Club Recommender for David Horsey

Matt Mills

Introduction

Club choice can be one of the biggest decisions a golfer makes on the course. Many inputs affect this decision such as pin location, conditions, distance to the pin, and the lie of the ball. In this paper I present a method to find the optimal club choice for David Horsey by using models for estimating the value of a shot and the impact of club type, lie, and distance to the pin. These were used to find the expected strokes gained from picking one club type over another given the lie and distance to the pin. The dataset was limited to the type of club (driver, fw wood, hybrid, iron, putter, wedge), the distance from the pin, and the type of lie of the ball from 14 tournaments. The methods for finding the value of each club should be generalizable to more information like club number, x/y position on the course, playing conditions, and other information. First a model is built to estimate the expected number of remaining strokes on a hole given Mr. Horsey's distance to the pin and current lie. Then the value of each shot is measured by comparing the expected number of strokes before and after the shot, with an expected strokes remaining of 0 once the ball is in the hole. Then a model is built that attempts to predict the expected shot value given the club used, current lie, and distance to the pin. This model can then be used to find the expected strokes gained from picking a certain club for a shot.

A Model for Estimating Strokes Remaining

The perspective of this model considers the quality of a player's position before a shot using this formula.

$$shot\_value = f(pin\_distance|current\_lie)$$

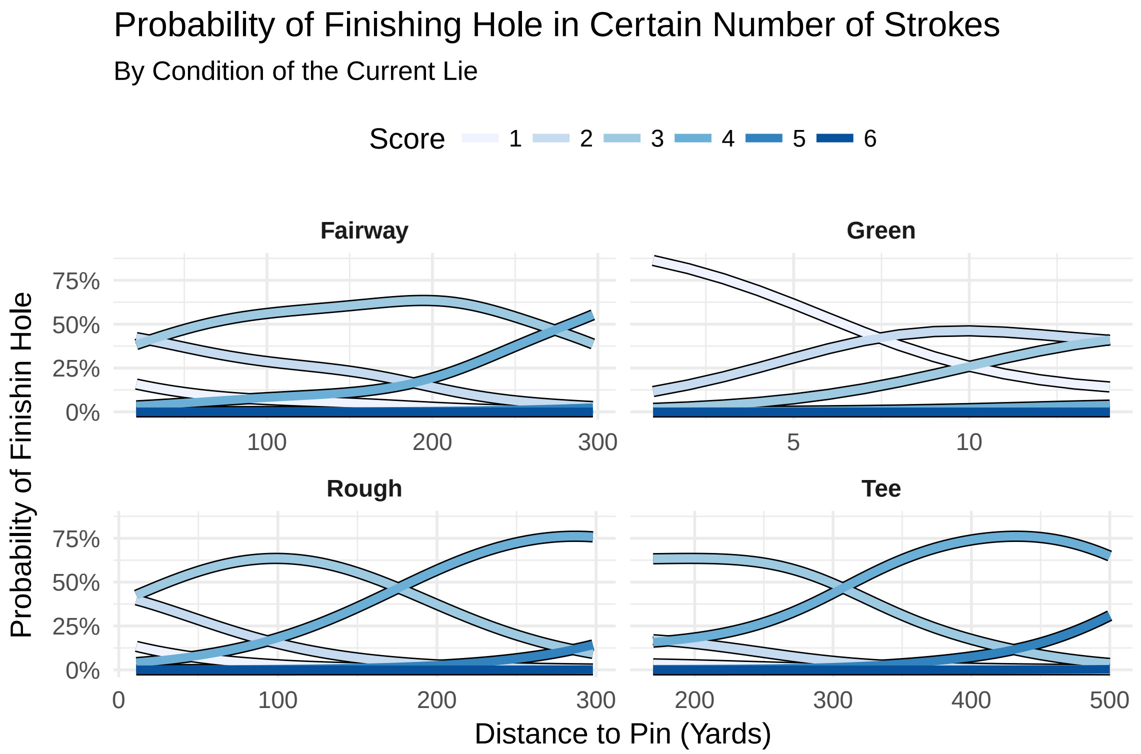
This model was estimated using an ordinal generalized additive model using seperate smoothing splines for each type of lie. In R this was fit using the `mgcv` package with the following call:

```r
shots <- shots %>%
  mutate(shots_remaining = score - shot_no + 1)
shots$lie_before_factor <- as.factor(shots$lie_before)
## GAM needs the by variable to be a factor

cats <- sort(unique(shots$shots_remaining))
ordered_model <- gam(shots_remaining ~ s(left_to_pin_before, by = lie_before_factor),
                     data = shots,
                     family = ocat(theta = cats[c(-1, -8)]))
## gam uses -1 as the first cut level and the last category doesn't need a cut
## level since it's the last one
```

This model was chosen because of the insight that can be gained by having seperate models for completing the hole in a certain number of shots. However a Poisson model will generate very similar predictions (correlation of .98).

An example of the model output can be seen below:

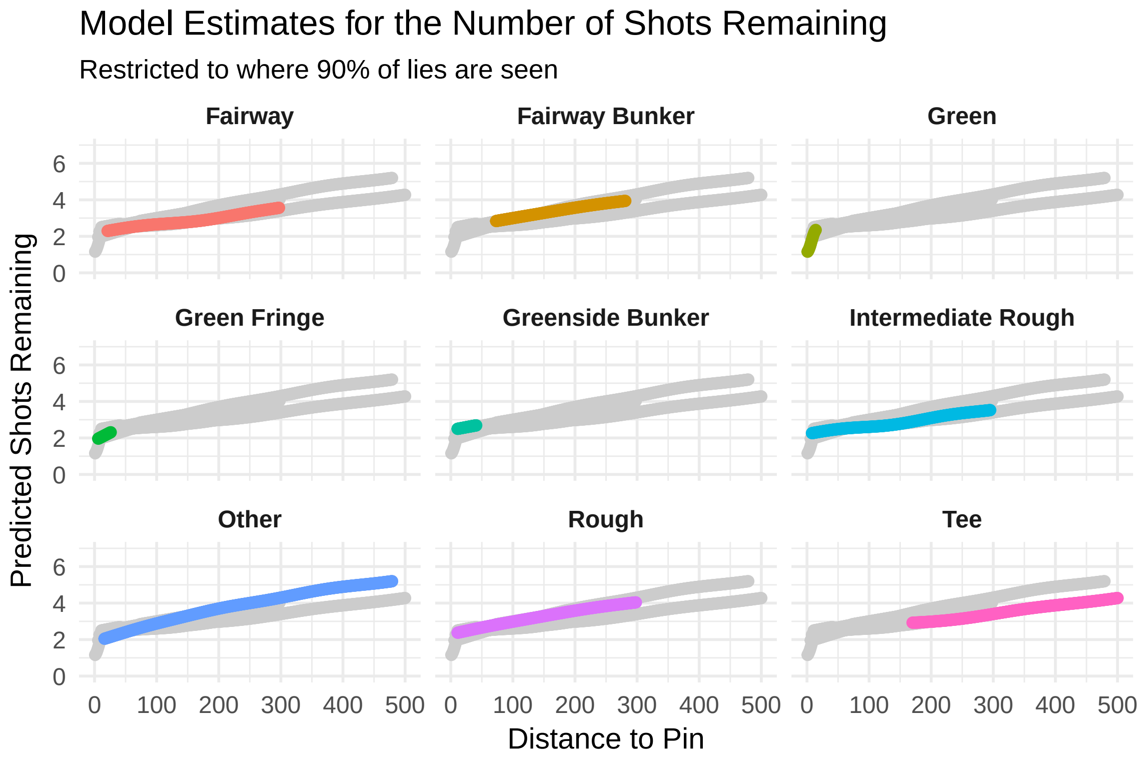


For example, this model estimates a putt from 5 yards out on the Green has around a 62% chance of being made, while a shot from the Fairway has ~8% chance of going in.

The probability distributions for the underlying remaining strokes on the hole are used to get an estimated number of shots remaining. Here is an example of the results of this model from the DP World Tour Championship in Dubai on the Par-5 18th hole in the 2nd round:

Shot Number	Distance to Pin (Yards)	Current Lie	Predicted Shots Remaining
1	540	Tee	4.47
2	259	Fairway	3.36
3	70	Fairway	2.56
4	2	Green	1.21

A full picture of the expected shots remaining for a givin lie and distance can be seen below.



It should be said that this expected shots model is actually a “Expected Shots Remaining given the courses David Horesy played in 2016 and the results of those shots”; It doesn’t have as nice of a ring to it but it is more accurate. With more shot data from more players and more tournaments the generalizability of this model should increase.

Strokes Gained During a Shot

Now that we have an estimate for the number of strokes remaining given a shot’s conditions we can find an estimate for the strokes gained once a shot has been completed.

$$strokes\_gained = expected\_shots\_remaining_{before} - expected\_shots\_remaining_{after}$$

Using the 18th hole of the 2nd round of the World Championships as our example again we can see how this strokes gained score is calculated:

Shot Number	Distance to Pin	Current Lie	Shots Remaining Before	Shots Remaining After	Strokes Gained
1	540	Tee	4.47	3.36	1.11
2	259	Fairway	3.36	2.56	0.81
3	70	Fairway	2.56	1.21	1.34
4	2	Green	1.21	0.00	1.21

A Model for Estimating the Strokes Gained for a Shot

Now that we have an estimate for the strokes gained during a shot we can attempt to measure the value of club choice. This is done by using the following model:

$$strokes\_gained = f(club\_choice|distance) + f(current\_lie|distance)$$

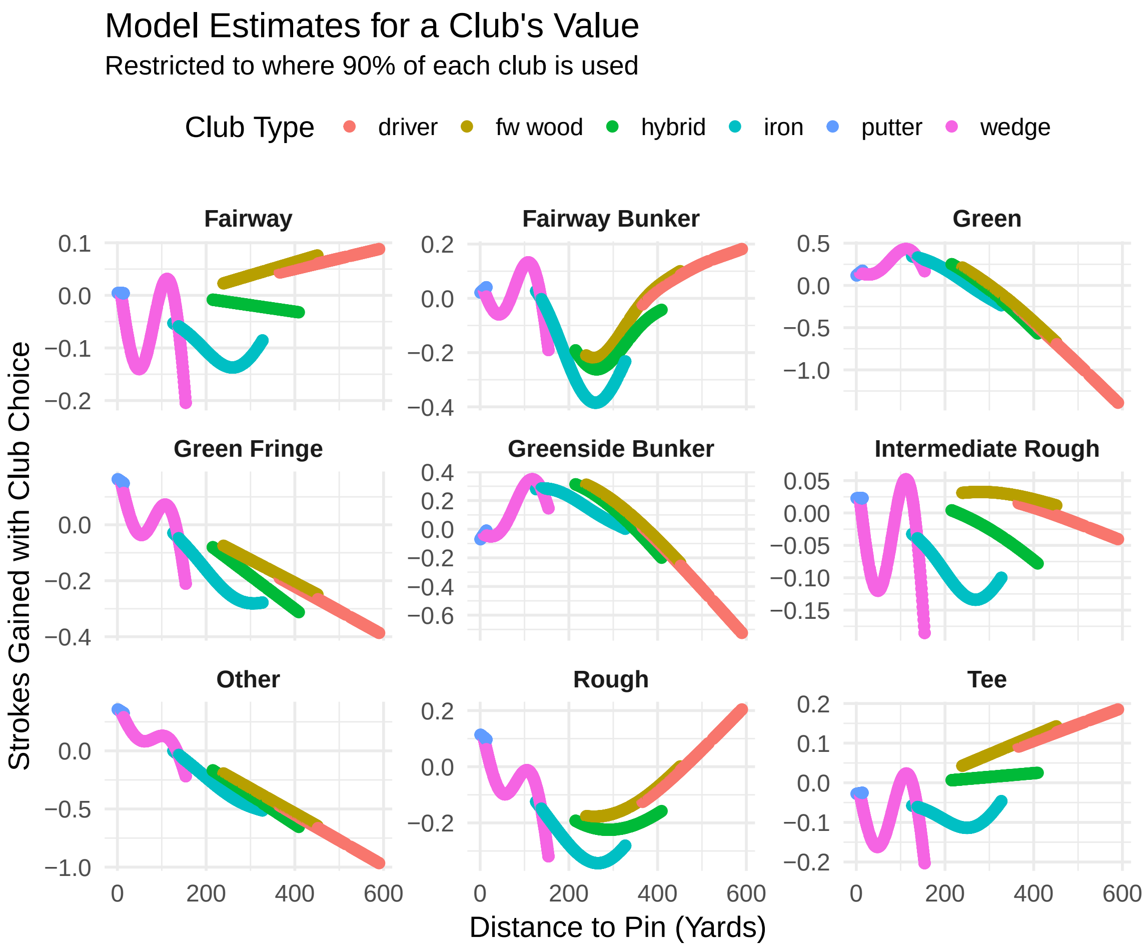
The rationale for this model is that certain club types are better at different distances and different lies. So given a lie and distance we should be able to optimize our club choice to give us the club that has been shown to increase the strokes gained for similar shots. It is important to note that we want to maximize the strokes gained on a shot; a good shot will decrease our expected shots remaining by "gaining" more strokes than expected on that shot.

This model is fit using another generalized additive model with different smoothing splines estimated for each club type and each lie as shown below. With this formulation we can derive an estimate for impact of club choice on the strokes gained during a shot given that shot's distance and current lie.

```
shot_value$club_factor <- factor(shot_value$club)

club_gam <- gam(shot_value ~ s(left_to_pin_before, by = club_factor) + s(left_to_pin_before, by = lie_be
                data = mutate(shot_value, lie_before_factor = factor(lie_before)))
```

Now that our model is built we can visualize the impact of club choice by different lies and distances:



My hypothesis coming in was that certain clubs would be better for some lies than others but that does not seem to be the case. According to this model the Fairway Wood always improved the strokes gained more than any other club no matter the lie, at least where the Fairway Wood was commonly used. With more data we could hopefully tell if this results is an artifact of the sample we were given, due to a biased club choice (fairway wood only being used on the best lies of each individual lie type), or if it actually exists.

Conclussions

There are a couple caveats with this analysis. The club choice model is dependent on different club types being used at similar distances so we have data to estimate the value of picking one club over another. In addition ideally we would be able to fit the Expected Strokes Remaining and Strokes Gained models on different datasets. However with only 14 tournaments worth of data for one player I decided to show the overall proof of concept using all data for both models.

However even with all the caveats I think that there is ample evidence that data-driven analysis could improve Mr. Horsey's understanding of his own game and hopefully even influence his decision making on the course during actual tournaments.