PGA Tour Putting Probability

Lei Shi and Shalima Zalsha

Southern Methodist University

Executive Summary

One of the skills that enable players to lower their golf score is the ability to successfully hit the ball into the hole once it is on the green, in other words to putt. However, a golfer's likelihood to successfully putt the ball into the hole depends on several factors such as the putting distance and the golfer's skills level. In order to predict the binary outcome of whether a putt would succeed, logistic regression models are developed based on a variety of factors from comprehensive shot-level data on Professional Golfers' Association Tour (PGA Tour).

In model development, the explanatory variables considered include distance, slope, score, leaderboard rank, hole sequence, player, and slope-distance interaction. Variables such as score, hole sequence, and leaderboard rank were created algorithmically using information available in the data set. Two logistic regression models were built separately: the first model (model 1) was fitted for Rounds 1, 2, and 3, in which most PGA Tour players participate. The second model (model 2) was fitted for Round 4, in which only top-finishers play. The models were obtained by backward variable selection method.

The results suggest that distance and hole sequence have a negative impact on putting's success probability in both models. In terms of scoring, putting for a par or worse contributes to a higher putting success. For par values, putting in a par-five hole is more likely to succeed compared to putting in par-three hole. In terms of slope of the putting green, compared to level shots, uphill slopes have a negative impact on success rate, and downhill slopes have an even worse impact on success rate.

For players of interest, Rickie Fowler, Jason Day and Jordan Spieth putt better compared to players who are not in the list (reference level). On the other hand, Bryson DeChambeau and Xander Schaeuffele putt with worse or indifferent success rate compared to other players.

Testing the models using 2018 tournaments data, the false positive rate for future prediction is 26% and 21% for model 1 and 2 respectively, while the false negative rate is 10% and 9% for model 1 and 2 respectively. Future models should consider incorporating weather information as predictors. Prediction for rookie or incoming players using this model should be interpreted with caution, since the model was developed based on established PGA Tour players' data.

1. Introduction

Golf is a sport in which players aim to hit a ball into a series of holes in a golf course with as few strokes as possible. In a golf tournament, the winner is typically the player that finishes the tournament with smallest number of strokes and thus with the lowest score. Putting is considered one of the most important skills to lower their scores. A golfer's likelihood to putt the ball into the hole depends on several factors such as the distance from the ball to the hole, the slope of the ball's travelling path, and the golfer's skills level. Many other factors other than putting distance could have an impact on the putting's success. However, previous putting probability study by Gelman (2002) did not include factors other than distance due to the lack of detailed and comprehensive data at the shot level.

In this study, logistic regression models are developed to predict the probability that a Professional Golfers' Association (PGA) Tour golfer will sink a putt based on a variety of factors. The models are built and tested using shot-by-shot PGA Tour tournaments data between 2014 and 2018. The data was provided by Shotlink®, a PGA Tour platform that collects and manages PGA Tour data for experimentation and non-profit studies.

2. Method

We focused our research to develop a statistical model that will predict the probability that a PGA Tour golfer will successfully make their putts given a variety of factors. In addition, understanding how putting probability is impacted by each of the factors is also an important aspect of the study. Therefore, we used logistic regression to describe and analyze the relationship between a binary response variable (whether a putt is success) and the putting conditions such as distance to the hole and the slope of the putting green. In this section, preliminary steps and methodology to build the putting probability model will be explained.

2.1 Data Description and Processing

The putting probability models was developed using a data set that consisted of every putting shot in a PGA tournament between 2014 and 2018; the data were provided by Shotlink®. Because the study only concerns putts, all shots attempted from outside the putting green were excluded. Moreover, the study data set also excludes match-play tournaments, in which players compete at hole-to-hole basis rather than by aggregated scores for the holes they play.

Existing Variables

Variables readily available in the data set are displayed in Table 2-1 below.

Table 2-1: Description of variables in the data set necessary for data processing and modeling.

Variable Name	Description	Format
Year	Year of the tournament	Numeric
Tourn	Tournament ID	Numeric
TournamentName	Name of the tournament	Character
Player	Player ID	Numeric
Round	Round number in a tournament	Numeric
PlayerFirstName	Player's first name	Character
PlayerLastName	Player's last name	Character
IntheHoleFlag	Whether the ball is in the hole	Numeric
Slope	The slope of the ball's path (level, uphill, downhill)	Character
Time	Military time when the shot was made (1-2400)	Numeric
FromLocationScorer	Area from which the shot was made (i.e. tee box, green)	Character
Shot	The shot number for the specific hole	Numeric

Additional Variables

In addition to the original variables listed in Table 2-1, we created the following three variables that were considered important:

- *Score*—a categorical variable indicating whether the shot is for a birdie or better, par, bogey, and double bogey or worse
- *HoleSequence*—a numerical variable identifying the order in which the holes were played
- *LeaderBoard*—a numerical variable indicating the player's rank or position relative to the other players in the tournament before the shot was made

2.2 Model building

The model is developed with the goal to describe and analyze the relationship between putting's success and several explanatory variables listed in the previous section.

Table 2-2 lists both dependent and candidate explanatory variables as well as t in the initial model.

Table 2-2: List of variable effects considered and how they were treated in the model.

Variable Type	Variable Name	Description
Dependent variable	Putting success (IntheHoleFlag)	Binary
Independent variable	Distance (Distance)	Quantitative
	Squared Distance (Distance2)	Quantitative
	Slope (Slope)	Nominal
	Score (Score)	Nominal
	Leaderboard rank (Leaderboard)	Discrete
	Hole sequence (HoleSequence)	Quantitative
	Player (Player)	Nominal
	Slope-Distance interaction	
	Slope-SquaredDistance interaction	

Two models were produced: a model for Rounds 1, 2, and 3 combined and a model for Round 4. Moreover, the *player* variable also requires re-categorization.

Tournament Round as Sub-Populations

In the regular format of a PGA Tour tournament, each golfer plays at least three rounds and only the top players earn the chance to play in Round 4. Therefore, players in Round 4 are typically better players than those in the previous rounds. Since participating players are very different, so are the competition dynamics, two independent logistic regression models were built: one for Rounds 1, 2, and 3 and one model for Round 4.

Selected Players Effect

Due to the large number of players in the PGA Tour between 2014 and 2018, the analysis was performed with a focus on several players. That is, we only consider effects for players in the pre-selected list. The remaining players were grouped together as *"others"* and serve as a baseline.

Both models are fitted with all variables listed in Table 2-2, and backward selection method to rule out variables whose coefficient is not statistically significant until the model cannot be further reduced. The models were fitted using 2014 to 2017 data as training set and were validated with 2018 data. The final model is explained in the Result section.

3. Results

3.1 Model 1: Putting probability model for Round 1, 2 and 3.

The fitted model for putting in Round 1, 2 and 3 is:

$$\begin{aligned}\log\left(\frac{\pi_i}{1-\pi_i}\right) = {} & 1.33 - .02\ (Distance_i) + 0.000013(Distance_i^2) - 0.0032(HoleSeq_i) \\ & -0.0029(Rank_i) + \sum_{j=1}^{2}\beta_{slope_j}\left(I_{i,slope_j}\right) + \sum_{k=1}^{4}\beta_{score_k}\left(I_{i,score_k}\right) \\ & + \sum_{l=1}^{20}\beta_{player}\left(I_{i,player_k}\right) + \sum_{m=1}^{2}\beta_{par\ value_l}\left(I_{i,par\ value_m}\right) + \\ & \sum_{p=1}^{2}\beta_{distance\times slope_i}\left(I_{i,slope_p}\right),\end{aligned}$$

where π_i is the probability of success of putting attempt i, and β is the estimated effect for each level of the categorical variables. These estimated effects are summarized in Table 3-1 below.

Table 3-1: Estimated effect for categorical variables in putting probability model for Rounds 1, 2, and 3.

Slope	Est.	Score	Est.	Par value	Est.	Distance×Slope	Est.
Level	-	Birdie	-	Three	-	Level	-
Downhill	-0.41	Par	1.39	Four	0.053	Downhill	0.001
Uphill	-0.08	Eagle-	-1.47	Five	0.454	Uphill	0.001
		Bogey	1.86				
		Double Bogey+	1.35				

- Indicates baseline.

Similarly, player specific effect was estimated and summarized in Table 3-2. The reference level for the player's effect is other players in the data that are not included in the list as explained in Section 2.2 Model building.

Table 3-2: Estimated effect ($\hat{\beta}$) for players in Round 1,2, and 3 ordered by the effect magnitudes.

Order	Name	$\widehat{\boldsymbol{\beta}}$	$e^{\hat{\beta}}$	Order	Name	$\widehat{\boldsymbol{\beta}}$	$e^{\hat{\beta}}$
1	Henrik Stenson	0.41	1.51	11	Patrick Reed	0.25	1.28
2	Rickie Fowler	0.39	1.48	12	Bubba Watson	0.23	1.26
3	Adam Scott	0.35	1.42	13	Justin Thomas	0.22	1.25
4	Sergio Garcia	0.35	1.42	14	Webb Simpson	0.20	1.22
5	Jordan Spieth	0.34	1.4	15	Dustin Johnson	0.19	1.21
6	Jason Day	0.33	1.39	16	Tiger Woods	0.16	1.17
7	Brooks Koepka	0.32	1.38	17	Tony Finau	0.11	1.12
8	Zach Johnson	0.29	1.34	18	Jon Rahm	0.09	1.09
9	Phil Mickelson	0.26	1.3	19	Bryson DeChambeau	0.02	1.02
10	Justin Rose	0.26	1.3	20	Xander Schauffele	-0.06	0.94

3.2 Model 2: putting probability model for round 4.

Fitted model for Round 4 is:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 1.99 - .02\,(Distance_i) + 0.00002(Distance_i^2) - 0.0021(HoleSeq_i)$$
$$-0.0006(Rank_i) + \sum_{j=1}^{2}\beta_{slope_j}\left(I_{i,slope_j}\right) + \sum_{k=1}^{4}\beta_{score_k}\left(I_{i,score_k}\right)$$
$$+ \sum_{l=1}^{20}\beta_{player}\left(I_{i,player_k}\right) + \sum_{m=1}^{2}\beta_{par\ value_l}\left(I_{i,par\ value_m}\right) +$$
$$\sum_{p=1}^{2}\beta_{distance\times slope_i}\left(I_{i,slope_p}\right),$$

where π_i is the probability of success of putting attempt i, and β is the estimated effect for each level of the categorical variables. These estimated effects are summarized in Table 3-3 below.

Table 3-3: Estimated effect for categorical variables in putting probability model for round 4.

Slope	Est.	Score	Est.	Par value	Est.	Distance×Slope	Est.
Level	-	Birdie	-	Three	-	Level	-
Downhill	-0.58	Par	0.96	Four	0.016	Downhill	0.002
Uphill	-0.87	Eagle-	-1	Five	0.283	Uphill	0.004
		Bogey	1.35				
		Double Bogey+	0.89				

- Indicates baseline.

Player specific effect is summarized in Table 3-4. The reference level for the player's effect is other players in the data that are not included in the list as explained in section 2.2 Model building.

Table 3-4: Estimated effect ($\hat{\beta}$) for players in round 4 ordered by the effect magnitudes.

Order	Name	$\widehat{\boldsymbol{\beta}}$	$e^{\widehat{\beta}}$	Order	Name	$\widehat{\boldsymbol{\beta}}$	$e^{\widehat{\beta}}$
1	Jason Day	0.48	1.62	11	Zach Johnson	0.26	1.3
2	Rickie Fowler	0.48	1.62	12	Henrik Stenson	0.22	1.25
3	Jordan Spieth	0.45	1.57	13	Justin Thomas	0.21	1.23
4	Tiger Woods	0.40	1.49	14	Dustin Johnson	0.21	1.23
5	Phil Mickelson	0.39	1.48	15	Xander Schauffele	0.17	1.19
6	Brooks Koepka	0.33	1.39	16	Justin Rose	0.13	1.14
7	Patrick Reed	0.32	1.38	17	Sergio Garcia	0.12	1.13
8	Bubba Watson	0.29	1.34	18	Tony Finau	0.10	1.11
9	Adam Scott	0.29	1.34	19	Jon Rahm	0.04	1.04
10	Webb Simpson	0.26	1.3	20	Bryson DeChambeau	-0.05	0.95

3.3 Effect Interpretations.

The estimated effects (β) is the change in log odds for successful putting, while keeping all other variables constant. In other words, the odds for successful putting will change at e^{β} rate. For numeric variables, for every one-unit increase, the odd for successful putting will increase (or decrease) by e^{β} times. For example, the effect of distance was estimated to be $.02$ in model 2. Therefore, keeping all other variables constant, increasing the distance by one inch will

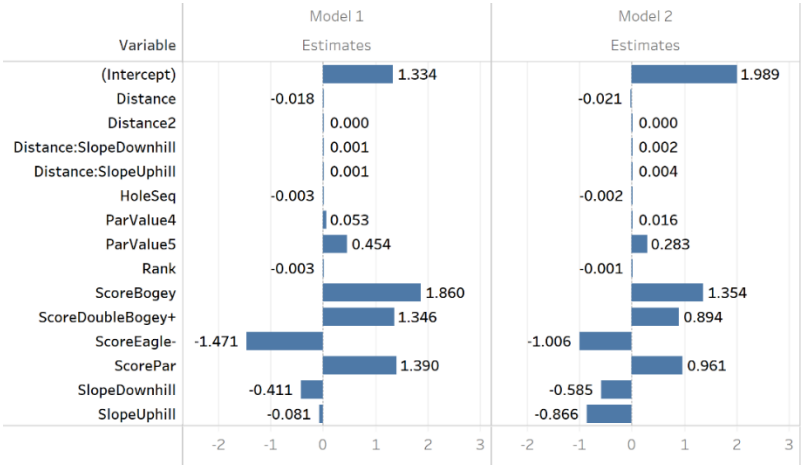
increase the odd for successful putting by $e^{.02} \approx 1.02$ times. Therefore, increasing the distance by 12 inches will increase the odd for successful putting by $e^{.02(12)} \approx 1.27$ times.

For categorical variables. being a member of a specific factor level makes the odds of success e^{β_k} times of being at the reference level for the factor. For example, the estimated player effect for Jason Day in model 2 is 0.48. It means that, with all other variables held constant, the odd for successful putting is $e^{.48} \approx 1.62$ times of for Jason Day, compared to the reference level for the player effect, which includes the other players who are not in the pre-selected list of players, as explained in Section 2.2 Model building.

3.3 Effect Comparisons

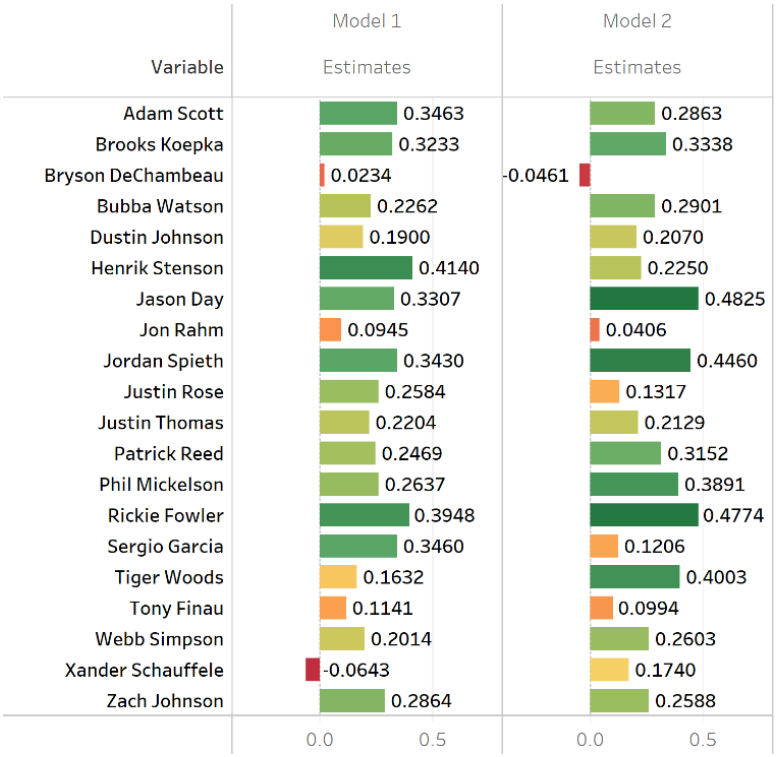
The estimated effects for all variables except for player are summarized in Figure 1, and the estimated effect for the players are summarized in that follows. As shown in both figures, the effects are consistent in direction but slightly differ in magnitudes for the two models. As can be seen in Figure 1, both distance and hole sequence have a negative impact on putting's success probability. On the other hand, par value five have a positive impact on putting's success probability. Moreover, uneven slopes have a negative impact on putting's success probability, but the impact is worse for downhill slopes compared to uphill slopes.

Figure 1: Effect estimates for all variables other than the player effect



As illustrated in Figure 2, Rickie Fowler, Jason Day and Jordan Spieth have higher success probability on putting compared to players not in the list (reference level). Meanwhile, Bryson DeChambeau's and Xander Schaeuffele's success probability are relatively low; and other players in the list are not significantly different from the reference level.

Figure 2 : Estimated player effect from the logistic regression model



3.4 Model Performance for future prediction.

To test the performance of the model, the data set is split into training and test sets. All putting data prior to 2018 were used as a training set, whereas the 2018 data were used to test the prediction of future putting outcomes. False positive rate is 26% for model 1 and 21% for model 2, while their false negative rates are 10% and 9%. Both models are a little too optimistic in predicting future of putting success, as reflected by the relatively higher false positive rates than false negative rates.

4. Discussion

While the models were built carefully using the available data, there are still limitations, including lack of weather information and the prediction scope is limited to long-term PGA players.

Variables in the model building process was chosen carefully to achieve an optimal model. However, not all variables desired in the model were included in the ShotLink® data set. For example, weather is a factor that is important to consider when predicting a player's performance. Therefore, weather should be considered in future analyses for putting probability, which will require an additional source of information related to weather; that source of information can be join with the ShotLink® data set so that weather-related variables are part of the analysis.

The training data set consists of long-term PGA players, players, and rookie players are underrepresented in the data because most PGA Tour players are professional golfers. Therefore, although the model may perform well to predict putting outcomes for professional golfers, the model may not predict putting probability for a rookie or a rising player. Therefore, predictions for players who are different from a typical PGA golfer should be used with caution.

5. References

Gelman, A., & Nolan, D. (2002), A Probability Model for Golf Putting. *Teaching Statistics, 24*, 93-95. doi:10.1111/1467-9639.00097