.09	4
fairway0	0.19	9	0.07	6
bunker	0.18	10	0.04	10
rough0	0.17	11	0.04	11
fairway180	0.17	12	0.09	3
rough30	0.13	13	0.05	7
other	0.07	14	0.01	14

Table 9: Pearson's correlation coefficients of within round performance in each skill category with overall round performance. 208,000 rounds worth of data was used. Predictive Correlation is the correlation of skill ranking and future tournament results – the same numbers as Table 7.

In predicting future performance, the ranking of players' abilities to drive the ball is most valuable. In explaining the results of single rounds, players' performances on the green are most valuable. A logical explanation of this phenomenon is that although putting performance is very important in determining relative performance on tour, there is more randomness involved in individual players' performances on the green during the course of one round. Measures of skill in tee shots and the long game are more useful towards prediction indicating more predictability of performance in these areas on a tournament-to-tournament basis. It is thus concluded that while putting performance is important in determining results, putting skill is relatively ephemeral compared with skills in the long game.

8. Conclusion

Data provided by the PGA TOUR is rich in both quality and quantity. The availability of these data allows researchers, players and fans more intimate understanding of the game. The problem of estimating the skills of the players on tour in the different aspects of the game is an interesting problem for Data Science.

This paper provides a novel approach to this problem through using direct and indirect comparisons of players' skills based on shots taken on the same hole and day in lieu of a universal baseline estimate of difficulty of shots. This approach is demonstrated to be more successful in predicting future results than the currently used method. In thirteen of the fourteen categories, skill measures computed using the new method are more correlated with future success than the current standard, with eleven of the fourteen comparisons statistically significant.

Currently, the PGA TOUR computes skill measures for four categories, considering entire seasons as time periods. In this paper, more fine-grained categories and more frequently updating measurements are computed that allow for even more detailed understanding of the relative skills of players.

In addition to inventing a method that produces more accurate estimations of the skills of players, this paper applies rigorous Data Science concepts like training and testing towards improving the current system. It is hoped that this work will help establish a rigorous standard of testing and prediction towards estimating the skills of golfers, and in applications of Data Science to golf in general.

References:

Broadie, M. 2012. *Assessing Golfer Performance on the PGA TOUR*. Interfaces, vol. 42, no. 2, p. 146-165.

Fearing, D., Acimovic J., Graves S. 2010. *How to Catch a Tiger: Understanding Putting Performance on the PGA TOUR*. MIT Sloan Research Paper No. 4768-10.

Park, J., Newman, M. 2005 *A Network-based ranking system for US college football*. Journal of Statistical Mechanics: Theory and Experiment, vol. 2005, no. 10, p. 100-14.

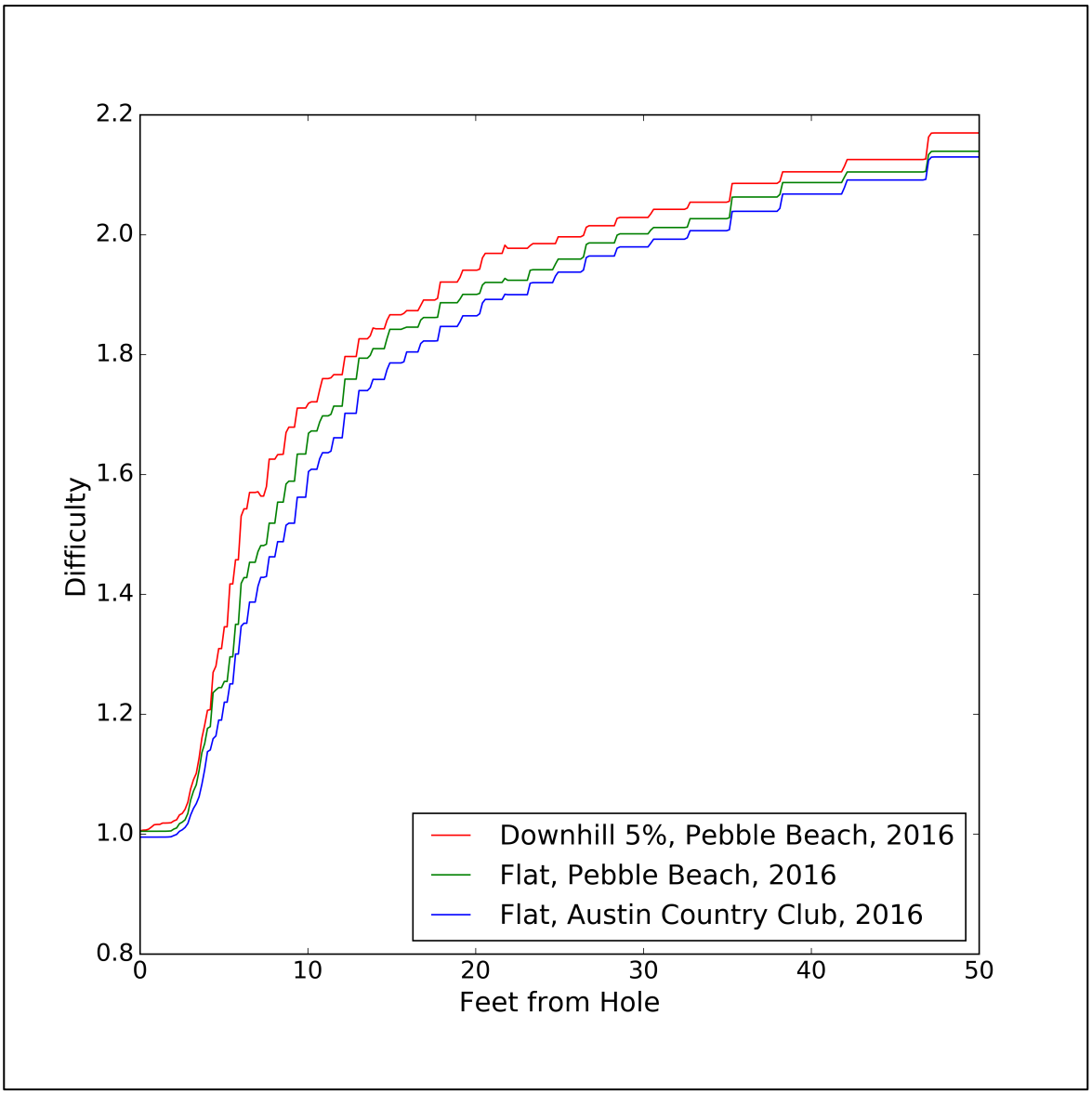
Steiger, J. H. 1980. *Tests for comparing elements of a correlation matrix*. Psychological bulletin, vol. 87, iss. 2, p. 245.

Stöckl, M., Lamb P., Lames M. 2013. *The ISOPAR Method*. Journal of Quantitative Analysis in Sports, vol. 7, iss. 1.

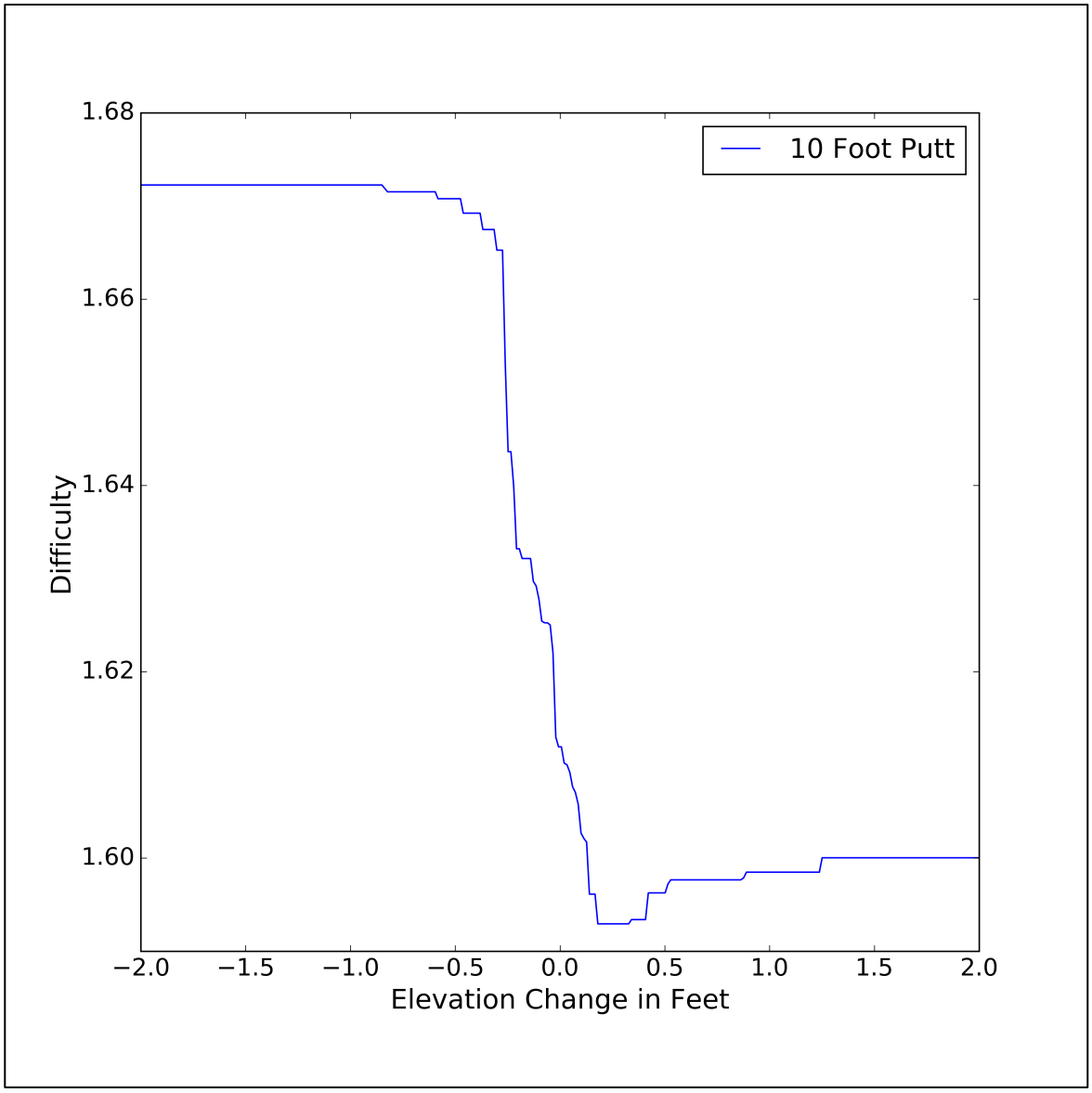
Yousefi, K., Swartz, T. 2013. *Advanced Putting Metrics in Golf*. Journal of Quantitative Analysis in Sports, vol. 9, iss. 3.

Zou, G. Y. 2007. *Towards using confidence interval to compare correlations*. Psychological Methods, vol. 12, iss. 4, pp. 399-413.

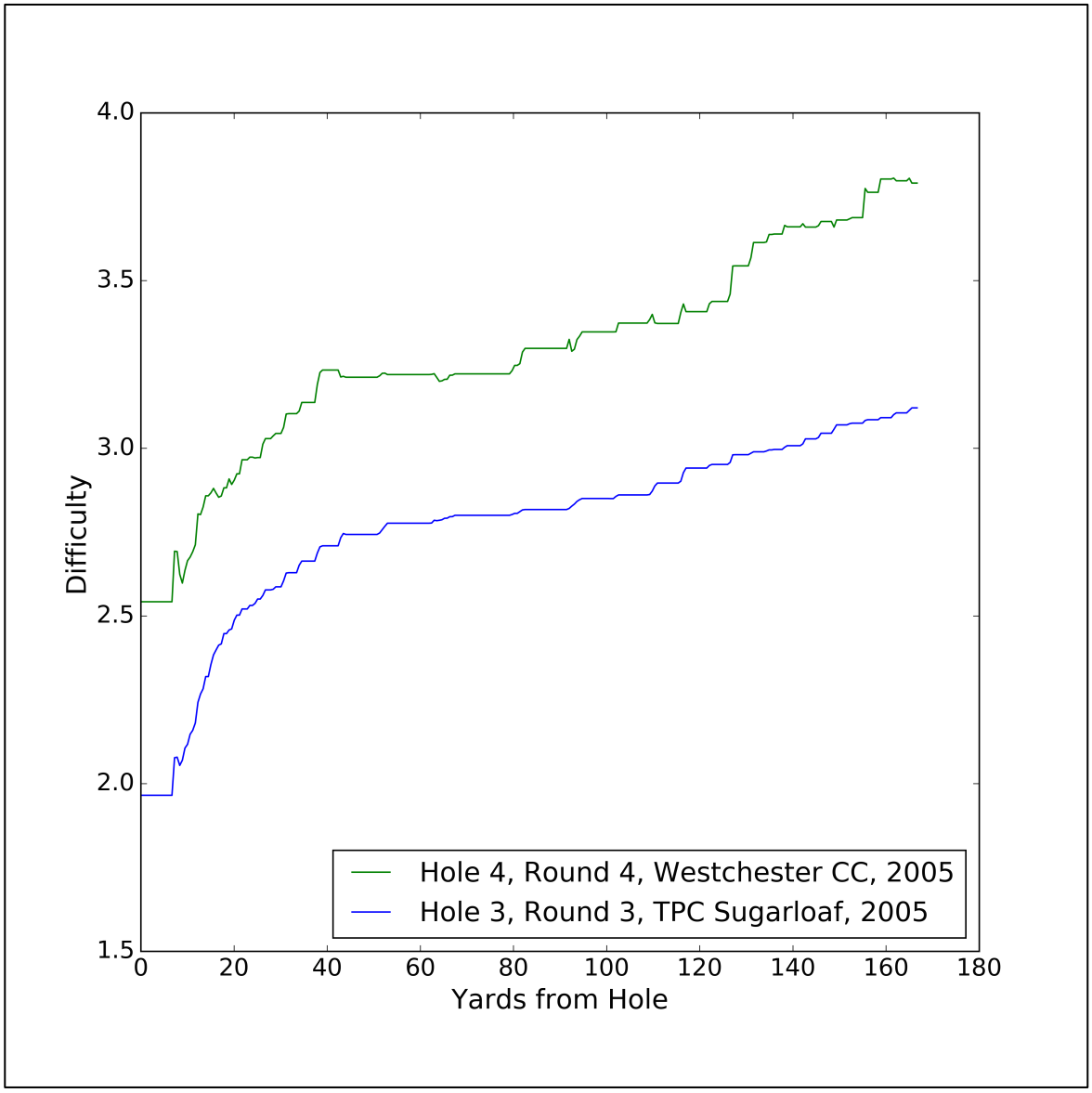
Appendix A. Plots and Tables



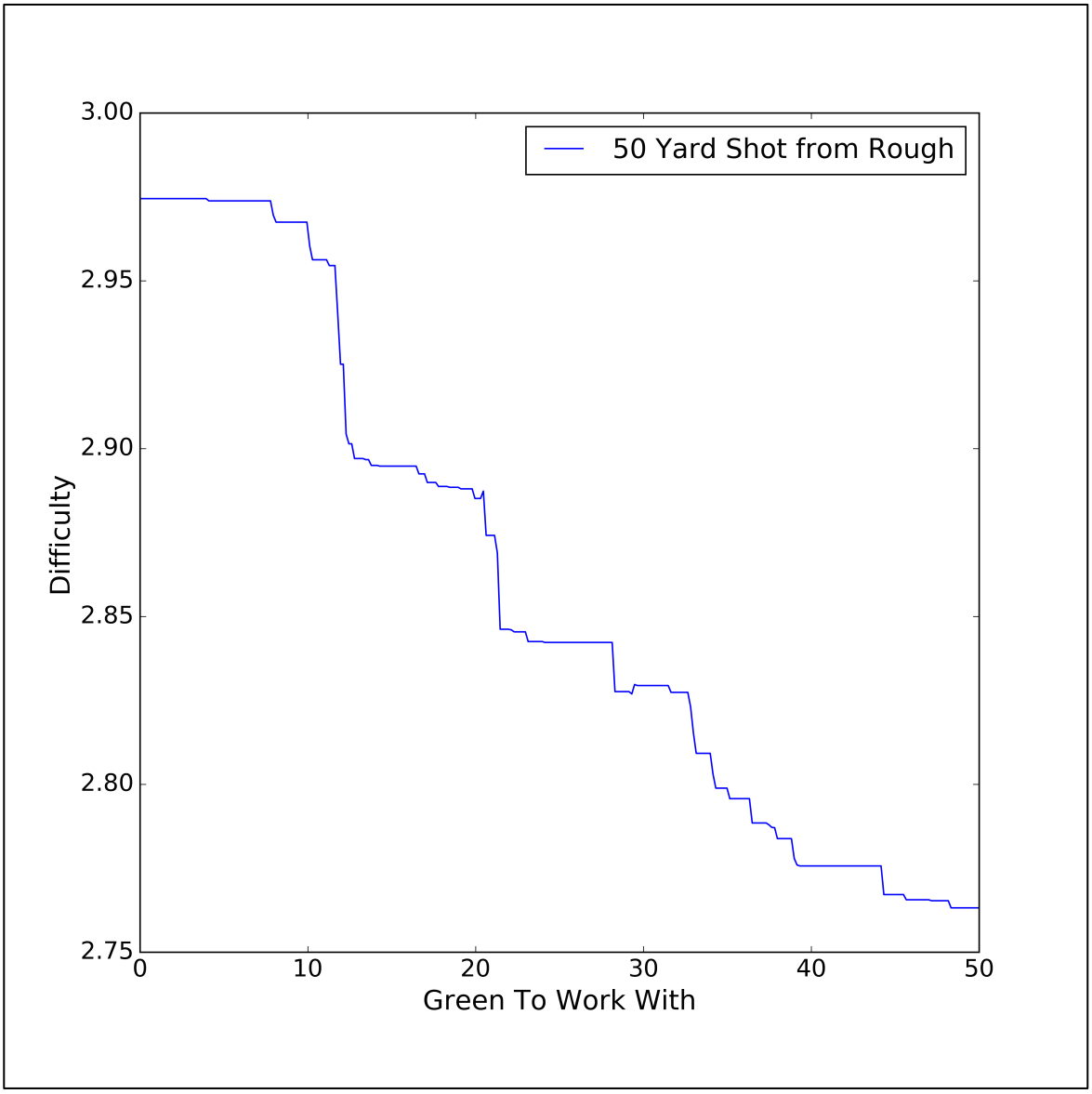
The model's predictions for difficulty of shots from the Green for three different situations at various distances. Important to note that it is not disambiguating the strength of the players playing on the courses from the difficulty of the course.



The model's prediction of difficulty for a 10-foot putt at no course – this can be interpreted as an average across all courses. Elevation change varies on the x-axis. The plot shows that the easiest putts are very slightly uphill.



The model's prediction of difficulty for shots from the Primary Rough for two different hole-rounds. These two hole-rounds were selected because they show the biggest difference in predicted difficulty across all the data. Again, it must be stressed that there is no effort made to disambiguate between strength of players and difficulty of course since the goal of the model is to allow the fair comparison of two shots taken on the same turf, hole, and round. Thus difference in assigned difficulty could be an indication of either true difficulty or a weak field or some combination of the two.



The model's prediction of generic (not specific to one hole-round) difficulty for a 50 yard shot from the Primary Rough with varying amounts of the Green To Work With.

Top 10 in Six of the Categories After the 2016 Season

Green or fringe, less than 5 feet

	Player	Rating
1	O'Hair, Sean	.9997
2	Senden, John	.9996
3	Todd, Brendon	.9995
4	Malnati, Peter	.9993
5	English, Harris	.9991
6	Merritt, Troy	.9991
7	Kisner, Kevin	.9990
8	Laird, Martin	.9990
9	Kim, Michael	.9988
10	Romero, Andres	.9988

Tee Shots, Par 3s

	Player	Rating
1	Matsuyama, Hideki	.9784
2	Stenson, Henrik	.9767
3	Kokrak, Jason	.9759
4	Rose, Justin	.9747
5	McIlroy, Rory	.9744
6	Casey, Paul	.9741
7	Garrigus, Robert	.9730
8	Watson, Bubba	.9729
9	Mickelson, Phil	.9727
10	Reavie, Chez	.9719

Green or fringe, between 10 and 20 feet

	Player	Rating
1	Day, Jason	.9846
2	Baddeley, Aaron	.9839
3	Kuchar, Matt	.9809
4	Choi, K.J.	.9808
5	Spieth, Jordan	.9803
6	Snedeker, Brandt	.9795
7	Donaldson, Jamie	.9793
8	Byrd, Jonathan	.9774
9	Hoffman, Morgan	.9767
10	Stricker, Steve	.9760

Rough, more than 125 yards

	Player	Rating
1	Watson, Bubba	.9852
2	Finau, Tony	.9800
3	Rose, Justin	.9776
4	Fowler, Rickie	.9769
5	Mickelson, Phil	.9766
6	Kaymer, Martin	.9766
7	Donald, Luke	.9766
8	Na, Kevin	.9759
9	Flores, Martin	.9751
10	Senden, John	.9751

Bunker

	Player	Rating
1	Weir, Mike	.9432
2	Day, Jason	.9388
3	Haas, Bill	.9388
4	Harrington, Padraig	.9326
5	Kuchar, Matt	.9314
6	Donald, Luke	.9310
7	Snedeker, Brandt	.9305
8	Na, Kevin	.9291
9	Appleby, Stuart	.9288
10	Garcia, Sergio	.9285

Tee Shots, Par 4s and 5s

	Player	Rating
1	McIlroy, Rory	.9882
2	Johnson, Dustin	.9873
3	Rahm, Jon	.9843
4	Holmes, J.B.	.9798
5	Watson, Bubba	.9793
6	Garcia, Sergio	.9787
7	Rose, Justin	.9779
8	Finau, Tony	.9776
9	Scott, Adam	.9775
10	Lovemark, Jamie	.9773

Top 10 in Six of the Categories After the 2008 Season

Green, between 10 and 20 feet

	Player	Rating
1	Woods, Tiger	.9878
2	Wilson, Dean	.9832
3	Crane, Ben	.9830
4	Mattiace, Len	.9820
5	Ames, Stephen	.9820
6	Heintz, Bob	.9819
7	Allan, Steve	.9818
8	Parnevik, Jesper	.9817
9	Atwal, Arjun	.9816
10	Jacobson, Freddie	.9811

Tee Shots, Par 3s

	Player	Rating
1	Woods, Tiger	.9689
2	Lehman, Tom	.9643
3	Mediate, Rocco	.9627
4	Mickelson, Phil	.9625
5	Kim, Anthony	.9620
6	Funk, Fred	.9603
7	Els, Ernie	.9594
8	Bohn, Jason	.9591
9	Quinney, Jeff	.9583
10	Campbell, Chad	.9583

Green, more than 20 feet

	Player	Rating
1	Faxon, Brad	1.011
2	Olazabal, Jose Maria	1.010
3	Woods, Tiger	1.010
4	Garcia, Sergio	1.008
5	Baddeley, Aaron	1.007
6	Byrd, Jonathan	1.007
7	Els, Ernie	1.006
8	Geiberger, Brent	1.006
9	Donald, Luke	1.006
10	Curtis, Ben	1.006

Tee Shots, Par 4s and 5s

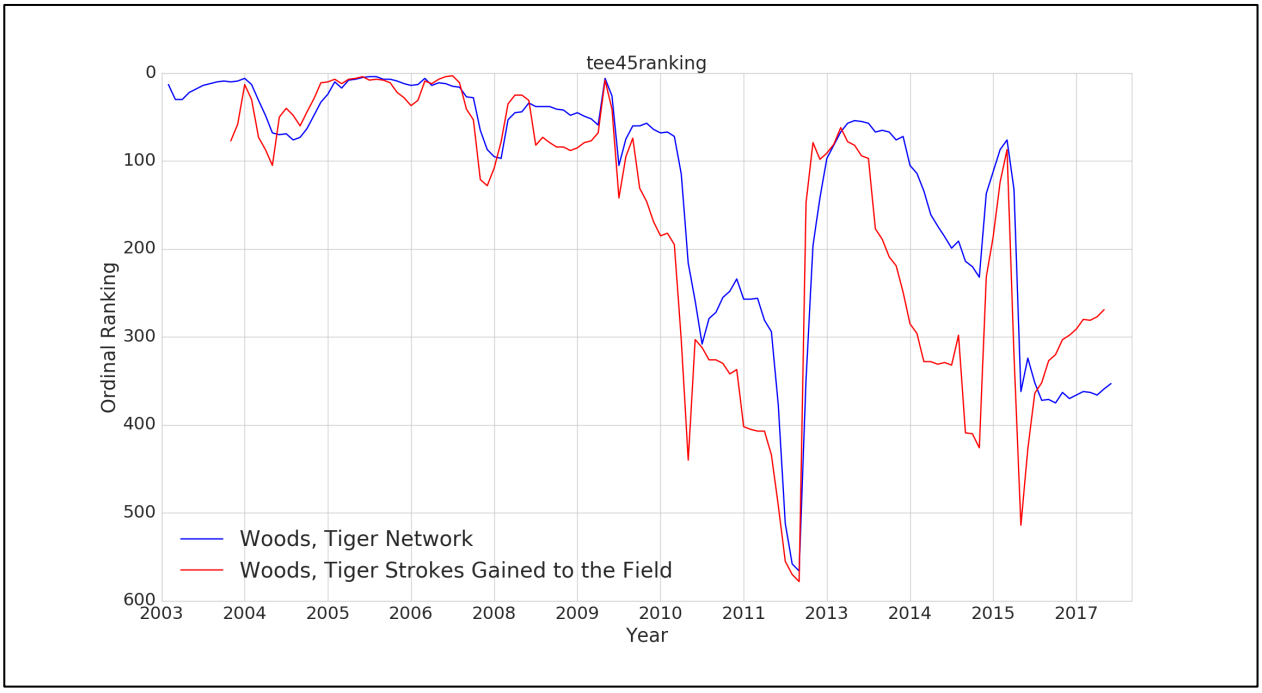
	Player	Rating
1	Holmes, J.B.	.9809
2	Watson, Bubba	.9783
3	Weekly, Boo	.9781
4	Wetterich, Brett	.9761
5	Warren, Charles	.9745
6	Allenby, Robert	.9744
7	Scott, Adam	.9740
8	Mahan, Hunter	.9740
9	Durant, Joe	.9739
10	Glover, Lucas	.9738

Rough, between 30 and 125 yards

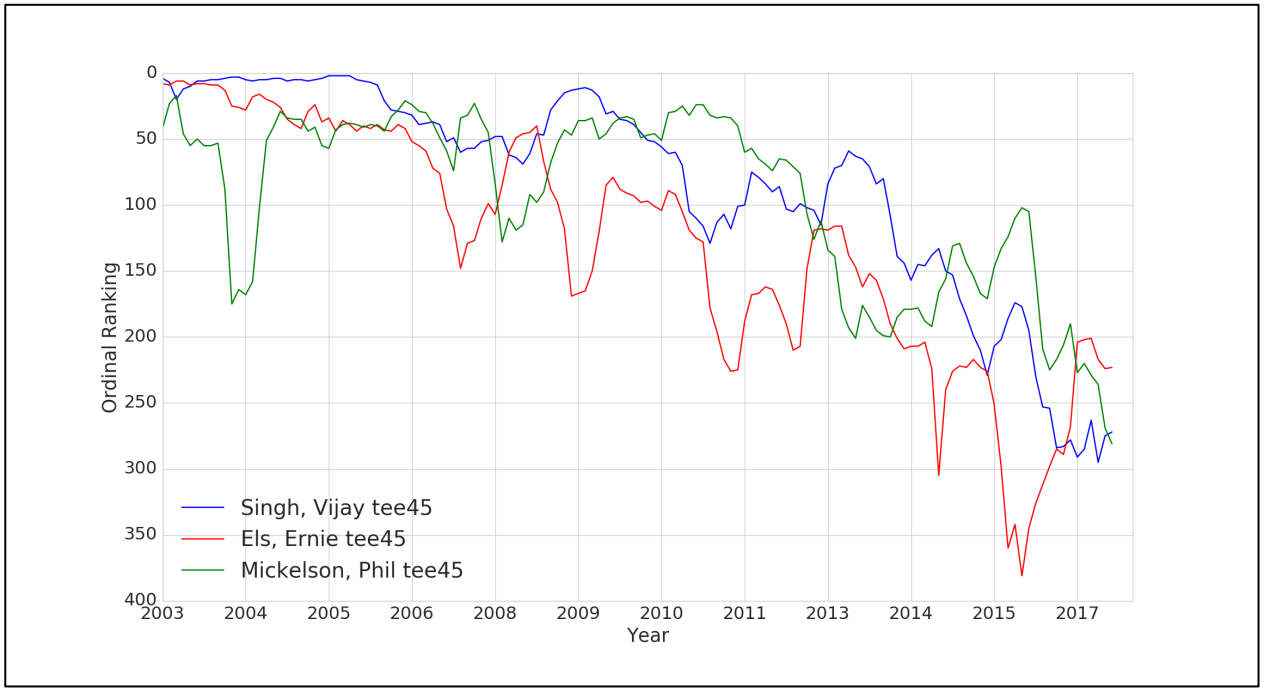
	Player	Rating
1	Mickelson, Phil	1.036
2	Duke, Ken	1.017
3	Kim, Anthony	1.015
4	Hayes, J.P.	1.014
5	Bryant, Bart	1.012
6	Woods, Tiger	1.011
7	Choi, K.J.	1.011
8	Verplank, Scott	1.010
9	Olazabal, Jose Maria	1.010
10	Sabbatini, Rory	1.009

Fairway, more than 180 yards

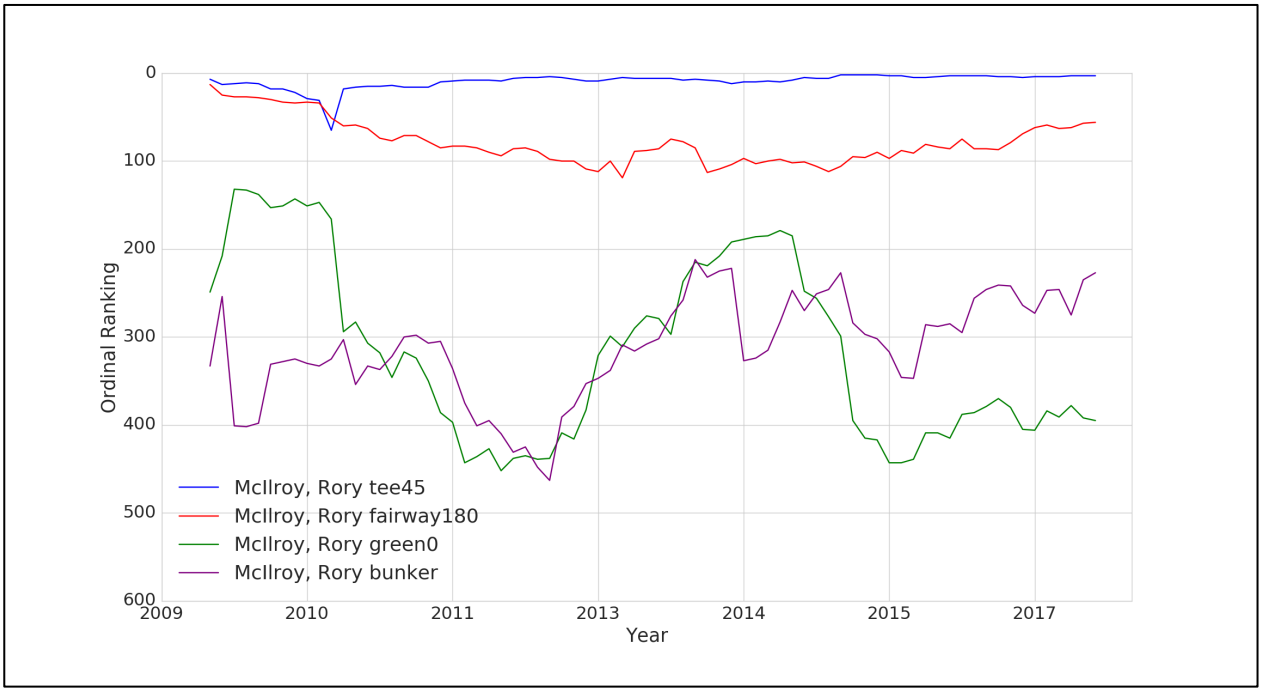
	Player	Rating
1	Woods, Tiger	.9776
2	Mickelson, Phil	.9740
3	Garcia, Sergio	.9738
4	Garrigus, Robert	.9697
5	Stadler, Kevin	.9674
6	Els, Ernie	.9672
7	Haas, Bill	.9664
8	Mahan, Hunter	.9652
9	Trahan, D.J.	.9644
10	Watson, Bubba	.9637



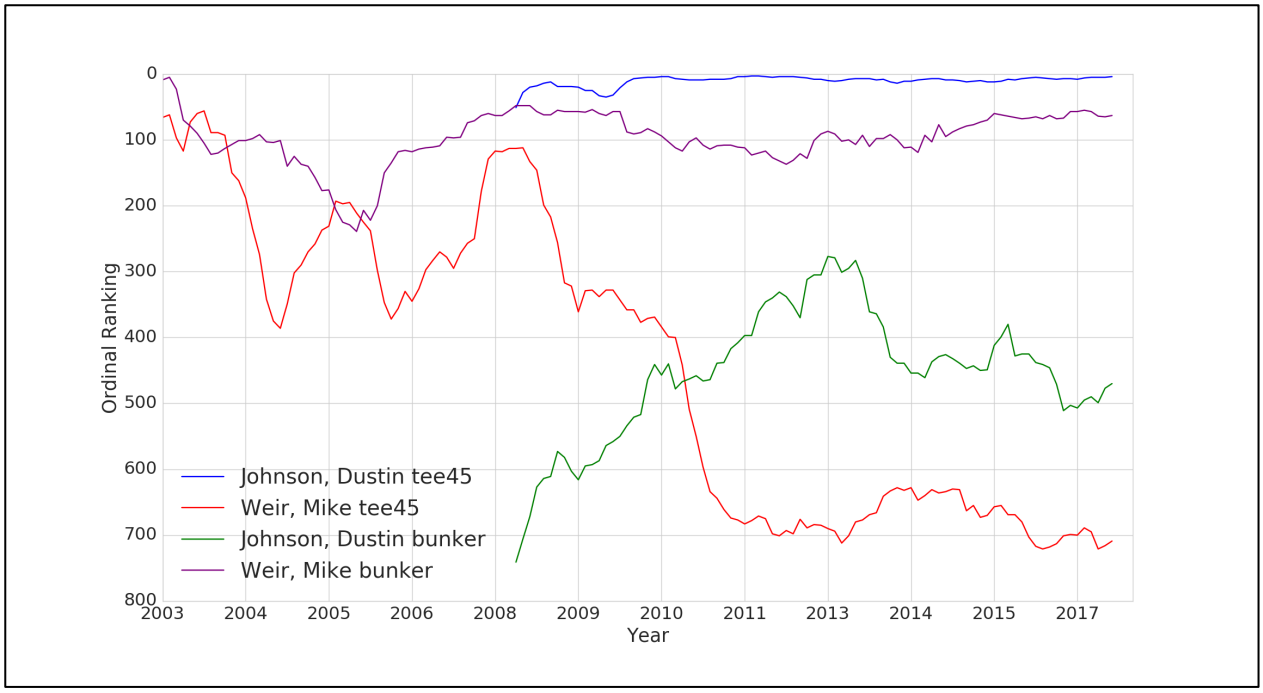
Tiger Woods' ranking in tee shots on par 4s and 5s using the Network system and the Strokes Gained to the Field method. The Network system mostly ranks Tiger higher, which is consistent with the fact that it takes into account strength of schedule.



Tee Shot par 4s and par 5s ranking for a few notable golfers.



Rory McIlroy's ranking in a few categories.



Comparing the strength's and weakness in the games of Dustin Johnson and Mike Weir.

Appendix B. Preprocessing Steps

There are both anomalies present in the raw data due to errors during collection, as well as missing data that are useful for analysis. A number of preprocessing steps were applied to fix errors and impute missing data. There are many player-holes in the data for which there are more shots recorded than the score of the player on the hole. In order to maintain the integrity of the data, all player-holes for which the number of shots in the data did not match the recorded score of the player on the hole were dropped.

Additionally, neither the coordinates of the tee box nor of the hole are present in the data. However, distance from the hole and distance that the ball traveled is present in the data. Thus, the coordinates of the hole and the tee box can be imputed. Lastly, any player-hole for which there was any shot for which the distance traveled was not in reasonable agreement with the coordinates recorded was dropped. Dropping the entire player-hole when there was an anomaly made the downstream analysis much easier. These cleaning steps reduced the size of the data by about 15% leaving just over 14 million anomaly-free shots. All code to reproduce this cleaning procedure is available.[8]

[8] https://github.com/adamwlev/Rank_a_Golfer

Appendix C. The Unintended Consequence of fitting a Complex Model to the same observations it is meant to access the difficulty of

In comparing the quality of two shots, both the estimates of difficulty of the shots at the current locations and at the locations that the balls travel to are important. With a large amount of detail available to a model – identification of a specific hole during a specific round perhaps – the potential to overfit the data is a concern. This level of detail and the fact that there might be very few shots taken from a specific turf during a specific round on a specific hole produces an unintended consequence.

If a model is fit to all the available data and then is used to produce estimates of difficulty for each of the shots in the data set, the model will have 'seen' all of the observations before making 'predictions'. The true outcome of each observation will thus have an effect on the 'prediction' for the observation. If a model contains very few features and plenty of observations, this is not much of an issue since the effect that a single data point has out of thousands (or millions) of observations in low dimensional space is typically minimal for most algorithms. However, with more and more features in a model, the effect of a single observation on the prediction that results from using the exact same combination of features can be substantial. In high dimensional space, the density of observations is low so the effect of one data point can be much higher. The unintended consequence is that the most important observation in assessing the difficulty of a shot is the actual true value of the number of shots taken from the location. This true value is not an unbiased indicator of the difficulty of a shot because it is not a shot taken by a random golfer; it is taken by the same golfer who took the previous shot.

An example will help illustrate the issue: if two players both take a shot from the fairway on the same day and on the same hole and one player ends up on the green while another player ends up in the bunker, estimates of difficulty from both of the spots on the fairway, the spot on the green and the spot in the bunker are needed to compare the quality of the two shots. First, to take an extreme example, let's say the player who hit it in the bunker hits it from the bunker in to the hole in one shot. If a complex model that is used to produce the estimate of difficulty of the shot from the bunker has been fit to the fact that a player took only one shot from the bunker on this hole and on this day, it will likely

underestimate the true difficulty of this shot. This consequence will manifest itself by overestimating the quality of the player's shot from the fairway to the bunker and underestimating the quality of the player's shot from the bunker to the hole. More critically, if this player is consistently a superior bunker player so that he is constantly taking relatively few shots to get in the hole from the bunker, driving down the estimated difficulty of shots from the location of his bunker shots, the system will consistently overestimate the quality of the shots that land him in the bunker while underestimating the quality of his bunker shots and thus underestimate his true skill in bunker play.

In this paper, this consequence is mitigated by not using a model that has been trained using the data from a specific hole and round to produce the estimates of difficulty for the shots from this hole and round. The shots taken on the same hole and the same round are considered to be a group. Estimations of difficulty for a group are produced using models that are trained using a subset of the data not including the group. Specifically, a 15-fold grouped cross-validation-prediction strategy is implemented. The choice of fifteen folds balances the desire to include as much information relevant to the estimation of difficulty - such as observations on the same course, the same hole (in other rounds or years) and during the same round - as possible with computational burden.