**1. Introduction**

The game is golf is a test of a variety of skills – driving off the tee, approach shots from the fairway or rough, and putting to name a few. Precise estimation of the skill of the players in the various aspects of the game is useful for a variety of reasons. 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.

This paper improves what is currently being done to estimate the skill levels of the players on the PGA Tour. 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 possibilities towards understanding the professional game in greater depth statistically. Up until detailed shot-level data was collected, it was impossible to quantify how the distinct skills determine golfers’ results.

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 everyone’s 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 3 of this paper.

Quantifying players’ skills is a pursuit of estimating latent variables. Competitive play is not setup in a way that makes this estimation 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.

The challenges involved in this modeling problem will be detailed first. Then a novel approach will be given. This approach is then backed with evidence that demonstrates its success.

**2. Dataset**

As mentioned, the dataset 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 available and will be 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.

|  |  |
| --- | --- |
| **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% |

|  |  |
| --- | --- |
| **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 here

means number of unique hole-day combinations.

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

*2.1 Preprocessing Steps*

Like with any data collected by humans, there were plenty of anomalies present in the data. There were many player-holes in the data for which there were more shots recorded than the score of the player on the hole. These extra shots resulted from errors in the recording of the data. 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 were present in the data. However, distance from the hole and distance that the ball travels is present in the data. Thus, the coordinates of the hole and the tee box could be imputed from the data. 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.[[1]](#footnote--1)

**3. Strokes Gained**

Before the detailed shot-level data and the Strokes Gained concept, statistics used to quantify specific skills in golf included Driving Distance, Fairways Hit, and Greens In Regulation (GIR) to name a few. To illustrate the ambiguity 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 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:

.

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, what is it good for?**

Modeling the difficulty of a shot is challenging for a few reasons. The first of which is that 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.

Additionally, when fitting a model for difficulty that contains information that distinguishes between different courses, there is a potential for erroneous interpretation 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).

Similarly, attempts to use spatial clustering or nearest-neighbor type algorithms 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.

For these reasons, producing unbiased measurements of difficulty of a shot is very challenging. In the rest of this section, previous attempts at this task will be outlined. Then, a new model with a subtle change in intention will be presented. This model will sacrifice unbiased measurements of difficulty of any single shot in favor of fairly estimating the skill of the golfers *relative* to one another in the network of all PGA Tour players.

*4.1 Previous models for difficulty of a shot*

Broadie (2011) 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 (2011), a model for distinguishing between course-round difficulty and player skill was done at a global level – estimating total strokes gained without allowing for the possibility that particular types of shots might be more or less difficult at certain courses or certain players 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 the author subsequently, strokes gained statistics currently used on tour are adjusted by the average strokes gained performance of the field for each category of shot for 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) model 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 (2011), 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. (2011) 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 either of the biases involved with this approach that were discussed above – the varying quality of a field and the bias for desirable locations to more frequently contain the shots of more capable players.

Finally, Yousefi and Swartz (2012) 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 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 shot and that change through time, the number of parameters to estimate can become immense.

Instead of attempting this type of model, this paper acknowledges that the estimates of difficulty of a given shot on a given hole and on a given day are not directly useful in comparing the quality of the shot to the quality of a different shot that was taken on a different hole and on a different day. 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 5 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, there are plenty of different approaches to making a model. Different models could be made using the data collected from each hole-day. Different models could be made using only the data from each tournament. There are different options. Building a model using only limited data – such as the data only from the hole-day or tournament – has the potential advantage of being less biased, but would come with a cost of having 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*

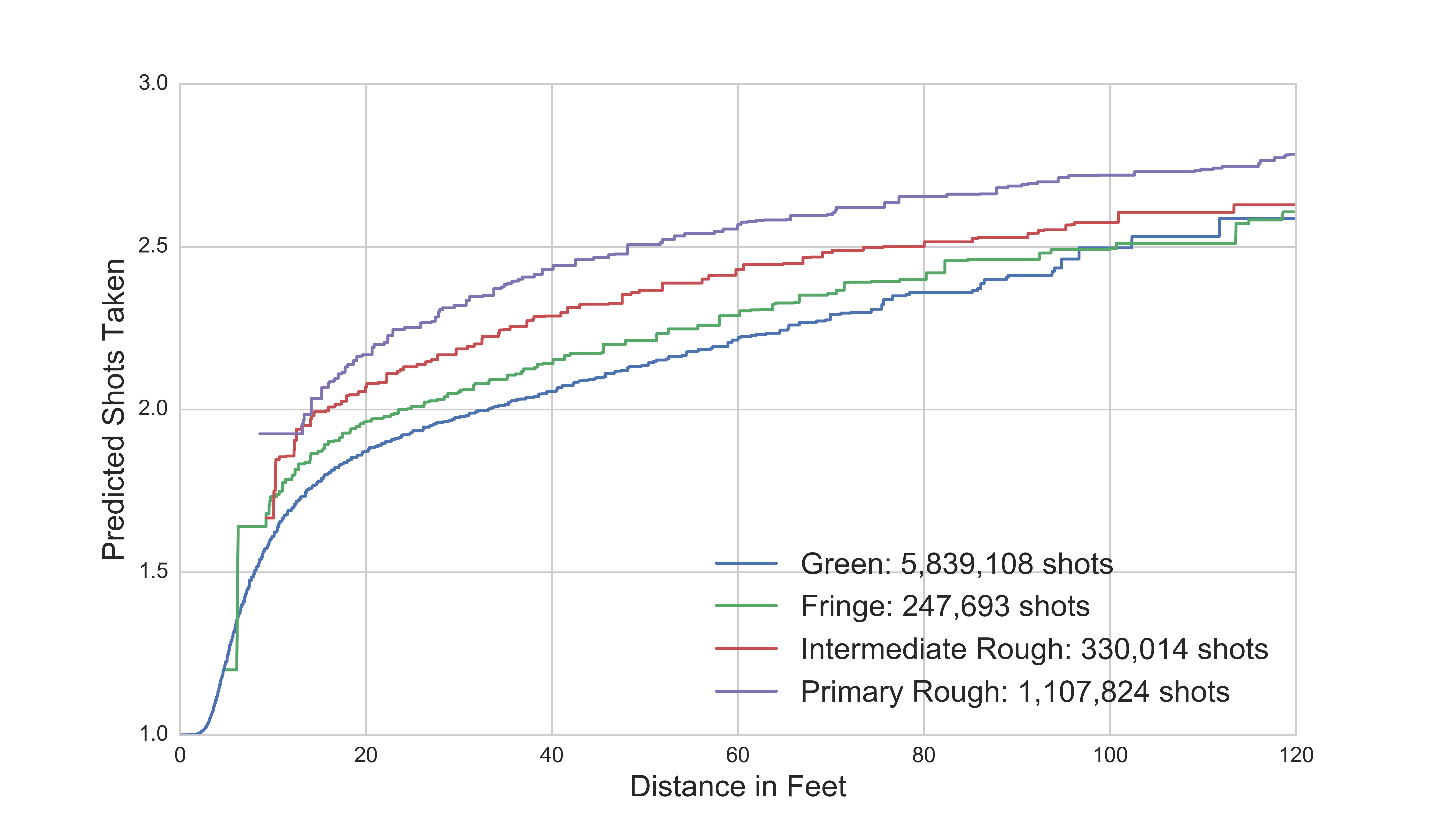
Distance is the most important feature for predicting difficulty of shot. The relationship between distance and difficulty of a shot is highly non-linear and is different for different turfs. Fitting various regression models is possible to visualize this relationship. One that fits the problem reasonably well is isotonic regression. Isotonic regression is a non-parametric method 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 are 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.

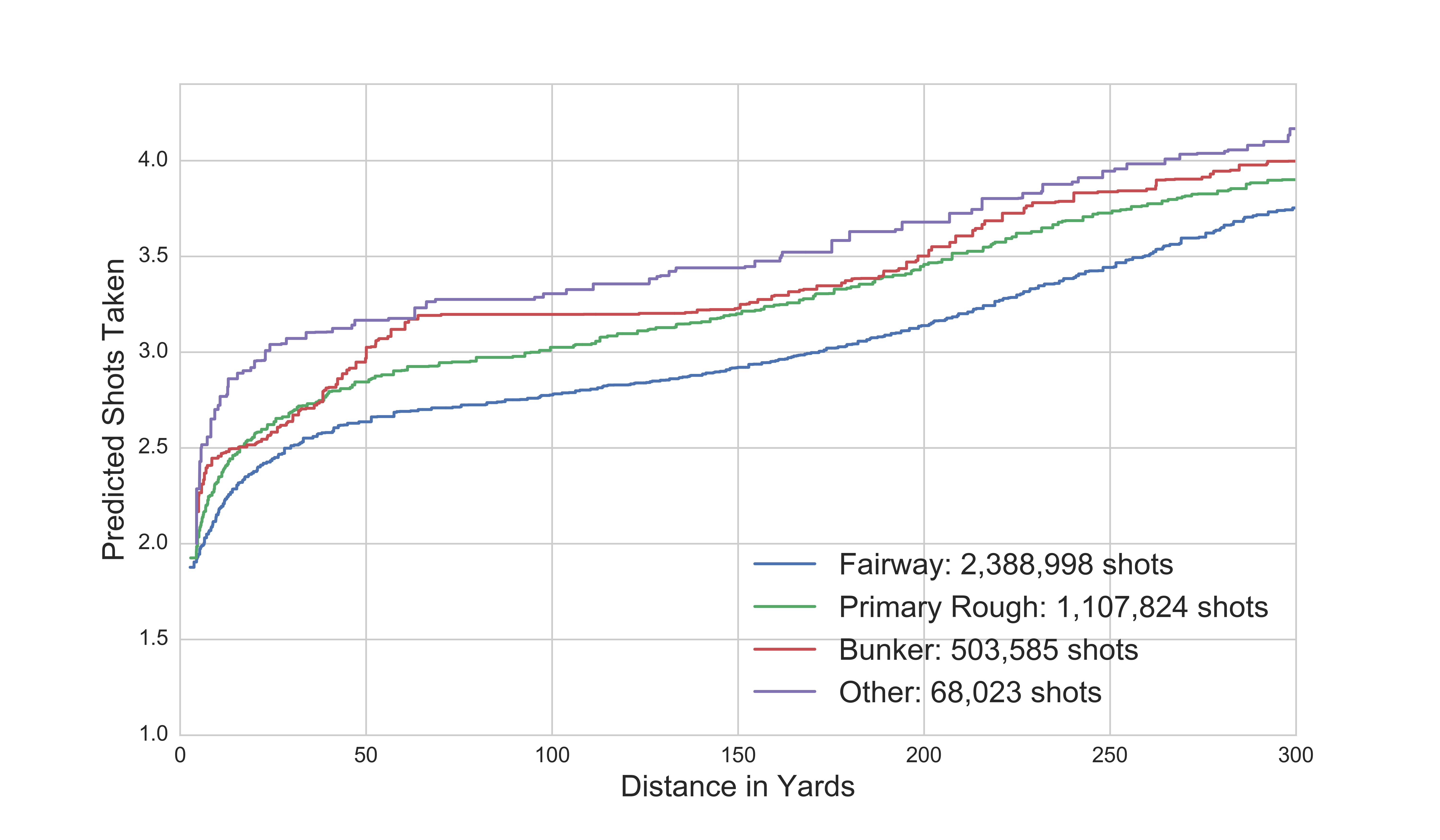


Figure 2: Isotonic Regression models for shots taken from longer distances. Close inspection reveals that shots from the Primary Rough are more difficult than shots from the Bunker for distances between 20 and 40 yards, while less difficult for other distances.

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. In principle one model for all shots could have been fit with indicator variables for each turf. 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 (non-course or round specific) features to predict difficulty of a shot is limited. Slope 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.

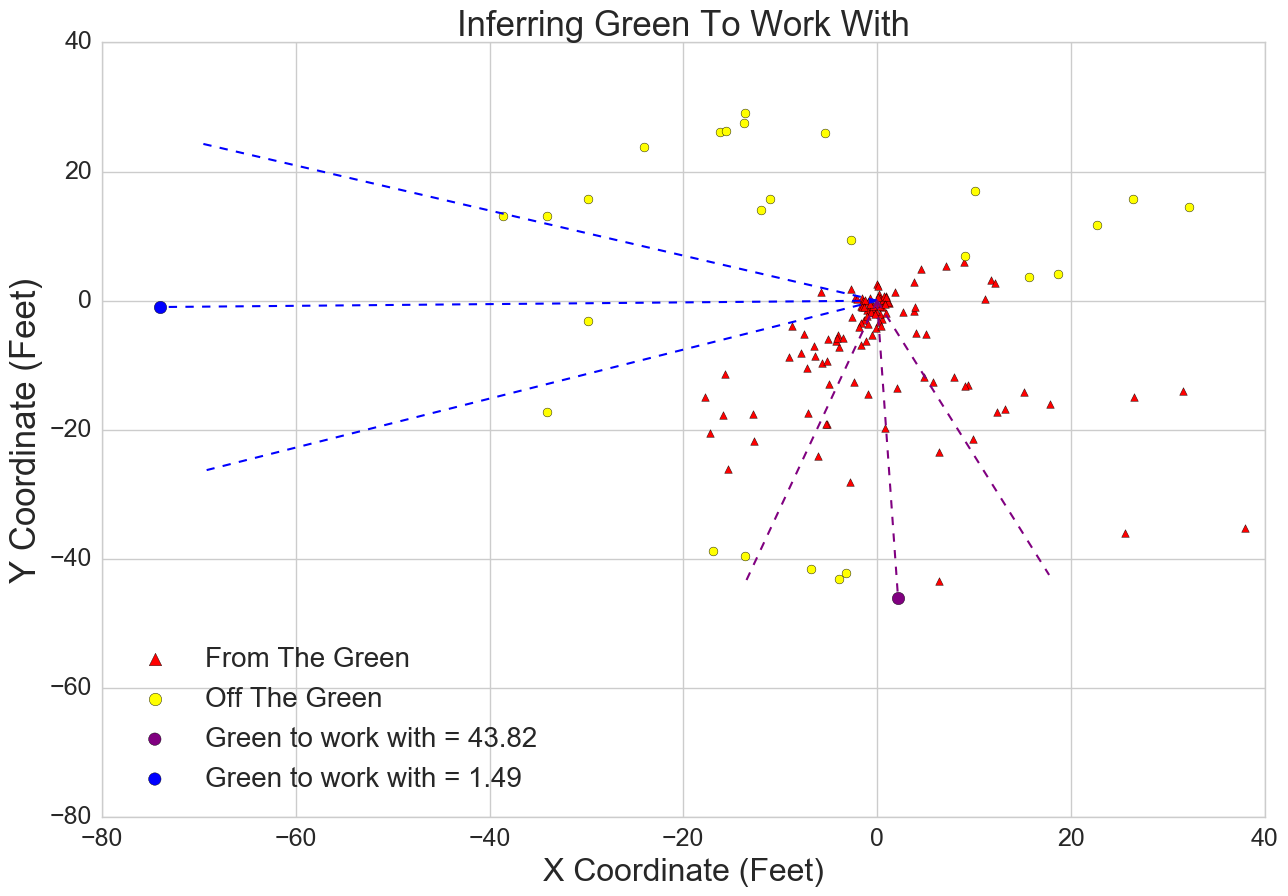


Figure 3: The purple point has an inferred Green To Work With of about 44 feet, while the blue point has an inferred Green To Work With of only 1.5 feet. This corresponds with a more difficult angle for the blue point.

The algorithm 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 more difficult to apply spin to the ball and control the run out. Results displaying the added benefit of slope and Green To Work With will be displayed in the next section.

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 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.[[2]](#footnote-0)

Special attention was paid to the strategy used to produce the estimates of difficulty to be used in the subsequent skill estimation process. As mentioned earlier, the model must produce estimates of difficulty that can be used to fairly compare shots taken on the same turf, on the same hole, and on the same day. In comparing the quality of two shots, both the estimates of difficulty of shots at the current locations and at the locations that the balls travel to are important. With the level of detail available to a model – identification of a specific hole during a specific round – 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 observation, this is not much of an issue since the effect 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 small. 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 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.

This unintended consequence is better explained using the following example: 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 the 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 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.

This consequence must be 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. This is the strategy taken in this work. 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 the course, hole (in other years) and round - as possible with computational burden.

This strategy naturally leads to the choice of grouped cross-validation for tuning the parameters of the Gradient Boosting Algorithm. This process was handled in an automated fashion utilizing a Bayesian Optimization library.[[3]](#footnote-1) For each category of shot, the optimal set of features was chose using a greedy approach.

1. https://github.com/adamwlev/Rank\_a\_Golfer [↑](#footnote-ref--1)
2. The XGBoost library was chosen because of the ease with which it handles large datasets. [↑](#footnote-ref-0)
3. https://github.com/fmfn/BayesianOptimization [↑](#footnote-ref-1)