Journal of Quantitative Analysis in Sports YYYY; aop

Research Article

McKay Gerratt*, Doeke Buursma, and Brian Hartman

Equitable Handicap System with Application in Disc Golf

https://doi.org/10.1515/sample-YYYY-XXXX
Received Month DD, YYYY; revised Month DD, YYYY; accepted Month DD, YYYY

Abstract: Handicaps are designed to allow players of different skill levels to compete on a level playing field. The current handicap system in golf does not equalize the chance of winning between players of different skill level, being biased towards more consistent players in small fields and towards less consistent players in large fields. In this paper, we develop a fair handicap system which equalizes each player's probability of winning. We use a simulation study to illustrate the need for a new system and how our system can overcome the issues in the current system. We then apply our system to disc golf using the richest score database available. We develop score predictions using a neural network and then use those score predictions to estimate a fair handicap for each player.

Keywords: Handicap, Equal Probability, Neural Network, Disc Golf, Golf

1 Introduction

A handicap is designed to enable athletes of differing skill levels to compete on an even playing field. The United States Golf Association (USGA) states its handicap system's purpose is "...to enhance the enjoyment of the game of golf and to give as many golfers as possible the opportunity to compete, or play a casual round, with anyone else on a fair and equal basis." (USGA, 2022). There are forms of handicapping in practically every sport. Many sports, such as tennis and pickleball, have divisions in an attempt to let players compete against similarly-skilled athletes.

The problem with many handicap systems is that they lack probabilistic backing. They do make the competitions more fair, but fall short of allowing each competitor an equal probability of winning. We will use golf as an example to show that the present handicap system may not actually give an equal probability to each competitor.

A golfer's handicap index (HI), used to calculate the number of strokes removed from the total needed to complete a course in a given round, is a function of the best 10 of their previous 20 rounds, as follows

$$HI = 0.96 * \frac{1}{10} * \sum_{i=1}^{10} D_{(i)}$$

Here $D_{(1)}$, ... $D_{(10)}$ are the smallest differentials of the golfer's previous 20 rounds, calculated as

$$D_i = (X_i - R_i) * \frac{113}{S_i} \quad \text{for } i \in [1, 2, ..., 20]$$

where i is the i^{th} previous round played. X_i is the golfer's adjusted score for round i, R_i is the course rating for the course where round i was played, and S_i is the slope rating for the course where round i was played (Lackritz, 2011).

***Corresponding author: McKay Gerratt,** Brigham Young University, Department of Statistics, Provo, UT, USA. e-mail: mckayger@gmail.com
Doeke Buursma, UDisc, LLC, North Branch, MN, USA, e-mail: doeke@udisc.com
Brian Hartman, Brigham Young University, Department of Statistics, Provo, UT, USA. e-mail: hartman@stat.byu.edu

The golfer's adjusted score differs from the gross score in that there is a maximum number of strokes per hole as determined by the golfer's handicap. For example, a player with a handicap of 25 has a maximum number of strokes per hole of 8 (Lackritz, 2011). The course rating can be viewed as the expected number of strokes that an expert golfer would take to complete 18 holes on the course (which should be at or near the total par for the course). The slope rating ranges from 55 to 155 and is meant to reflect the difficulty of the course to non-expert players, with 113 being the standard course difficulty (Kupper et al., 2001).

Given their handicap index, a golfer may then calculate their handicap for a particular course as

$$H = HI * \frac{S}{113}$$

where S is the slope rating for the course.

The system is a fairly flexible one. It accounts for how well the golfer performs relative to an expert player, and also adjusts a player's handicap for the course difficulty. In addition, the system only considers recent rounds, which may be more indicative of the golfer's true skill. Finally, the handicap index of the golfer may be used to measure themselves against others.

One thing to notice about the system is that it solely depends on the player; it does not give weight to the number or consistency of the other golfers in the field, which introduces potential inequalities in the chance of winning. Because the system gives weight to only the top half of the previous rounds played, it assumes that the players will play well, thus the handicap may not accurately reflect the skill of the player and cannot take into account the skill and consistency of the others in the field. There are many papers analyzing the fairness and effectiveness of handicapping systems, especially in golf. Even in the 1970s, researchers examined the fairness of the USGA system in two-player games and determined that the system provides the lower handicapped player with a competitve advantage (Scheid, 1971, 1972, 1975, 1977, 1979). McHale (2010) recently agreed with that conclusion. Several other papers mention that even if the handicaps are the same, a more consistent golfer has an advantage over a less consistent one (Smith and Prockow, 1981; Kupper et al., 2001). Many other papers have examined the statistical properties of the handicap index (Hall and Swartz, 1981a,b; Mosteller and Youtz, 1992; Scheid, 1990; Bingham and Swartz, 2000).

Swartz (2009) develops a handicap system for the Royal Canadian Golf Association which accounts for the statistical properties of handicaps. He incorporates the impact of playing against many players and presents a model with improved fairness in both one on one matches and in large tournaments. His handicap system is easily interpreted and accounts for the difference in fields. We agree (and will further show with simulation studies) that to create an equitable handicap for a competitor in any individual sport you need to account for the skill levels and consistency of those in the field. This does make the handicaps less intuitive, but the increased complexity comes with the benefit of a truly fair handicap.

In this paper, we extend Swartz (2009) with a model which provides more fairness and precision. We also have two advantages because of our particular application. First, with the proliferation of smartphones and handicapping apps, the calculations used to compute a handicap can be much more complicated than they could be in 2009. We are able to give up some interpretability for increased accuracy. Second, because our data is in disc golf, there is no currently established handicap system. Not having an established handicap and it's inertia allow us to deviate significantly from the system in golf.

In the next section, we illustrate the impact of field size and skill with a simple simulation study. We describe the game of disc golf in section 1.4. We then describe our model for predicting scores in section 2. In section 3 we describe our handicapping model. We conclude in section 4.

1.1 Simulation Study - Probability of Winning

We used normal distributions to model the scores of players in different scenarios. This could be thought of as time for a runner, points for a tennis player, or strokes for a golfer. We assume that, like in golf, the lowest score wins. In these simulations, all of the expected scores are the same (because if they weren't, a

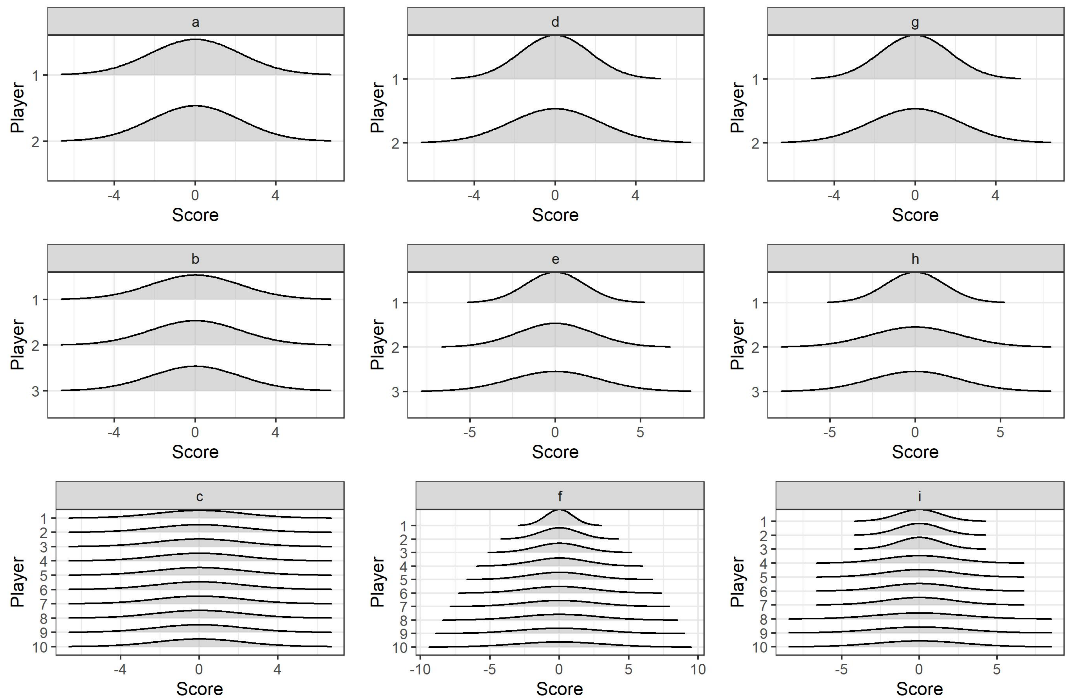


Fig. 1: Scenarios Without Shift

simple shift would make them all equal). For each of the 9 different scenarios, we varied the number of players (2, 3, and 10) and the variance pattern (equal variance, even spread of variances, and groups of variance). Each player's score distribution is centered at zero.

Figure 1 depicts the 9 scenarios while Table 1 depicts the probability of each player winning (or having the smallest value), as well as upper and lower Monte Carlo error bounds. In both the figure and table, the number of players is held constant across rows while the variance scenario is held constant across the columns. There are a few striking patterns in the probabilities.

First, there are 2 situations where each player has an equal probability of winning, without shifting the mean of the distribution. If there are two competitors, each has the same probability of winning, whether or not the variance is the same (scenarios a, d, and g). Regardless of number of players, if everyone competing has the same distribution, then each has an equal probability of winning (scenarios a, b, and c). We can also see that in the grouped scenarios, each player within a group has the same probability of winning.

Second, in scenarios e, f, h, and i, the player with the lowest variance has the smallest probability of winning, and the probability of winning increases up through the player with the highest variance. This clearly shows that even if players have a handicap that gives each the same expected score, the probability of each winning is not equal.

1.2 Simulation Study - Shift Needed to Equalize Probability

In every scenario presented in the previous section, all players had the same expected score, but had unequal probabilities of winning. We now want to calculate what adjustment to the center of the distribution is needed, relative to the player with the highest variance, in order for each player to have an equal probability of winning. For each scenario depicted in Figure 1, we calculated this adjustment by minimizing the difference in probabilities of the players most and least likely to win. The results are depicted in Figure 2 and the shift needed for equal probability is displayed in Table 2, as well as upper and lower Monte Carlo error bounds.

		Equal Variance				Even Variance				Grouped Variance			
Num	Play	Var	MC Lower	Est	MC Upper	Var	MC Lower	Est	MC Upper	Var	MC Lower	Est	MC Upper
2	1	5	0.497	0.500	0.503	3	0.497	0.500	0.503	3	0.497	0.500	0.503
2	2	5	0.497	0.500	0.503	5	0.497	0.500	0.503	5	0.497	0.500	0.503
3	1	5	0.330	0.333	0.336	3	0.301	0.304	0.307	3	0.295	0.298	0.301
3	2	5	0.331	0.334	0.336	5	0.332	0.335	0.338	7	0.348	0.351	0.354
3	3	5	0.330	0.333	0.336	7	0.357	0.360	0.363	7	0.348	0.351	0.354
10	1	5	0.098	0.100	0.102	1	0.016	0.017	0.018	2	0.039	0.041	0.042
10	2	5	0.098	0.100	0.102	2	0.037	0.038	0.039	2	0.039	0.041	0.042
10	3	5	0.098	0.100	0.102	3	0.057	0.059	0.060	2	0.040	0.041	0.042
10	4	5	0.098	0.100	0.102	4	0.077	0.078	0.080	5	0.103	0.104	0.106
10	5	5	0.098	0.100	0.102	5	0.095	0.097	0.099	5	0.103	0.105	0.106
10	6	5	0.098	0.100	0.102	6	0.112	0.114	0.116	5	0.103	0.105	0.107
10	7	5	0.098	0.100	0.102	7	0.127	0.129	0.131	5	0.102	0.104	0.106
10	8	5	0.098	0.100	0.102	8	0.141	0.143	0.145	8	0.151	0.153	0.156
10	9	5	0.098	0.100	0.102	9	0.154	0.156	0.159	8	0.151	0.153	0.155
10	10	5	0.098	0.100	0.102	10	0.166	0.169	0.171	8	0.151	0.153	0.156

Tab. 1: Probability of winning by scenario

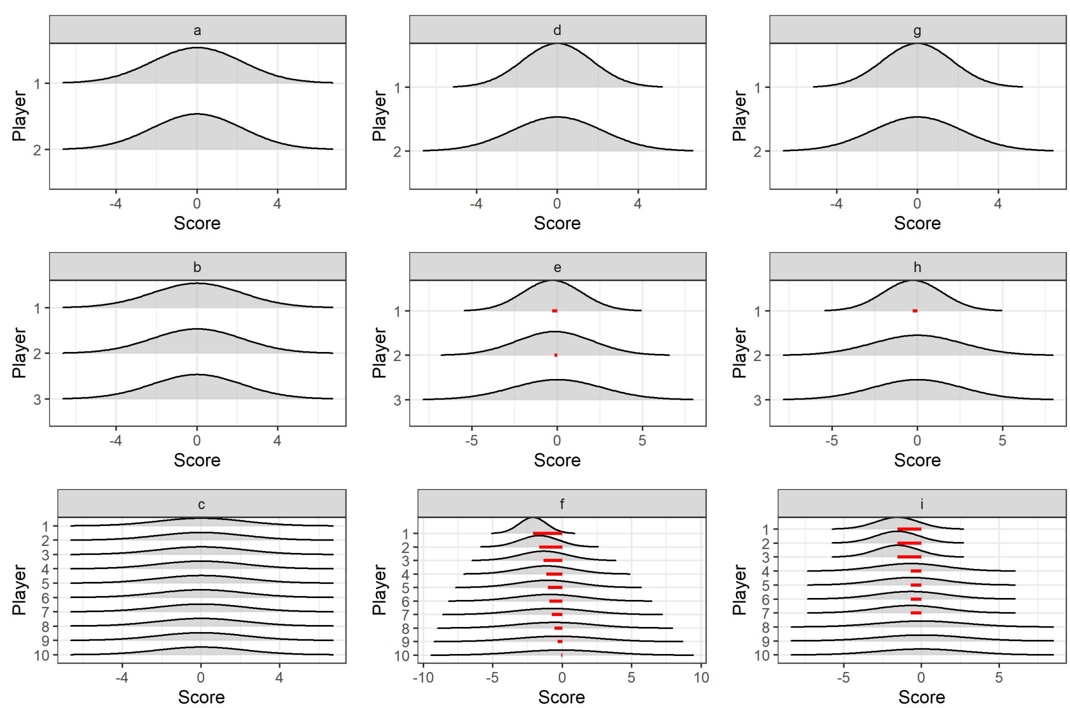


Fig. 2: Scenarios With Shift

		Equal Variance				Even Variance				Grouped Variance			
Num	Play	Var	MC Lower	Est	MC Upper	Var	MC Lower	Est	MC Upper	Var	MC Lower	Est	MC Upper
2	1	5	0.000	0.000	0.000	3	-0.013	0.000	0.013	3	-0.014	-0.001	0.012
2	2	5	0.000	0.000	0.000	5	-0.013	0.000	0.013	5	-0.013	0.000	0.013
3	1	5	0.000	0.000	0.000	3	-0.320	-0.294	-0.267	3	-0.292	-0.269	-0.246
3	2	5	0.000	0.000	0.000	5	-0.182	-0.154	-0.125	7	0.000	0.000	0.000
3	3	5	0.000	0.000	0.000	7	0.000	0.000	0.000	7	0.000	0.000	0.000
10	1	5	0.000	0.000	0.000	1	-2.320	-2.041	-1.761	2	-1.582	-1.556	-1.530
10	2	5	0.000	0.000	0.000	2	-1.911	-1.598	-1.286	2	-1.582	-1.556	-1.530
10	3	5	0.000	0.000	0.000	3	-1.592	-1.280	-0.968	2	-1.582	-1.556	-1.530
10	4	5	0.000	0.000	0.000	4	-1.374	-1.092	-0.811	5	-0.715	-0.687	-0.659
10	5	5	0.000	0.000	0.000	5	-1.229	-0.969	-0.710	5	-0.715	-0.687	-0.659
10	6	5	0.000	0.000	0.000	6	-1.083	-0.849	-0.614	5	-0.715	-0.687	-0.659
10	7	5	0.000	0.000	0.000	7	-0.901	-0.681	-0.461	5	-0.715	-0.687	-0.659
10	8	5	0.000	0.000	0.000	8	-0.698	-0.496	-0.295	8	0.000	0.000	0.000
10	9	5	0.000	0.000	0.000	9	-0.455	-0.265	-0.076	8	0.000	0.000	0.000
10	10	5	0.000	0.000	0.000	10	-0.093	0.000	0.093	8	0.000	0.000	0.000

Tab. 2: Shift needed for equal probability of winning

		Equal Variance				Even Variance				Grouped Variance			
Num	Play	Var	MC Lower	Est	MC Upper	Var	MC Lower	Est	MC Upper	Var	MC Lower	Est	MC Upper
2	1	5	0.499	0.500	0.501	3	0.696	0.697	0.697	3	0.696	0.697	0.697
2	2	5	0.499	0.500	0.501	5	0.303	0.303	0.304	5	0.303	0.303	0.304
3	1	5	0.332	0.333	0.334	3	0.606	0.607	0.608	3	0.702	0.703	0.704
3	2	5	0.333	0.333	0.334	5	0.263	0.264	0.265	7	0.147	0.148	0.149
3	3	5	0.332	0.333	0.334	7	0.128	0.128	0.129	7	0.148	0.149	0.149
10	1	5	0.099	0.100	0.100	1	0.388	0.389	0.390	2	0.223	0.224	0.225
10	2	5	0.099	0.100	0.100	2	0.207	0.208	0.209	2	0.223	0.223	0.224
10	3	5	0.100	0.100	0.101	3	0.131	0.132	0.132	2	0.223	0.224	0.225
10	4	5	0.100	0.100	0.101	4	0.087	0.088	0.088	5	0.065	0.065	0.066
10	5	5	0.100	0.100	0.101	5	0.061	0.062	0.062	5	0.065	0.065	0.066
10	6	5	0.100	0.100	0.101	6	0.043	0.044	0.044	5	0.065	0.065	0.066
10	7	5	0.099	0.100	0.100	7	0.030	0.030	0.030	5	0.064	0.065	0.065
10	8	5	0.099	0.100	0.101	8	0.021	0.022	0.022	8	0.023	0.023	0.023
10	9	5	0.099	0.100	0.100	9	0.015	0.015	0.016	8	0.022	0.023	0.023
10	10	5	0.099	0.100	0.100	10	0.011	0.011	0.011	8	0.022	0.023	0.023

Tab. 3: Probability of winning using USGA handicap

Unsurprisingly, the scenarios where either only two players are competing or all players have the same variance required no shift. In the scenarios with more than two players with differing variance, we can see that the players with the smallest variances require the biggest shifts in order to have the same probability. It is also interesting that players require different shifts depending on the number and variance of the other players in each scenario. Take, for example, the players with a variance of 5 in scenarios e, f, and i. The shift in each scenario differs greatly, ranging from a change of 0.154 to 0.969.

1.3 Simulation Study - USGA Handicap Simulation Study

For comparison, we thought it would also be insightful to show how the USGA handicap performs in evening the playing field. For this simulation, we first simulated the handicap for different skill levels. We assumed that the differentials of each player follow a normal distribution centered at 0. Using 10,000 iterations, we calculated the average handicap for each consistency level. We then found the probability of each player winning in the same scenarios as above. Table 3 shows how more consistent players now have an advantage over less consistent players. Simulations in Kupper et al. (2001) also support this conclusion.

The simulation study provides intuition behind why the number of players and the relative consistency in the playing field is important to factor into a handicap system. The same player will need a different handicap depending on the number and consistency of the players he is competing against, even if they have the same expected score.

In this paper we propose a simple structure on which handicap systems for many individual sports can be built. The structure consists of two parts:

- Modeling Individual Ability - No handicap system can be accurate without an accurate measure of a competitor's ability and consistency.
- Modeling Relative Ability - After individual ability is measured, each player in the field needs to be compared and given a suitable handicap given the ability and consistency of those in the group

Using this structure, we have developed a disc golf handicap system. The system has a similar purpose to that in golf, enabling players of differing skill levels to have a (near) equal probability of winning by adding a predetermined number of strokes to the final score of each player. Our proposed system has a few key advantages over the one currently used in golf; it not only accounts for the difficulty of the course being played on, but it also accounts for the skill, consistency, and size of the field of players.

1.4 Disc Golf

The reader will likely be somewhat familiar with the sport of golf, where a person aims to complete a series of holes by striking a ball with a golf club into each hole in as few strokes as possible. The sport of disc golf is very similar, where a disc is thrown in place of hitting a ball, and a basket is used in place of a hole in the ground.

Since disc golf was invented in the 1960's (PDGA, 2018), its popularity has slowly — but consistently — increased. When the global COVID-19 pandemic hit in 2020, the popularity of the sport grew dramatically and it shows no signs of returning to pre-COVID levels. In 2020 alone, there were an estimated 50 million rounds of disc golf played around the world (Hartman et al., 2021). In 2021, over 17 million rounds of disc golf were logged on UDisc (the leading disc golf score tracking app), which was an increase of about 50% over 2020 (UDisc, 2022). As the popularity of disc golf continues to increase, so does the relevance and need for a handicap system.

Using the structure stated above, we first predict the ability of each disc golfer using a neural network. We then use the output in an optimization process to estimate a handicap that gives each competitor an equal probability of winning on a given course.

2 Predicting Scores

The handicapping system for golf can be thought of—to an extent—as a predictive model: given a player's handicap index HI and a course's rating R and slope rating S, the player's expected score on the course is given by

$$E = R + HI * \frac{S}{113}.$$

Our proposed handicapping system will also depend on a model which predicts a player's score on a given course.

In the golf handicapping system, course and slope ratings are determined by course raters—humans who evaluate both quantitative (e.g., hole length, elevation change, prevailing wind) and qualitative (e.g., shape of the fairway, bunker placement) aspects of a course (USGA, 2022). Raters assign different course and slope ratings for each set of tees on a course.

There are roughly 40,000 golf courses in the world, each with several sets of tees (a common configuration is to have forward, middle, and back tees) (R&A, 2021). Therefore, roughly 200,000 course and slope ratings need to be manually evaluated and maintained. This is a tall task, but not insurmountable given golf's resources and time-tested infrastructure.

On the other hand, it is not uncommon for each basket on a disc golf course to occupy one of two or more positions in the ground at any given time. The choice of position can drastically alter the difficulty of the hole. A single 18 hole disc golf course with two basket positions per hole has $2^{18} = 262{,}144$ configurations. Courses with three basket positions per hole each have $3^{18} = 387{,}420{,}489$ configurations. There are roughly 15,000 courses in the world and there is no central body nor consistent criteria by which course difficulty is evaluated.

These practices and circumstances necessitate that a predictive scoring model for disc golf be hole-based (as opposed to course-based), and predictions can then be aggregated to the course level as needed. The task is to estimate the discrete distribution $p(\text{score} \mid \text{player, hole})$.

2.1 Evaluation Criteria

We evaluate our model and some naive baselines against several accuracy metrics. Since the task can be thought of as both a classification task (yield the score class among the choices 1, 2, 3, . . .) as well as a regression task (yield the expected score as a floating point value), we utilize metrics which apply to each perspective.

- **Top k Accuracy** (classification metric): For each sample, was the actual score one of the k most probable scores according to the model? This metric takes the value 1 if true, 0 if false, and averages across all samples. We evaluate this metric for $k = 1, 2, 3$. Higher is better.
- **Mean Absolute Error** (regression metric): For each sample, what was the absolute difference between the expected score according to the model and the actual score? This metric averages these absolute differences across all samples. Lower is better.
- **Root Mean Squared Error** (regression metric): For each sample what was the squared difference between the expected score according to the model and the actual score? This metric averages these squared differences across all samples, then takes the square root to return to the original units. Lower is better.

We use 5-fold cross-validation. To generate the folds, we use the following procedure:

1. Randomize the order of the data.
2. Apply a stable sort by player, then by hole.
3. Cyclicly assign fold indices (1, 2, 3, 4, 5, 1, 2, 3, 4, 5, . . .).

This procedure ensures that players are evenly distributed between folds, and that holes are roughly evenly distributed between folds.

2.2 Naive Baselines

We compare our model's performance against the following naive baselines.

- **Global Frequency (GF)**: Use the global scoring frequencies in the training set as the prediction for every sample in the validation set. That is, every sample in the validation set receives the same predictions.
- **Player Frequency (PF)**: For each player, use their overall score frequencies in the training set as the prediction for all of that player's samples in the validation set. That is, each player receives a constant prediction regardless of the hole. Samples from players which appear in the validation set but not the training set are handled according to GF.
- **Hole Frequency (HF)**: For each hole, use its overall score frequencies in the training set as the prediction for all of that hole's samples in the validation set. That is, each hole receives a constant

prediction regardless of the player. Samples from holes which appear in the validation set but not the training set are handled according to GF.

- **Player and Binned Distance Frequency, (PBDF)**: Group holes into 4 equally sized bins by distance. For each (player, distance bin) pair, use that pair's overall score frequencies in the training set as the prediction for all of that pair's samples in the validation set. Pairs which appear in the validation set but not the training set are handled first according to PF (if the player appears in the training set), then by GF.

2.3 The Model

The data set we used consists of 34,438,664 hole plays by $n_P = 244,534$ players on $n_H = 156,827$ holes and 10,044 courses throughout the world. The vast majority of disc golf courses exist in and throughout the United States, but there are courses in at least 79 countries with high concentrations in Canada and Northern Europe, especially Norway, Sweden, and Finland (UDisc, 2022).

In this data set, scores of 9 are about as common as holes-in-one, and scores of $n = 10$ or more are about half as common, occurring in 0.03% of samples. We use this fact to justify the simplification of considering scores greater than 10 as 10s. The same simplification is applied to the naive baselines.

Our model is a neural network which can be summarized by the following diagram:

$$\begin{array}{ccccc} H \xrightarrow{\iota_H} \mathbb{R}^h & & & & \\ & \searrow^{\mathrm{cat}_H} & & & \\ & & \mathbb{R}^{h+p} \xrightarrow{f} \mathbb{R}^n \xrightarrow{\sigma} \Delta^{n-1} & & \\ & \nearrow_{\mathrm{cat}_P} & & & \\ P \xrightarrow{\iota_P} \mathbb{R}^p & & & & \end{array}$$

where

- H and P are the sets of holes and players, respectively,
- ι_H and ι_P are categorical embeddings into real vector spaces of dimensions h and p, respectively (h and p are hyperparameters of the model),
- $\mathrm{cat} = \mathrm{cat}_H \times \mathrm{cat}_P$ is concatenation of vectors,
- f is any parametrized differentiable transformation—we use an affine linear transformation, i.e., a dense linear layer with bias,
- $\Delta^{n-1} = \{(y_1, y_2, \ldots, y_n) \mid 0 < y_i < 1,\ y_1 + \cdots + y_n = 1\} \subseteq \mathbb{R}^n$ is the probability simplex, and
- $\sigma : (x_1, x_2, \ldots, x_n) \mapsto (\sum \exp(x_i))^{-1}(\exp(x_1), \exp(x_2), \ldots, \exp(x_n))$ is the softmax function.

The model hyperparameters are h, p, n, and the choice of f. We have discussed our choice of $n = 10$ and f as an affine linear transformation above, and we use $h = p = 3$.

The categorical embeddings ι_H and ι_P can carry both fixed and trainable parameters. For example, our data set contains hole distances so we fix the first axis of ι_H to be (log-standardized) hole distances.[1] If m_H and m_P are the number of fixed hole and player features, then ι_H and ι_P carry $n_H(h - m_H)$ and $n_P(h - m_P)$ trainable parameters, respectively. If f is an affine linear transformation, then it carries $n(h + p + 1)$ trainable parameters.

As the model is differentiable, it can be trained against the crossentropy loss function using any variation of gradient descent. For the below results, we used adaptive moment estimation (Kingma and Ba, 2014). Note that although the number of trainable parameters carried by ι_H and ι_P is large, they yield sparse gradients.

1 Other features of holes and players could in principle be included here if known, e.g. elevation difference between tee and target, or player age.

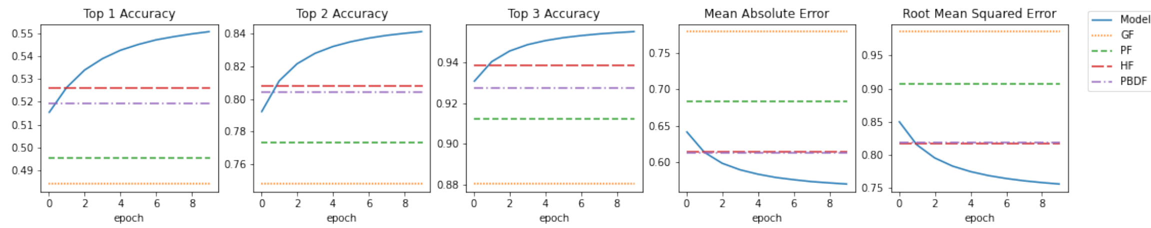
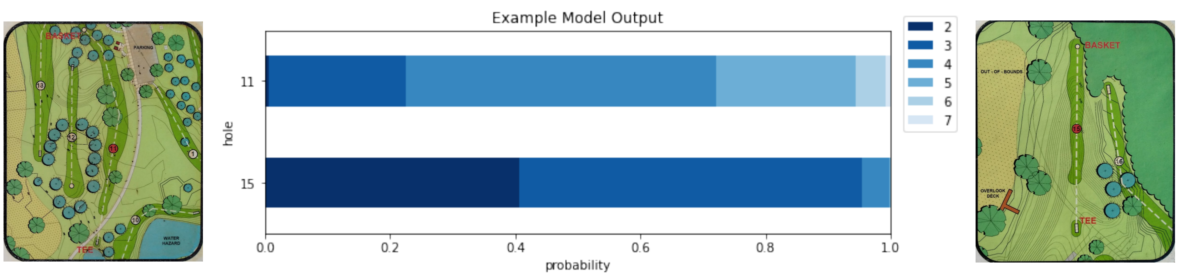


Fig. 3: Model evaluation over the course of 10 epochs on the validation sets, average of 5 folds, against naive baselines.



(a) Hole 11

(b) Model predictions on holes 11 and 15

(c) Hole 15

Fig. 4: Sample model predictions. The darkest shade indicates the probability of scoring a 2 on the hole. Holes-in-one and scores greater than 7 have negligible probability.

One improvement we hope to make is to introduce covariates which are not directly tied to players or holes, the most prominent example being weather conditions (wind speed and direction, temperature, humidity, precipitation, etc.). This requires only minor changes to the model architecture, although the tuning of hyperparameters becomes more complicated.

If c is the dimension of the covariate space, then we propose the modified model represented by the following diagram.

$$\begin{array}{ccccccc} H \xrightarrow{\iota_H} \mathbb{R}^h & & & & & & \\ & \searrow^{\mathrm{cat}_H} & & & & & \\ \mathbb{R}^c & \xrightarrow{\mathrm{cat}_C} & \mathbb{R}^{h+c+p} & \xrightarrow{f} & \mathbb{R}^n & \xrightarrow{\sigma} & \Delta^{n-1} \\ & \nearrow_{\mathrm{cat}_P} & & & & & \\ P \xrightarrow{\iota_P} \mathbb{R}^p & & & & & & \end{array}$$

Note that the hyperparameters h and c would need to be re-tuned, and the function f would likely need to be more sophisticated than an affine linear transformation to properly encapsulate potentially non-monotonic relationships between covariates and scoring outcomes (e.g. temperature).

2.4 Results and Examples

Metric results are in figure 3. We see that our model outperforms the naive baselines within the first few epochs.

As a first example, figure 4 shows model predictions for the 2nd author on holes 11 and 15 at Paw Paw Park Disc Golf Course in Holland, MI. Hole 11 is 164m (538ft) with trees throughout. The hole design elicits higher scores and more scoring variance when compared to hole 15 which is only 77m (253ft) with no obstacles. The output distributions of the model have expected scores of 4.1 and 2.6 on the two holes, respectively, and standard deviations of 0.86 and 0.58.

Figure 5 shows model predictions for four players on each hole at Johnson Park Disc Golf Course in Grand Rapids, MI. Player 1 is the 2nd author, Player 2 is the 2nd author's brother-in-law (who, anecdotally,

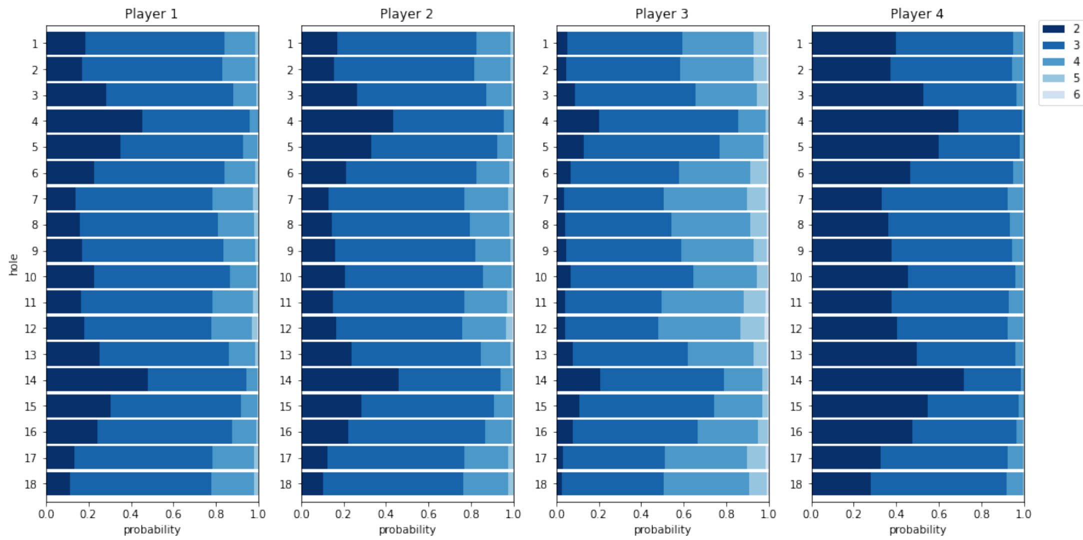


Fig. 5: Model predictions for four players of varying skill levels on each hole at Johnson Park Disc Golf Course in Grand Rapids, MI. The darkest shade indicates the probability of scoring a 2 on the hole. Holes-in-one and scores greater than 6 have negligible probability.

is of similar skill to the 2nd author), Player 3 is the 2nd author's sister, and Player 4 is a professional disc golfer. The sums of the expected scores for the entire round for the four players are, respectively, 52.7, 53.3, 61.0, and 46.7, and the standard deviations are 2.71, 2.73, 3.15, and 2.45.

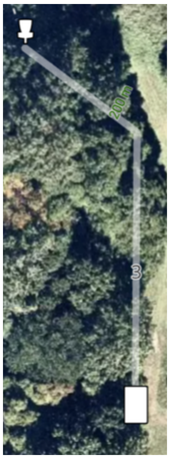
To illustrate the predictive power of the model, note that during the time the data was collected, Player 1 lived in the Grand Rapids, MI area near the course in question, while Players 2 and 3 lived in Australia. The data set does not contain any plays by Players 2, 3, or 4 at Johnson Park Disc Golf Course.

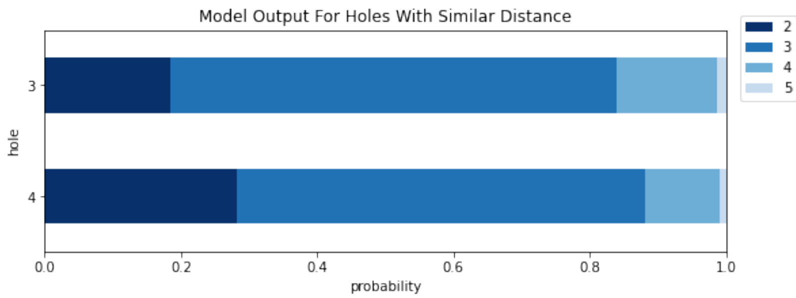
To illustrate the fact that the model is capable of learning hole characteristics other than distance, we observe Figure 6 which shows the model predictions for the 2nd author on two holes with similar distance but different characteristics. The first (hole 3) is 62.5m (205ft), up a steep hill, with a sharp "dogleg left" fairway shape. The second (hole 4) is 68m (223ft), has nearly zero elevation difference, and no obstacles. The expected scores on the two holes are 3.0 and 2.85, respectively, with standard deviations 0.63 and 0.65.

2.5 Toward a Handicapping System

The trained model works reasonably well. But there is a step missing between a predictive model and a handicapping system. Namely, a handicapping system requires online learning. Player skill can improve or deteriorate. Obstacles such as trees can be added, removed, or modified, either by design or by nature. A handicapping system should be kept up-to-date with these changes.

In our model, the categorical embeddings ι_H and ι_P quantify the characteristics of players and holes, while f describes the relationship between quantified player and hole characteristics and scoring outcomes. We can make the assumption that once the parameters carried by f have been determined, they do not change over time. This corresponds to freezing the weights present in f. Then, we can apply online learning using classical stochastic gradient descent to only the parameters carried by ι_H and ι_P. This has the additional qualitative advantage that training the model on new data only affects the players and holes which appear in the new data.





(b) Model predictions on holes 3 and 4

(a) Hole 3

(c) Hole 4

Fig. 6: Model predictions for the 2nd author on holes 2 and 3 at Johnson Park Disc Golf Course in Grand Rapids, MI. The darkest shade indicates the probability of scoring a 2 on the hole. Holes-in-one and scores greater than 5 have negligible probability.

3 Handicap System

We will now use the predicted scores to build the handicap system. The data used to develop the system come from predictions from the neural network on 10,000 players and the 50 most played courses in 3 states (Colorado, Florida, and Utah) and Finland (where disc golf is also popular). This was done to obtain a wide variety of course difficulties, terrains, and layouts.

We converted the by-hole score probabilities for each player to by-course score probabilities, so that we had the probabilities of total scores for each player on each course, and could calculate the expected score and variance for each player.

We began developing the model for each player's handicap as a function of the field's expected scores and variances.

$$H_{ij} = E_{ij} - P^{(1)} * V_{ij}$$

Here H_{ij} is the handicap, or number of strokes, added to the score of Player i (the player with the i^{th} smallest variance) playing at course j, E_{ij} is the difference in expected score of Player i and the player in the field with the most variability at course j, and V_{ij} is the difference in variance of Player i and the Player n, or the player in the field with the most variability at course j. We expect $P^{(1)}$ to change depending on the field size and course difficulty.

In order to determine the relationship between $P^{(1)}$ and the number of players in the field, we ran several scenarios of field size N from 3 up to 150 players. For each iteration, we drew N players from the population and sampled 10,000 scores for each, adding some random jitter to make it a continuous distribution. We then found the values of $P^{(1)}$ in this equation that minimized the difference in probability of winning between all players in the field. This was done repeatedly for each scenario of N players. Figure 7 shows the results from this exploration, organized by location, with the number of players plotted against the average estimated value of $P^{(1)}$.

There is a clear pattern displayed here. Each line displays the results for a particular course at the location. For each course, as the number of players in the field increases, the average value of $P^{(1)}$ increases as well, flattening out at about $\frac{1}{2}$. By optimizing on the value of a second parameter $P^{(2)}$, we found that the following equation fits each curve nicely.

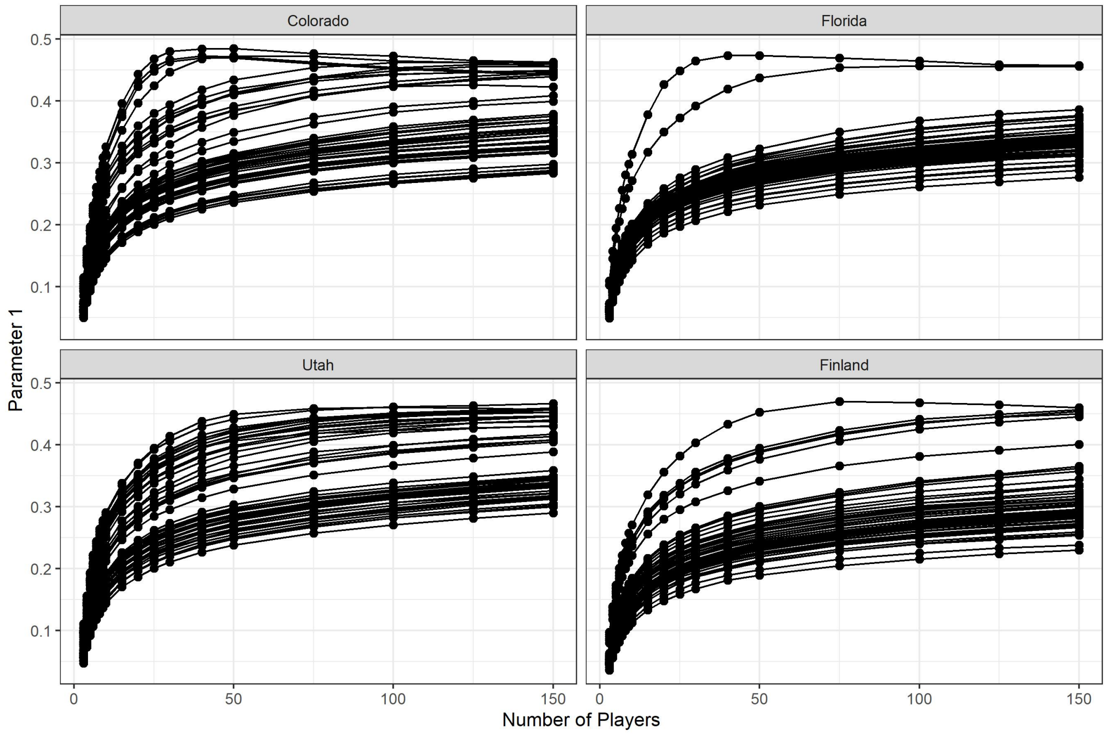


Fig. 7: Model Layer 2 Results

$$P^{(1)} = A\left(1 - \left(\frac{2}{N}\right)^{P^{(2)}}\right)$$

Here N is the number of players in the field. A is the horizontal asymptote, under which the curve is constrained. 2 in $\frac{2}{N}$ was calculated to constrain $P^{(1)}$ to be zero when the number of players is 2. $P^{(2)}$ is an additional parameter to be estimated that makes an adjustment for each individual course.

We found that a horizontal asymptote of $\frac{1}{2}$ fit the data well. Using that asymptote in the equation, we found the relationship between $P^{(2)}$ and the average course variance (calculated by averaging all of the predicted variances for all 10,000 players on the particular course) as depicted in Figure 8.

We decided on the following equation

$$P^{(2)} = \frac{B}{C_j^D}$$

where C_j is the average variance in the score at course j, and B and D are parameters to be estimated. This line is shown in Figure 8

We can see that for higher average course variances that our model appears to underestimate $P^{(2)}$. This is not a very serious issue as this had the best fit of all the functional forms we considered in terms of sample mean squared error. In addition, we show in the next section that the model performs very well, even on those courses with a high variance.

By minimizing the sum of squared error, we estimate B to be approximately 1.3738 and D to be approximately 0.7330.

3.1 Proposed Model

Our proposed model has the following form.

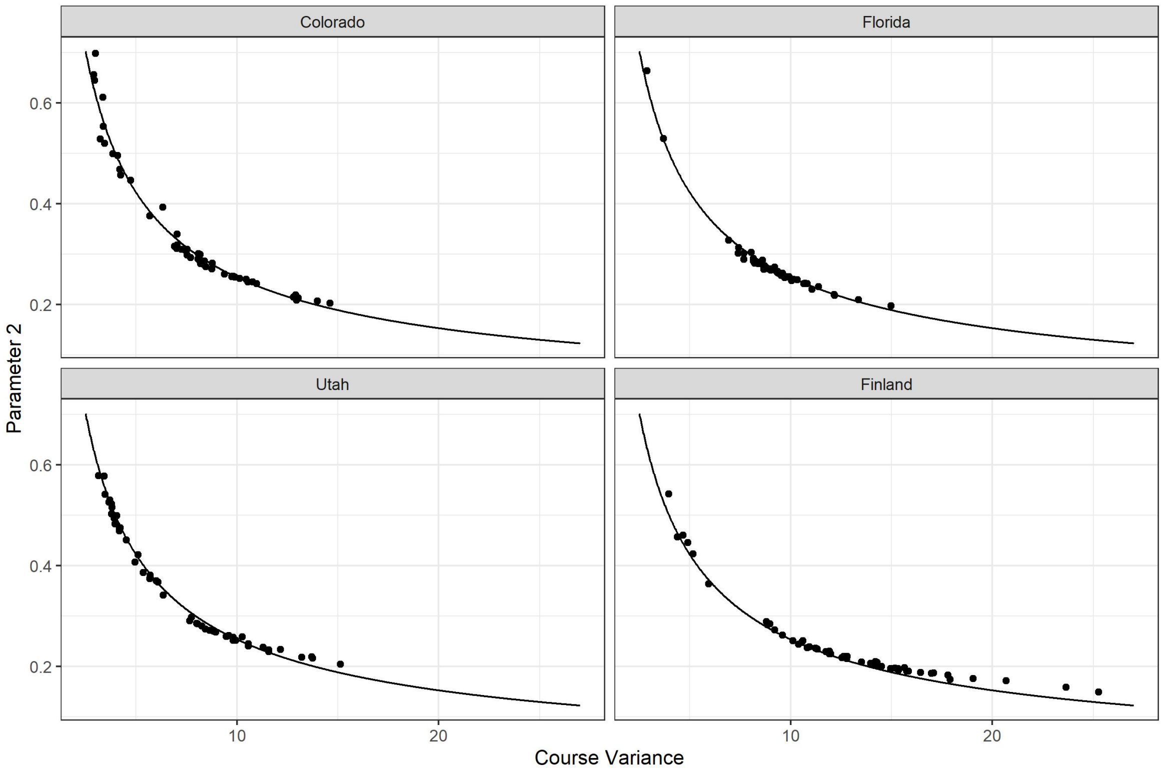


Fig. 8: Model Layer 3 Results

$$\begin{aligned}\hat{H}_{ij} &= \lfloor E_{ij} - \hat{P}^{(1)} * V_{ij} \rceil \\ \hat{P}^{(1)} &= \frac{1}{2}\left(1 - \left(\frac{2}{N}\right)^{\hat{P}^{(2)}}\right) \\ \hat{P}^{(2)} &= \frac{1.3738}{C_j^{0.7330}} \end{aligned} \tag{1}$$

In the next section we will evaluate how well this model performs in creating an equal playing field.

3.2 Model Evaluation

In order to test how well our model performs, we repeatedly drew from the score distributions of N players (again, ranging from 3 to 150) from all 200 courses used in developing the model, calculated our estimated handicap (rounded to apply in a discrete case) and estimated the probability of each player winning. Figure 9 shows the average difference in probability from the player least likely to win to the player most likely to win after the estimated handicap is applied.

There are a few striking patterns. Overall our model adequately creates an even playing field for all competitors, as the maximum average disparity in probability is between 0.015 and 0.025 for each location, with the majority of scenarios being less than 0.01. It is interesting that the model performs the worst for a medium number of players, between 8 and 20, in each location. This may be because the problem is trivial for 2 players, and that increasing the number of players increases the probability of large disparities in variance. Then as the number of players increases, the probability of each player winning naturally decreases, and this seems to outweigh the effect of having more players to account for. It is also interesting to note

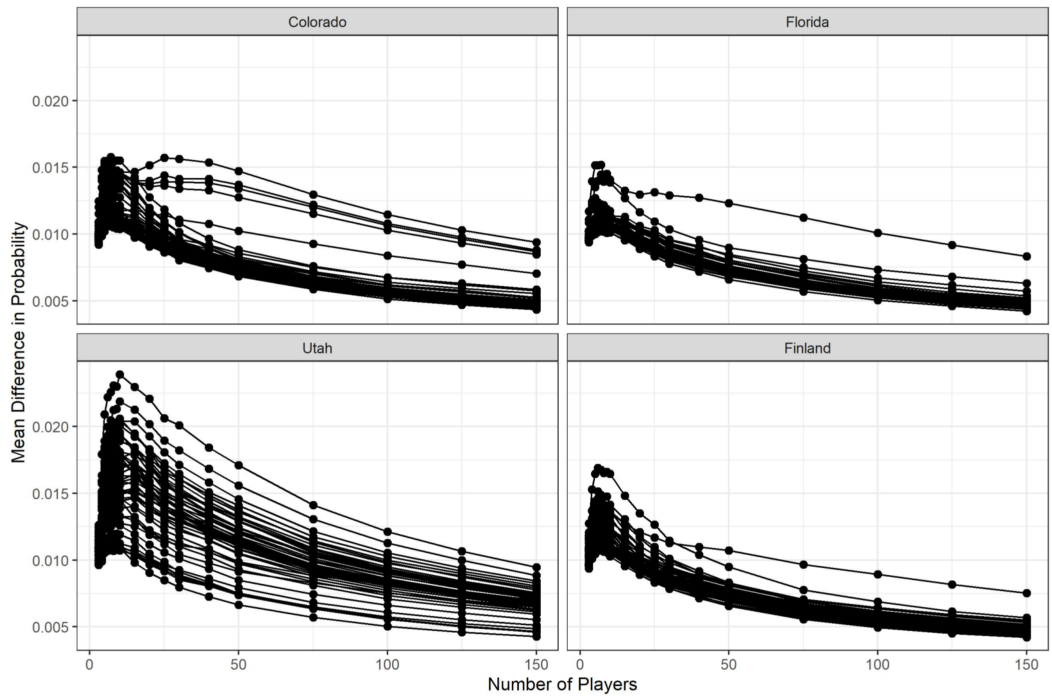


Fig. 9: Model Results

that the fitted model performs consistently worse in Utah as compared to the other locations. It may be that the courses in Utah are more varied than those in the other locations.

As a point of interest, we tested a naive model, where we assumed the handicap had the following form

$$\hat{H_{ij}} = \lfloor E_{ij} \rceil$$

where $\lfloor E_{ij} \rceil$ is the difference in expected scores for player i and the player in the group with the highest variance on course j, rounded to the nearest integer. Figure 10 displays the results.

When comparing Figures 9 and 10, it is clear that accounting for the number and variance of players is very important, as our fitted model does about 5 times better than the naive model, in terms of the average largest disparity in probability. We see the same pattern as before, that the model performs the worst for mid-sized groups. We also see that the model performs the worst in Utah.

In developing the model, we did not partition our players into train/test sets. However, the randomly selected fields of players that we used in developing our model were different from those that we used to test it. While a player's predictions may have been used in both developing and evaluating the model, their relative contribution is different depending on the other players in the field.

We did not test our fitted model on any courses outside of Colorado, Florida, Utah, and Finland, but we expect that our model will apply well to any location. We mentioned before that we chose these locations in order to get a wide variety of layouts, elevations, and terrains. We can see in Figure 8, the course variance has a very strong relationship with the handicap needed, with little variation around our estimate. As long as we have the average variance for the course, our model will give an accurate handicap for a group of players at that location. Moving forward, we will continue to test our model to ensure it applies to other locations.

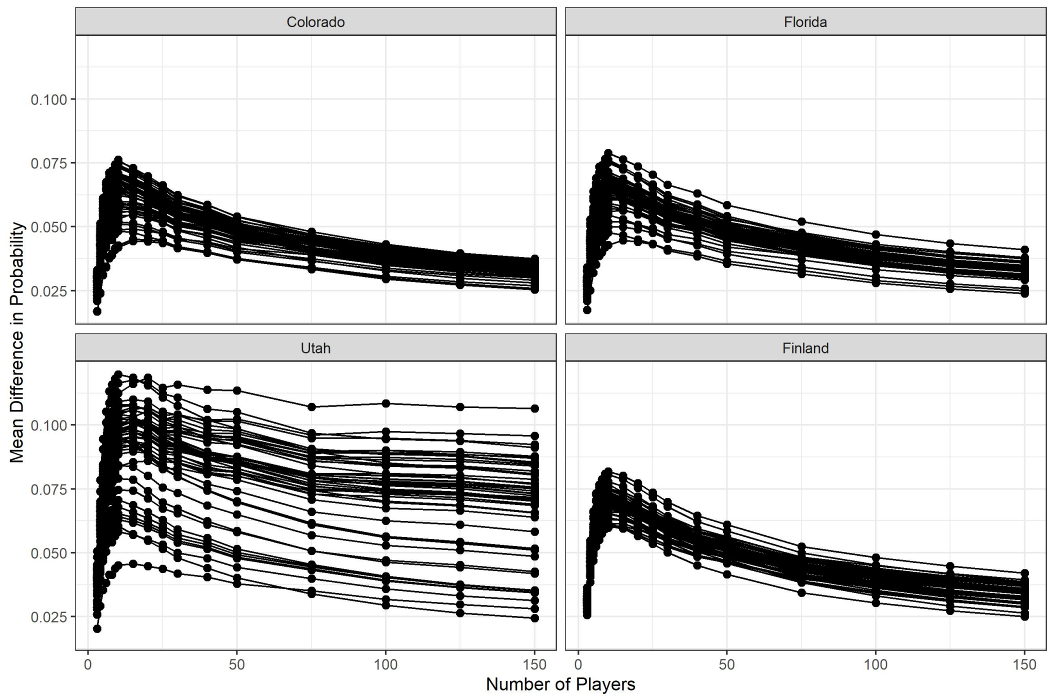


Fig. 10: Naive Model Results

4 Conclusion

It is essential to account for the size, skill, and consistency of the playing field in order to create a truly fair handicap system. Our model produces a handicap that gives each competitor a near equal probability of winning, fulfilling that purpose of a handicap system. Our model performs much better than other methods that do not account for player consistency and field size.

We have shown the model's success in an application to disc golf, but this handicap system is applicable to many sports. A neural network can be used to accurately predict the ability and consistency of each competitor; these predictions can then be used to give each player an equal probability of winning. We believe that this method is effective in creating a truly even playing field.

Acknowledgment: The authors appreciate Josh Lichti, Matt Krueger, and the rest of team UDisc for their commitment to innovation in disc golf and for building the technology necessary to collect the training data upon which our system is built. We thank Anthony Rentsch for many helpful discussions and valuable feedback. We also thank David Dahl and Gil Fellingham for their help with this manuscript.

References

Bingham, D. R. and T. B. Swartz (2000). Equitable handicapping in golf. *The American Statistician 54*(3), 170–177.

Hall, C. and C. Swartz (1981a). The effect of handicap stroke location on best-ball golf scores. *Mathematical Modelling 2*(3), 161–167.

Hall, C. and C. Swartz (1981b). The effect of handicap stroke location on golf matches. *Mathematical Modelling 2*(3), 153–159.

Hartman, B., B. Brunner, M. Christensen, and M. Gerratt (2021, Feb). Figuring out how much disc golf was played in 2020.

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kupper, L. L., L. B. Hearne, S. L. Martin, and J. M. Griffin (2001). Is the usga golf handicap system equitable? *Chance 14*(1), 30–35.

Lackritz, J. R. (2011). A new analysis of the golf handicap system: Does the better player have an advantage? *CHANCE 24*(2), 24–34.

McHale, I. G. (2010). Assessing the fairness of the golf handicapping system in the uk. *Journal of sports sciences 28*(10), 1033–1041.

Mosteller, F. and C. Youtz (1992). Professional golf scores are poisson on the final tournament days. In *1992 Proceedings of the Section on Statistics in Sports*, pp. 39–51. American Statistical Association Washington, DC.

PDGA (2018). Brief history of disc golf and the pdga.

R&A (2021). Golf around the world.

Scheid, F. (1971). You're not getting enough strokes. *Golf Digest*, 32–33.

Scheid, F. (1972). A least squares family of cubic curves with an application to golf handicapping. *SIAM Journal on Applied Mathematics 22*(1), 77–83.

Scheid, F. (1975). Nonlinear feature of golf course rating with application to handicapping. In *Siam Review*, Volume 17, pp. 388–389. SIAM PUBLICATIONS 3600 UNIV CITY SCIENCE CENTER, PHILADELPHIA, PA 19104-2688.

Scheid, F. (1977). An evaluation of the handicap system of the united states golf association. *Optimal strategies in sports*, 151–155.

Scheid, F. (1979). Golf competition between individuals. Technical report, Institute of Electrical and Electronics Engineers (IEEE).

Scheid, F. (1990). On the normality and independence of golf scores, with applications. In *Science and Golf: Proceedings of the First World Scientific Congress of Golf (editor AJ Cochran), London: E. and FN Spon*, pp. 147–152.

Smith, L. A. and I. Prockow (1981). Simulating a match play handicapping system. *Simulation & Games 12*(4), 417–450.

Swartz, T. B. (2009). A new handicapping system for golf. *Journal of Quantitative Analysis in Sports 5*(2).

UDisc (2022, Feb). The disc golf boom, one year removed.

USGA (2022). Rules of handicapping.