Analysis of David Horsey's Approach Shots in 2016

Scott Came

March 20, 2017

> *Believe it or not, approach shots... affect the scoring average of better players more than any other category.*
>
> –Golf Analytics Expert Mark Broadie, *Golf Digest*, 2015

This paper, written for the March, 2017 Hackathon sponsored by 15th Club, analyzes shot-level data from fourteen of European Tour player David Horsey's rounds in 2016. The analysis attempts to quantify Horsey's performance on approach shots, considering the context in which each shot was taken. Its most important finding is that shot distance, lie (Fairway versus Fairway Bunker versus Rough), and putting surface area are significant explanatory factors in predicting an approach shot's probability of landing on the green. Ambient weather (wind, barometric pressure, temperature) and the presence of greenside hazards do not seem significant in explaining whether shots land on the green, though this finding could well be due to the relatively small sample of data available for analysis.

Data Overview

The analysis leveraged a dataset provided by 15th Club, consisting of 3412 shot-level observations from 49 of David Horsey's European Tour rounds in 2016. This period was one of overall strong performance from Horsey, with six top-twelve finishes and two top-fives, including a T2 at the Turkish Airlines Open:

Tournament	Date	Score	Finish
BMW International Open	June 23	-9	T7
100th Open de France	June 30	+12	T139
AAM Scottish Open	July 7	+4	T92
D+D Real Czech Masters	August 18	-2	T47
Made In Denmark	August 25	+10	T138
KLM Open	September 8	-12	T4
Italian Open	September 15	-16	T5
Porsche European Open	September 22	-6	T49
Alfred Dunhill Links Championship	October 6	-12	T11
British Masters supported by Sky Sports	October 13	-10	T12
Portugal Masters	October 20	-15	T22
Turkish Airlines Open	November 3	-17	T2
Nedbank Golf Challenge hosted by Gary Player	November 10	+7	T44
DP World Tour Championship, Dubai	November 17	-11	T13

The shot-level dataset includes the following variables for each shot (variables irrelevant to our analysis are excluded):

- Tournament name
- Round Number
- Starting Tee for the round (e.g., if starting on the back nine, Starting Tee = 10)
- Hole Number
- Par for the hole (3, 4, or 5)
- Distance to the pin before the shot
- Distance to the pin after the shot
- Lie before the shot (e.g., Tee, Fairway, Fairway Bunker, Green, etc.)

The first step in the analysis is to take a subset of the shot-level dataset that includes only approach shots. The PGA Tour defines[1] "approach shot" (for statistics purposes) as a shot that does not originate from on or around the green, and ends up on or around the green or in the hole. The Tour further defines "on or around the green" as being within 30 yards of the edge of the green. Because the 15th Club dataset does not indicate distance from the edge of the green, but rather distance from the pin, the subsetting process will use a 35-yard distance from the hole in determining "on or around the green". Applying this threshold results in a dataset of 873 approach shots taken in the sampled 14 tournaments.

Course Feature Data

An important component of this analysis involves examination of the impact of course features on approach shot performance. The `golfcoursegeo` package[2] produces, for each hole:

- Area of the green (putting surface)
- Linearity of the green (i.e., degree of "oblongness")
- Number of greenside water hazards and bunkers
- Approximate elevation change from shot location (tee or fairway) to green

The `golfcoursegeo` package, created by the author, converts Google Earth (KML) map layers that adhere to the package convention into hole-by-hole datasets of golf course characteristics. The author created layers for a sample of the courses involved in the Horsey dataset; 463 of the sample approach shots have associated course/hole feature information. Except where otherwise indicated, all analysis in this paper has been performed on this subset of Horsey's shots, which includes every shot taken at the following tournaments:

- 100th Open de France
- BMW International Open
- British Masters supported by Sky Sports
- DP World Tour Championship, Dubai
- Nedbank Golf Challenge hosted by Gary Player
- Portugal Masters
- Turkish Airlines Open

To compute linearity of the green, the analysis considers the ratio of the area of the minimum-radius enclosing circle to the area of the green. Thus greens that are relatively equal-diameter in their dimensions will have a ratio closer to one than greens that are long and narrow, so that a higher linearity measure indicates a more "unequally proportioned" green.

Weather Data

The analysis also explored the effects of weather on approach shot performance, using hourly Automated Surface Observation Station (ASOS) data obtained from an online database[3] made available at Iowa State University. We manually found the nearest ASOS station (typically a commercial airport) to each golf course in the analysis, then retrieved the hourly observation that was closest to noon local time on each day

[1] http://www.pgatour.com/stats/stat.02325.html
[2] https://github.com/scottcame/golfcoursegeo
[3] https://mesonet.agron.iastate.edu/request/download.phtml

that Horsey played a round on that course. See the link below for a list of all available fields; this analysis considered only air temperature, wind speed, and barometric pressure. Please note that the distance from the course to the ASOS station ranged from a few to several miles, and that weather conditions at the time Horsey played any particular shot likely varied somewhat from the daily observation we considered. This is quite possibly why weather conditions proved insignificant in the models estimated below; improving upon the inclusion of weather data is a potential topic of future research.

Summary Statistics and Visualizations

The golfer's objective, when standing over an approach shot, is to finish that hole in the fewest remaining shots. Generally speaking (though not always[4]), that means attempting to advance the ball such that it comes to rest a minimum distance from the hole. And almost always, the player has a dual goal of hitting the ball on the green, if at all possible given the circumstances of the shot. Thus when assessing a player's performance on approach shots, two outcome measures (or, in modeling terms, dependent variables) of interest are: proximity to the hole (a continuous variable, measured in feet), and whether the ball lies on the green at the end of the shot (a binary Yes/No variable). These two variables are named `DistanceToPinAfter` and `ShotResult`, respectively.

The analysis will consider the following as explanatory (or independent) variables for these outcomes:

- `DistanceToPinBefore`: Distance from the hole at the start of the shot (in feet)
- `GreenArea`: The area of the green (in square yards)
- `GreenLinearity`: The linearity of the green (i.e., it's "oblongness")
- `ApproachElevationChange`: The elevation difference between the green and the spot from which the approach shot is taken (in feet)
- `GreensideHazards`: The presence of greenside water hazards
- `GreensideBunkers`: The presence of greenside bunkers
- `ShotLie`: Lie for the shot (e.g., Tee, Fairway, Fairway Bunker, Rough)
- `Par`: Par for the hole

Preliminary efforts initially included weather data in the analysis, but found that none of the explanatory variables included (air temperature, wind speed, and barometric pressure) were significant in explaining the dependent variables. In the interest of brevity, these variables will not be discussed further here.

Basic Summary Statistics and Frequency Counts

The following table gives basic summary statistics for all continuous variables in the analysis:

Variable	Min	25th %ile	Median	75th %ile	Max	Mean
DistanceToPinAfter	1.00	12.00	24.00	42.00	105.00	29.15
DistanceToPinBefore	108.00	381.00	498.00	577.50	960.00	481.90
GreenArea	251.50	602.00	697.80	775.80	1110.30	689.10
GreenLinearity	1.23	1.62	1.80	2.09	3.26	1.91
ApproachElevationChange	-41.81	-7.37	0.29	6.71	38.01	-0.17
GreensideBunkers	0.00	1.00	2.00	3.00	5.00	1.93
GreensideHazards	0.00	0.00	0.00	1.00	4.00	0.41

The following tables give frequency counts for the discrete variables in the analysis:

[4]For example, if a pin is tucked on the narrow part of a green, with water on one side and a bunker on the other, and a player is on the back nine in the final round with a three-shot lead, he or she may elect to "play it safe" and aim for the center of the green, rather than aiming at the flag. Typically, though, proximity to the hole is the goal.

ShotResult	Frequency
Off Green	132
On Green	331

ShotLie	Frequency
Fairway	258
Fairway Bunker	8
Intermediate Rough	50
Rough	39
Tee	108

Par	Frequency
3	104
4	270
5	89

In summary: nothing really unexpected here. Horsey hits most of his approach shots from the fairway, though four times he went for the green on par four holes, as well (every tee shot on a par three is considered an approach shot). He is getting onto the green the vast majority of the time with his approach shots, which is also to be expected from a professional golfer.

In initial modeling of explanatory factors, it quickly became clear that the change in elevation and the number of greenside bunkers and hazards were not significant factors in explaining the outcomes. So again in the interest of simplifying the presentation of results, these factors will not be considered further here.

Relationships of Key Variables

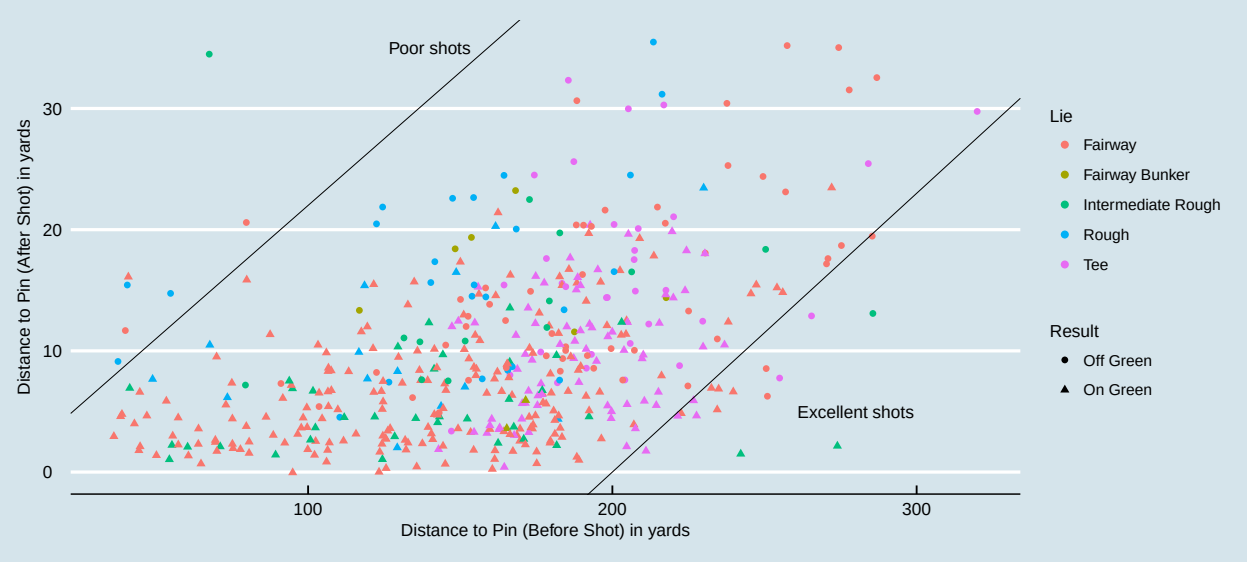
The regression models considered later will suggest significant relationships between the dependent variables (`DistanceToPinAfter` and `ShotResult`) and the length of the approach shot (`DistanceToPinBefore`) and the lie of the ball for the approach shot (`ShotLie`). But a sense of these relationships emerges from a simple table:

Distance	Lie	n	Mean Distance To Pin After	Mean # Remaining Shots	% On Green
250+	Fairway	13	66.77	2.31	23.08
250+	Intermediate Rough	3	33.00	2.33	33.33
250+	Tee	4	57.75	2.00	0.00
200-250	Fairway	28	39.61	2.18	57.14
200-250	Fairway Bunker	1	42.00	2.00	0.00
200-250	Intermediate Rough	3	30.67	1.67	66.67
200-250	Rough	4	86.25	2.75	25.00
200-250	Tee	37	37.19	1.97	64.86
150-200	Fairway	99	26.59	1.88	76.77
150-200	Fairway Bunker	5	38.40	1.80	40.00
150-200	Intermediate Rough	16	26.88	1.81	68.75
150-200	Rough	15	40.73	2.40	20.00
150-200	Tee	63	32.38	1.95	76.19
100-150	Fairway	70	17.94	1.71	91.43
100-150	Fairway Bunker	2	49.50	2.00	0.00

Distance	Lie	n	Mean Distance To Pin After	Mean # Remaining Shots	% On Green
100-150	Intermediate Rough	18	19.50	1.61	77.78
100-150	Rough	14	37.57	2.07	50.00
100-150	Tee	4	22.75	2.00	75.00
70-100	Fairway	25	15.72	1.52	92.00
70-100	Intermediate Rough	5	15.40	1.60	80.00
70-100	Rough	1	18.00	2.00	100.00
35-70	Fairway	23	13.13	1.57	95.65
35-70	Intermediate Rough	5	28.40	1.40	80.00
35-70	Rough	5	34.80	2.20	40.00

This table strongly suggests some fundamental relationships that are fairly intuitive to golf fans (and certainly to experienced players). First, the results of an approach shot (in terms of distance from the pin after the shot, and whether the ball rests on the green) are less favorable the longer the shot taken. And second, there is a significant penalty–in terms of hitting (or getting closer to) the green–as well as the expected number of subsequent shots on a hole–for being in the thick rough or a fairway bunker. This penalty seems to be between a third and half a stroke, depending on the shot distance.

A scatterplot provides a visual representation of the same relationships:



In this graphical depiction, the prominence of the light blue (rough) and the few dark green (fairway bunker) at the top of any given vertical slice of the graphic–and the prominence of circles rather than triangles for those observations–underscores that the rough and in fairway bunkers are not the ideal places from which to play an approach shot. It is also more immediately apparent in the graphic (versus the table) that Horsey had quite a few truly excellent shots–and also some that he would probably rather forget.

The next phase of our analysis attempts to model these relationships more rigorously, and will include course features in the model to determine the impacts of things like green area and greenside hazards.

Regression and Classification Models

Recall from the prior discussion that there are two approach shot outcomes of interest here: having the ball come to rest as close to the hole as possible, and having the ball come to rest on the green. Preliminary analysis suggests two factors that seem to influence these outcomes: the length of the approach shot, and

the lie of the ball for the approach shot. It would appear, from tabular and graphical analysis, that these explanatory variables are correlated with the outcomes; the next step is to estimate a least-squares regression for the `DistanceToPinAfter` dependent variable, followed by a logistic regression classification model for the `ShotResult` dependent variable.

The independent (explanatory) variables for these models will include:

- `DistanceToPinBefore`: Distance from the hole at the start of the shot (in feet)
- `GreenArea`: The area of the green (in square yards)
- `GreenLinearity`: The linearity of the green (i.e., it's "oblongness")
- `ShotLie`: Lie for the shot (e.g., Tee, Fairway, Fairway Bunker, Rough)

Initial regression runs showed that the counts of greenside hazards and bunkers were not significant explanatory factors for either of the outcome (dependent) variables. The par for the hole (i.e., whether the approach shot is a par-three tee shot versus a fairway shot on a par 4/5) turned out to be statistically insignificant as well. The analysis that follows will only present models fitted without these variables included.

Least Squares Regression: Explaining Variation in Proximity to the Hole

To analyze factors that explain the distance remaining to the hole after the approach shot, consider the following ordinary least squares regression:

$$DistanceToPinAfter = \beta_0 + \beta_1 DistanceToPinBefore + \boldsymbol{\beta_2} ShotLie + \beta_3 GreenArea + \beta_4 GreenLinearity + \epsilon$$

Note that the $\boldsymbol{\beta_2}$ represents a vector of coefficients on dummy variables representing the type of lie, with fairway lies held out as the reference value. Note also that the units of `DistanceToPinAfter` and `DistanceToPinBefore` are in feet, while `GreenArea` is in square yards.

Fitting this regression to the 463 approach shot observations results in the following coefficients, standard errors, and p-values:

Variable	Estimate	Standard Error	Pr(>|t|)
(Intercept)	-0.660	6.372	0.918
DistanceToPinBefore	0.061	0.006	0.000
ShotLie: Fairway Bunker	13.620	6.516	0.037
ShotLie: Intermediate Rough	-0.249	2.806	0.929
ShotLie: Rough	19.223	3.136	0.000
ShotLie: Tee	1.391	2.213	0.530
GreenArea	-0.005	0.006	0.380
GreenLinearity	0.924	2.158	0.669

An R^2 value for this regression of 0.27 indicates that this model explains 27 percent of the variation in `DistanceToPinAfter`. The F statistic value of 24 with 7 and 455 degrees of freedom leads to rejection of the null hypothesis that the model does not explain the variation in `DistanceToPinAfter`. In short: there is a great deal (73 percent, to be exact) of the variation in proximity to the hole that this model does not explain, but we can confidently reject the notion that it does not explain any of the variation.

As expected after the graphical presentation in the prior section, the explanatory variables `DistanceToPinBefore` (i.e., the length of the approach shot) and `ShotLie` (for fairway bunker and rough lies) are all statistically significant. The coefficient on `DistanceToPinBefore` suggests that for every 100 feet added to the length of an approach shot, one would expect the ball to land approximately six feet farther from the hole, controlling for lie and green size and shape. A six percent rate of growth in expected remaining distance makes sense, considering that golfers play longer shots with "longer" clubs to cover the additional distance, but that longer clubs come with less control and higher variance in where the ball comes to rest after the shot. It is not possible with the data available to test the hypothesis that some (or even most) of the additional six percent

distance remaining after the shot is due to lateral inaccuracy in the shot. This would be an interesting topic for further research and analysis.

The dummy variables for Fairway Bunker and Rough lies indicate that these lies add just over four and six yards, respectively, to the expected distance remaining after the shot, compared to shots of the same distance played from the fairway. There are several potential explanations for this finding, including:

- Thick rough often impacts the clubhead speed with which a golfer contacts the ball, resulting in less energy and less flight distance
- Grass between the club face and ball at impact affects spin rate, sometimes unevenly across the face, which affects shot control
- Fairway bunker lies sometimes compel a golfer to choose a higher-lofted club in order to clear the bunker lip, affecting shot distance
- Occasional "fried egg" lies in a fairway bunker can create impact issues similar to those created by thick rough, with the same consequences

The available data do not allow testing of these hypotheses, but doing so would be a valuable area of further research.

Interestingly, approach shots played from tees (whether from a par three tee, or from a par four where Horsey "went for it") and shots played from the Intermediate Rough are not statistically significantly different from Fairway shots. Finally, the model indicates that the size and shape (linearity) of the green are not statistically significant factors in explaining proximity to the hole on approach shots.

Logistic Regression: Explaining Variation in Success at Getting on the Green

Next the analysis will focus on explaining a different outcome: rather than analyzing proximity to the hole following an approach shot, we will look at whether Horsey succeeded at having the ball come to rest on the green. In many ways, this is a more valuable insight, since most golfers would rather have a 25-foot putt for birdie (having hit the green in regulation) than a bunker shot 20 feet from a tucked pin, needing to get up-and-down for par. The analysis will use logistic regression to *classify* each of Horsey's 463 approach shots as successful or not, in terms of the ball resting on the green after the shot.

Whereas least squares regression seeks to explain variation in a variable, like `DistanceToPinAfter`, that is continuous over a range of values, logistic regression explains variation in a discrete variable. A common scenario is to use logistic regression to explain a binary outcome, like `ShotResult`. The regression coefficients (indirectly) allow estimation of the impact of the dependent or explanatory variables on the odds of one outcome occurring versus the other.

This model fits the same explanatory variables as in the least squares regression model above, but now the dependent variable is the log-odds ratio of being on the green versus not:

$$\ln\left(\frac{Prob(OnGreen)}{1-Prob(OnGreen)}\right) = \beta_0+\beta_1 DistanceToPinBefore+\boldsymbol{\beta_2} ShotLie+\beta_3 GreenArea+\beta_4 GreenLinearity+\epsilon$$

Fitting this model to the sample of Horsey approach shots produces the following coefficient values (we have indicated, for each explanatory variable, the marginal effect of that variable on the odds of hitting the green):

Variable	Odds Ratio	Estimate	Standard Error	Pr(>|t|)
(Intercept)	25.353	3.233	0.900	0.000
DistanceToPinBefore	0.994	-0.006	0.001	0.000
ShotLie: Fairway Bunker	0.083	-2.494	0.848	0.003
ShotLie: Intermediate Rough	0.533	-0.629	0.386	0.104
ShotLie: Rough	0.095	-2.352	0.411	0.000

Variable	Odds Ratio	Estimate	Standard Error	Pr(>|t|)
ShotLie: Tee	0.991	-0.009	0.282	0.974
GreenArea	1.002	0.002	0.001	0.006
GreenLinearity	0.715	-0.335	0.282	0.235

As with the least squares regression analysis above, note that `DistanceToPinBefore` (i.e., the length of the shot), and the lie being in the Fairway Bunker or Rough are significant factors in explaining the odds of Horsey hitting the green with his approach shot. However, there is a new significant explanatory variable in this analysis: area of the green. This makes sense, because the area of the green has a direct impact upon whether a ball that is hit to within a specified distance of the hole is in fact on the green. This means, of course, that if Horsey and his caddie were primarily interested in hitting the green on a given hole–versus just trying to get it as close to the pin as possible–then this analysis enables them to factor in the size of the green in choosing among their tactical options.

It is easiest to interpret the effects of the factors by considering the marginal odds ratio effects. For example, holding all other factors constant, a one foot increase in the length of the approach shot (`DistanceToPinBefore`) reduces the odds of hitting the green by 0.6 percent. Similarly, though somewhat more alarmingly, hitting an approach shot from the rough reduces the odds of coming to rest on the green by 90 percent.

An example further illustrates how the odds calculation works. Consider Horsey standing in the fairway, 200 yards from the flag, contemplating a shot into a perfectly round, 500 square yard green[5]. The model indicates that his odds of hitting the green with that shot are[6]:

$$Odds = e^{3.233-0.006*(3*200)+0.002*500-0.335*1} = 1.34$$

Thus Horsey's odds of hitting the green on this relatively straightforward shot are just shy of seven-to-five. However, consider what happens if we move his ball a few feet behind him into the rough, and ask him to play the same distance shot from there instead. His odds now change to:

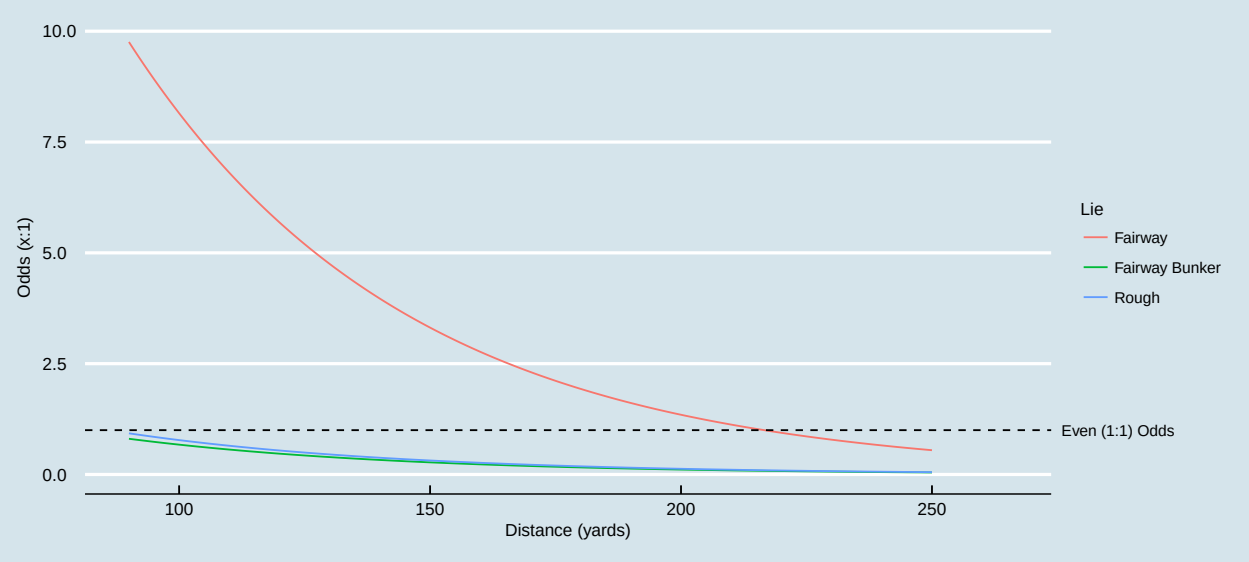
$$Odds = e^{3.233-0.006*(3*200)-2.352*1+0.002*500-0.335*1} = 0.128$$

The odds have shifted dramatically, and are now almost eight-to-one *against* him hitting the green. As every experienced golfer knows, there are usually significant consequences to playing an approach shot from thick rough or a fairway bunker!

The impact of approach shot length and bunker/rough lies is summarized in the following graphic:

[5]A 500 square yard, perfectly round green is just over 25 yards wide.

[6]Note that we have omitted multiplying the `ShotLie` dummy variable values by zero in the exponent.



Of course, it would be reasonable to ask whether this model, overall, is better than nothing at predicting success at green-hitting. There are R-squared like measures for logistic regression, but the more common measure of model quality is the Likelihood Ratio test. It compares the likelihood of observing the estimated model to the likelihood of the "null hypothesis" model that includes only the intercept term. The test for the model fitted here gives a p-value of zero, meaning that the estimated model is significantly better at explaining the observed variation than a model with only the intercept term.

Another measure of goodness-of-fit for a logistic regression (or any classification model) is to look at the *confusion matrix*, which displays the number of true and false positives and negatives predicted by the model. The confusion matrix for the logistic regression model above is:

Predicted Outcome	Actual Outcome	Shots	Percent
On Green	On Green	311	67.17
On Green	Off Green	79	17.06
Off Green	Off Green	53	11.45
Off Green	On Green	20	4.32

Overall, the model fits the actual data pretty well; it accurately predicts the actual outcome 79 percent of the time.

Conclusion and Recommendations for Strategy

To a large extent, this paper has confirmed, and quantified, several basic principles that any experienced golfer (professional or amateur) or serious fan of the game understands about approach shots:

- Expected proximity to the hole increases with the length of the shot into the green
- A player typically can get the ball closer to the hole, and has a much better chance of hitting the green, from the fairway versus the rough or fairway bunkers
- Small, tight greens are tough to hit, especially from longer distances

In quantifying the specific marginal effects of factors like length of shot, lie, and course characteristics for a specific player (in our case, European Tour member David Horsey), this paper has shown how relatively simple statistical modeling techniques can help pinpoint expectations and, hopefully, help a player and his or her caddie make better course management decisions.

The analysis here also suggests potentially fruitful topics of further research. First and foremost, it would be instructive to compare the model coefficients for David Horsey to his fellow competitors, both on the European Tour and the PGA Tour. It would also be helpful to collect more observations of intermediate rough and tee lies, to see if a larger sample would boost the significance of those variables. It is disappointing that weather data proved insignificant in the models fitted here; perhaps more frequent observations, that introduced greater variance, would open up possibilities there.

The author thanks 15th Club for sponsoring this Hackathon, and for providing the data that allowed for the fitting of these models. It has been a lot of fun!

To explore the R/RMarkdown source code used to create this paper, see https://github.com/scottcame/15th-club-hackathon-2017.

Author Contact Information

Scott Came

Olympia, Washington, USA

scottcame10@gmail.com

360-529-2938