STAT 515 Applied Statistics and Visualizations for Analytics

How important are tee-shots and do players get “hot” from the tee?

by:
Matthew Oldham

May 2017

Table of Contents

1 Introduction

US president Woodrow Wilson perfectly sums up most people's feeling towards golf with the following quote "an ineffectual attempt to put an elusive ball into an obscure hole with implements ill-adapted to the purpose". However, the popularity and value of golf is huge, with it adding about $70 billion a year to America's economy (The Economist, 2015). For all but a few superstars of the game, the secret to consistently low scores remains an unsolved mystery, with most golfers guided by the heuristic "drive for show, putt for dough". This paper adds to the discussion of identifying the determinants of golf performance by attempting to quantify the affects that tee-shots have on finishing inside the top 5, 10 of a US PGA Tour tournament. This is achieved by employing numerous classifications models on a dataset from the US PGA Tour that includes the strokes-gained and driving rank metrics.

This paper also investigates whether US PGA tour players exhibit "hot" or "cold" streaks from the tee. These streaks in general have become popularized as "hot" or "cold" hands, and relate to when a player has an identifiable run of performance above or below what is expected, i.e. the player is "hot" or has gone "cold". The existence of such events has generally been debunked, but by utilizing the US PGA SHOTLINK data, a novel approach is undertaken in this paper.

2 Background

2.1 Theory

The debate surrounding whether professional athletes exhibit streaky performance has continued in earnest since Gilovich, Vallone, and Tversky (1985) published their seminal paper, in which they investigated the phenomenon of the "hot hand" in professional basketball. The term "hot hand" refers to when a player's performance is significantly better than would be expected by chance (Livingston, 2012). The effect manifests itself in the belief amongst players, and spectators, that a player is more likely to make their next shot if they have just made a shot, and less likely if they have just missed, hence they tend to perform in streaks.

Researchers have also tried to understand whether streakiness exists in golf. The justification for researching golf is that compared to continuous sports; such as basketball, a player has more time to reflect on past performance and prepare for the next shot in a more systematic manner (Clark, 2005b). Arkes (2016) further reinforces the argument that golf is an ideal sport to investigate the existence of the "hot hand" because of the standardized scoring and shot system. In addition, given a round is played over 18 holes the role that randomness plays in the outcome tends to be removed (Arkes, 2016).

The initial papers of Clark (2003a), and (2003b), researched whether players had a tendency for their sub-par rounds to be clustered together. A sub-par round being where a player completes a round (18 holes) in fewer strokes than the predetermined allowable strokes for the course (the course par). While the results indicated, players tended to have their sub-par rounds clustered together, the difficulty of the course was found to be the main determinant. Clark (2005a) and (2005b) expanded the approach to look at the hole-to-hole streakiness of players. Consistent with the initial findings of Gilovich et al(1985), Clark found no evidence to indicate that a player was more or less likely to score under or over par following recording an under/over par score.

In an approach closer to what is undertaken in this paper, Livingston (2012) reassessed the issue of the "hot hand" in golf by examining the separate effects of bad (above par) or good (below par) outcomes on one hole on the probability of a bad or good performance of the next hole. The paper found that evidence of the "hot (and cold) hand" can be found but it tends to be hidden because the focus of the analysis had previously been at an overall mean impact level rather than accounting for the experience of the players, and by allowing for this there is greater evidence supporting the existence of streaky play.

Arkes (2016) makes a further contribution to the research topic, when amongst numerous changes, he introduced performance measures that resulted in greater variation than sequential shots of under/over par rounds, and produced a model that included all players rather assessing each player individually. These measures grouped the player's score relative to par for either a block of 3 6, 9 or 18 holes. Despite the new approach, the evidence of the "hot hand" proved elusive but there was evidence supporting the "cold hand", which is important because as the author suggests golf is a game of misses.

A key innovation in golf has been the introduction of the strokes-gained metric, as developed by Broadie (2012). As explained in Broadie (2012), "strokes-gained represent the decrease in the average number of strokes to finish the hole from the beginning of the short to the end of the shot, minus one to account for the stroke taken". Compared to traditional metrics it is superior because it compares a player's performance to the rest of the field and can isolate individual aspects of the game. Traditional golf statistics, such as greens in regulation and putts per green, are influenced by a player's performance on shots other than those being measured. Broadie utilized a variance decomposition to conclude that the long game explains about 72% of the variability of golfers' overall skills (Levin, 2017). However, Broadie (2012) notes that variability does not equate with importance, thus leaving open the question as to what skills are more important in determining the results on the PGA Tour.

By using the correlation between a player's skill level for various shots type and future results, Levin (2017) suggested that it is tempting to state that tee-shots on par 4s and par 5s is the most important golf skill. However, per the author it is more complicated because there are two reasonable but different meanings of "importance", one in terms of the season, and one with regards to a tournament. Utilizing the different definitions lead to different orderings of the skill categories. To predict future performance (performance across the season) Levin (2017), found that a player's ability to drive the ball was most valuable. However, in explaining the results of a single round, a player's performance on the green was most valuable. The explanation being that there is more randomness within a round compared to across a season, therefore, putting performance was more important in determining relative performance. Alternatively, given randomness tends to even out across a year, tee-shots and the long game were found to be more important.

2.2 Data Description

The objective of this paper is to investigate whether a player gets "hot" (or "cold") from the tee and if so what are the ramifications of their driving performance on overall performance. To achieve the research objective, data from the US PGA's SHOTLINK system for each player from the 2014, 2015, and 2016 seasons, and their strokes-gained from the tee was analyzed. There are over 140,000 records for each season. Data pertaining to how each player performed in each tournament was sourced from the PGA Tour website. This data was matched with the strokes-gained data to determine the influence of driving performance on where a player finished.

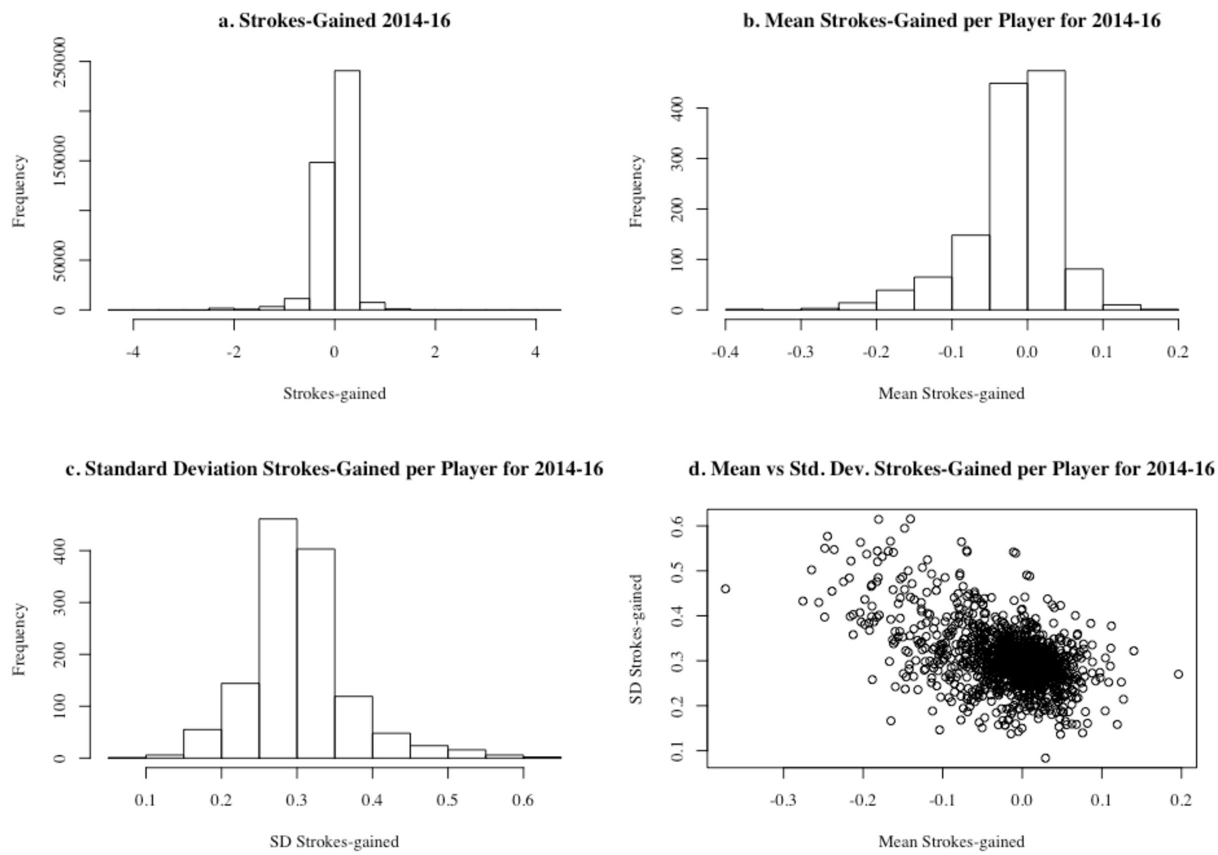

Figure 1: Graph (a) plots the raw distribution of strokes-gained from the tee, with graph (b) providing the distribution after adjusting for a player's season average. The standard deviations for the player's strokes-gained is seen in graph (c) and is plotted against the players' mean in graph (d).

While the strokes-gained, metric is useful, an issue arises that requires additional steps to be included for the analysis in this paper. Because the strokes-gained is effectively a comparison against the average player, so as illustrated in Figure 2a, the mean for strokes-gained from the tee is centered around 0. Therefore, players that are comparatively better drivers, Dustin Johnstone for instance, is more likely to record "hot" streaks as a function of being an above average driver. To overcome this, each player's driving performance needs to be compared to their own expected strokes-gained from the tee for a given season. This allows the analysis to detect, or otherwise, if a player recorded a meaningful period of superior performance other than what could be normally expected from that player.

To make this adjustment, a player's strokes-gained for each tee-shot (SGFT) has their season long average strokes-gained from the tee subtracted. In addition, an additional metric (SGFTB) was calculated where the SGFT was converted to a binary string based on whether the SGFT was positive, that is if SGFT was greater than 0 (a "hot" shot) then SGFTB was 1, otherwise it was 0. Figure 1(b), a histogram of the average strokes-gained from the tee for each player for the 2014-16 seasons, shows the effect of assessing the average performance for each player. The distribution is left skewed, with more players averaging a negative strokes-gain off the tee across the season, when compared to the distribution of just shots (Figure 1(a)).

Figure 1(c) illustrates the histogram of the standard deviations. The rationale for analyzing the is to assess whether some players are more erratic than others, and whether this was detrimental to their success. The distribution is also skewed but this time it is right skewed, indicating players in general tend to be more eradicate off the tee. The erratic players will be of interest to see whether this eradicate behavior off the tee was detrimental to their finishing position in the tournament. An initial analysis of the relationship between a player's mean strokes-gained and the standard deviation is provided in Figure 1(d). The results indicate that the more eradicate a player, the lower their strokes-gain result is.

3 Modelling Approach

3.1 Hot Hand Detection

To identify whether any players produced a "hot" or "cold" streak the Wald-Wolfowitz (1940) runs test, a test for randomness among continuous data, was utilized. The test was implemented in R (2017) using the randtest package (Caeiro & Mateus, 2014). The variable used to try and detect a streak was the player's strokes-gained for a given tee-shot (SGFT) minus their average strokes-gained from the tee for the season (SASG). A "hot"("cold") shot is defined as a shot where a player records a strokes-gained result better (worse) than their season average.

The default threshold used in the test is the sample median for the player. One downside to the test is that it does not automatically indicate whether the streak, that occurred, was a "hot" or "cold" streak. To help determine this the package provides a plot such as Figure 2. In this example the player experienced a positive streak towards the end of the season. Further analysis of this was undertaken in Tableau.

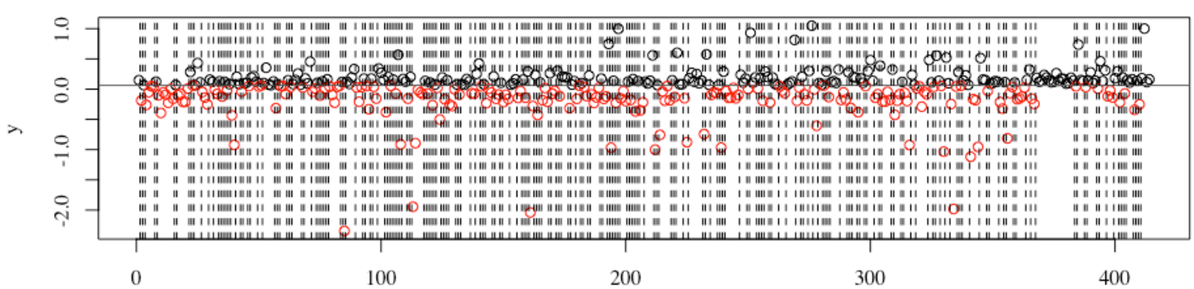

Figure 2: An example of a runs test where a player did have a hot streak. The x axis represents the player's shots ordered sequentially, and the y axis is the player's strokes gain per shot – minus their season average.

3.2 Tournament Placing

To answer the second research, of assessing the effects of a player's driving performance, a variety of separate statistical models were utilized to assess whether a player's chance of finishing in the Top 5, Top 10 or Top 20 of tournament improved based on their driving performance. This approach reduces the question to a classification problem, that is, did the player finish in the Top 5 etc. or not? The methods used were logistic regression, and classification and regression tree (CART) models, where both a bagging and random forest implementations were included.

4 Results

4.1 Hot Hand Detection

The results of the runs-test for the three seasons are illustrated in Figure 3. A player is included in the "Streaks" category if they returned a p-value of .05 or less from the Wald-Wolfowitz runs test. The percentage of players that returned a statistically significant streak for the entire sample was 6.1%, with the result for 2014, 2015, and 2016 being 6.1%, 4.8%, and 7.2% respectively. While not overwhelming, the data does suggest that there is evidence of players exhibiting streaky performance from the tee.

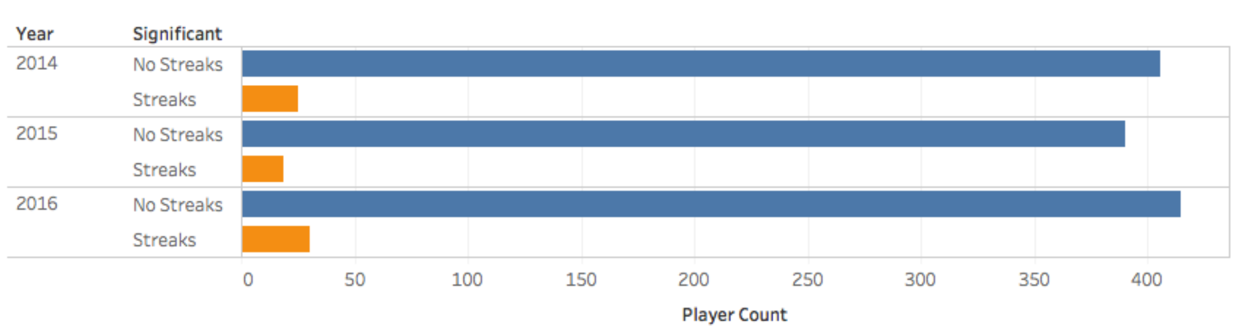

Figure 3: The breakdown of players recording a streak. *Chart produced in Tableau*

Figure 4 provides comparative boxplots that illustrates the spread of the largest positive and negative streaks for each player, with the data spilt by whether the players recorded a statistically significant streak (either a "hot" or "cold" streak) in each year. It is evident that the length of the players' "hot" streaks are longer than their "cold" streaks, as seen by the top chart which plots the maximum positive run for the players. This suggests that players are aware that they are performing below their expectation and can adjust their approach to their subsequent tee-shots, to limit further poor shots, while riding their luck when they are performing well.

Not surprisingly, the median streak is longer in the sample where players returned a significant streak. However, there is some evidence that some players recorded an extended streak, yet they were not statistically significant. Using the "Player Scatter" sheet (as seen in Appendix A) in the included Tableau file, each player can be identified in terms of their maximum "hot" and "cold" streak and whether it was statistically significant.

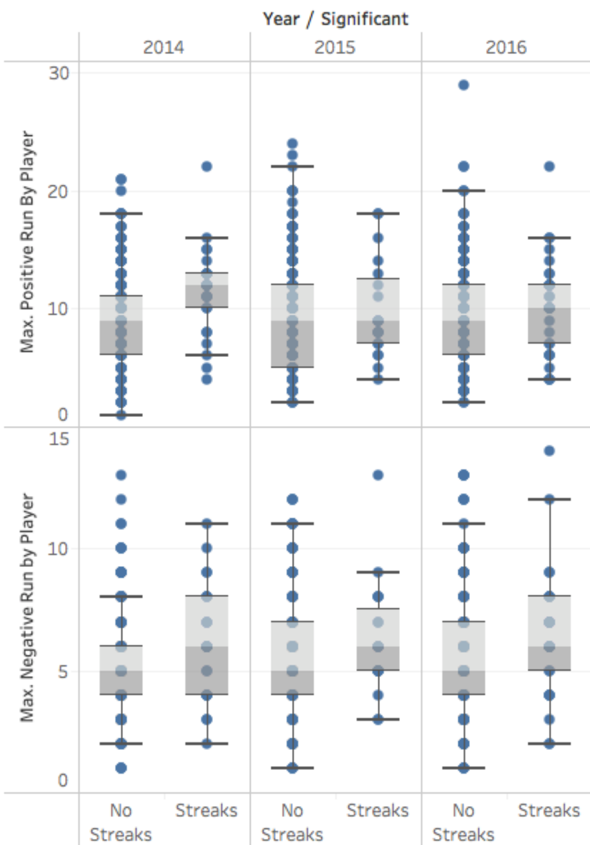

Figure 4: Boxplot for each of the seasons showing the distribution of player streaks and whether they were significant or not. *Chart produced in Tableau*

4.2 Tournament Placing

Before specifying the logistic regression, the available data was assessed for its usefulness. In addition to the strokes-gained data, the PGA Tour Data provides several other driving statistics associated with a player's performance at a specific tournament. To assess the predictive power of the variables, a correlation plot as per Figure 5 was created. Total Driving Rank, which is a tournament specific metric, is calculated as a combination of a player's driving distance and accuracy for the tournament, is correlated with Top 20 finishes. The correlation is negative because the best player is ranked 1, with other players ranked in a descending ordering, and if a player misses the cut their rank is assigned as 999. The plot also confirms the findings reported from Figure 1(d), but this time for individual tournaments that players with a higher stroked gained average (MSG), have a low standard deviation for this measure (sdSG). The definition of the other metrics are:

- Mean for strokes-gained (MSG): This metric is the mean of a player's strokes-gained from the tee for the tournament.

- Mean run – binary (Mrun): the average for a given tournament of a player's binary strokes-gained per tee-shot (SGFTB), where 1 meant a tee-shot was better than their season long average (SASG), and 0 meant a tee-shot worse than their SASG. For example, if a player had 72 tee-shots all better(worse) than their SASG then Mrun would equal 1(0) for the tournament.

- Mean run – shot (Mrun1): the average for a given tournament of a player's strokes-gained per tee-shot minus their SASG.

- Standard deviation of strokes-gained (sdSG): This metric is the standard deviation of a player's strokes-gained from the tee for a given tournament. Given the correlation between this metric and sdRun1, sdRun1 was not considered.

- Standard deviation for SGFTB (sdRun): This metric is the standard deviation of a player's binary strokes-gained per tee-shot for a given tournament.

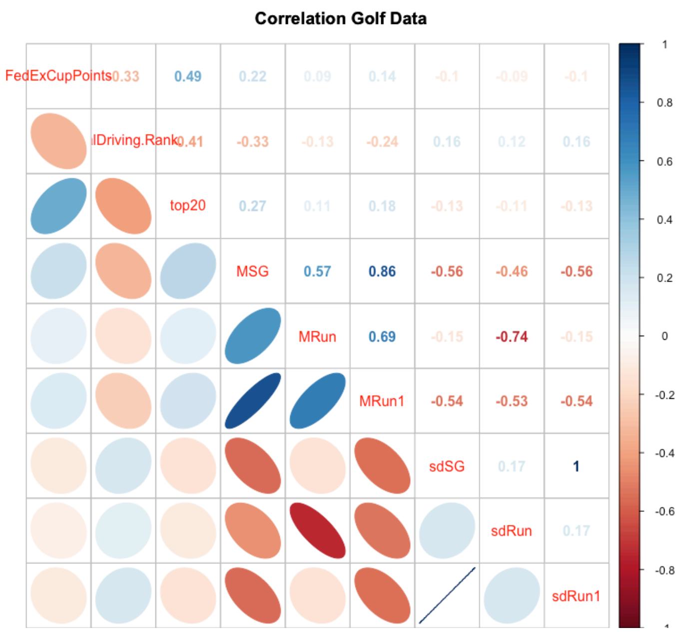

Figure 5: Correlation plot for selected variables

The results of running logistic regression models for predicting the probability of finishing in the Top 5, 10 or 20 are contained in Table 1. The model was specified by including all variables seen in Figure 5, and then running a backward induction algorithm to identify the most relevant and statistically significant variables, with the final model being manually specified. It should be noted that the intercept was removed based on the rationale that if you do not tee off you will have 0% of finishing in the relevant places.

Table 1: Results of the logistic regression

	Top 5	Top 10	Top 20
MSG	11.629122*	10.784744*	12.777862 *
MRun1	-9.203037 *	-6.763943 *	-7.608702*
sdSG	-6.252326*	-4.111417 *	-1.675738*
Total Driving Rank	-0.024127 *	-0.021071*	-0.014270 *

The variable interpretations are:

- **Mean Strokes-gained (MSG):** This is a measure of a player's performance from the tee for an tournament, and indicates that players that generate positive strokes-gained from the tee are more likely to finish within in the Top 5, 10 or 20 for any given even.

- **Mean strokes-gained above/below the player's tournament average (MRun1):** This result is one that makes little initial sense. As the metric measures the mean difference between a player's strokes-gained for each hole in a specific tournament and their average for the tournament, it suggests that if a player on average performs above their tournament average they are less likely to finish within in the Top 5, 10 or 20 for a given tournament. It is hard to rationalize this result, with a possible explanation being that the metric needs to be adjusted for course difficulty and/or the quality of the field, or there is an issue with multicollinearity given the high correlation between the two.

- **Standard Deviation in strokes-gained across the year (sdSG):** This measures the variance in a player's ability across the tournament in terms of their strokes-gained, and indicates that players that are more consistent with their tee-shots are more likely to finish within in the Top 5, 10 or 20 for any given tournament.

- **Total Driving Rank for the tournament:** The lower player's driving rank (that is the better they performed), this being a relative measure for the specific tournament, then the higher the probability of finishing in one of the prescribed places.

Having specified the model, the question turns to assessing the accuracy of the models, and utilizing the regression equations to classify whether a player finishes inside or outside the top 5, 10 or 20. As per Figure 10, which shows the validation accuracy for the models with varying cut-off thresholds, given the large flat portions of these charts it suggests that the models may have only limited ability in predicting where a player finishes.

With the cut-off values set as per the headings in Table 2, the following confusion matrices were returned. The initial impression is that the models do a reasonable job of classifying players that finish outside the relative places, but not so well in classifying players that finish inside the relative places once you move beyond the Top 5. It should be noted that there is an issue with the data set being unbalanced, in that for a given tournament there are around 150 players, with a large percentage of those players finishing outside the Top 20. Future research in the area may address this by using techniques such as Synthetic Minority Oversampling (SMOTE) or re-sampling with a uniform distribution to increase the fidelity of the models.

Table 2: The confusion matrix from the logistic regression

Predicted		**Cut off value = .30 Observed**		**Cut off value = .40 Observed**		**Cut off value = .40 Observed**	
		Not Top 5	Top 5	Not Top 10	Top 10	Not Top 20	Top 20
	Not Top X	2221	107	2086	193	1842	337
	Top X	0	2	31	20	61	90

The initial impressions that the models require farther work is confirmed in Table 3. The table provides various metrics relating to not only the performance of the logistic regressions but also a bagged tree and random forest model. Precision, defined as TP / (TP+FP), which can be interpreted as the proportion of true top 20 finishers among the predicted Top 20 finishers, is the preferred metric in this instance. With results around 50%, it indicates that the models need to be refined, with resampling techniques mentioned previously the most obvious avenue.

Table 3: Model performance metrics

	Top 5	Top 10	Top 20 - LR	Top 20 - BT	Top 20 - RF
Misclassification	4.6%	10%	17%	18%	18%
Accuracy	95.4%	90%	83%	82%	82%
Sensitivity	1.8%	9%	21%	24%	28%
Specificity	95.4%	91.5%	84.5%	84.7%	85.3%
Precision	100%	39%	60%	50%	49%

Having established some minor validity for the general research approach, a bagged tree and random forest model were employed to see if these approaches could provide some further insight to whether a player finished in the Top 20. The rationale for focusing solely on the Top 20 comes from the fact that the sample is slightly more balanced given the larger number of players finishing inside the predicted class.

Figure 6 illustrates the importance scores from the CART models for each of the variables in terms of determining whether a player finished inside the top 20. The clear finding is that from the driving metrics utilized, the driving rank of a player provides the greatest insight into whether a player finishes in the Top 20. Supporting the finding that a player's strokes-gained from the tee is important, is the fact that the MSG variable records the second highest importance measure.

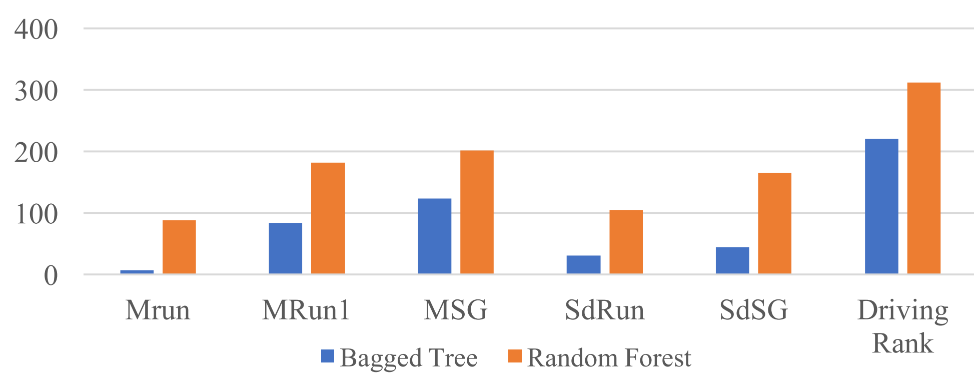

Figure 6: The importance scores for each of the variable as per the two CART methods

The resulting bagged tree produced by the Rpart package, as seen in Figure 7, shows that the driving rank is the first split. In terms of the accuracy of the bagged tree and random forest model, Table 4 provides the confusion matrix. The resulting performance measures of the models, as seen in Table 3, are not that dissimilar to those recorded in the logistic regression model. However, yet again the models are not overly helpful in providing insight into whether a player will finish inside the top 20.

Table 4: Confusion matrix for the bagged tree and random forest models

Predicted		**Bagged Tree Observed**			**Random Forest Observed**	
		Not Top 20	Top 20		Not Top 20	Top 20
	Not Top 20	1802	325		1778	306
	Top 20	101	102		125	121

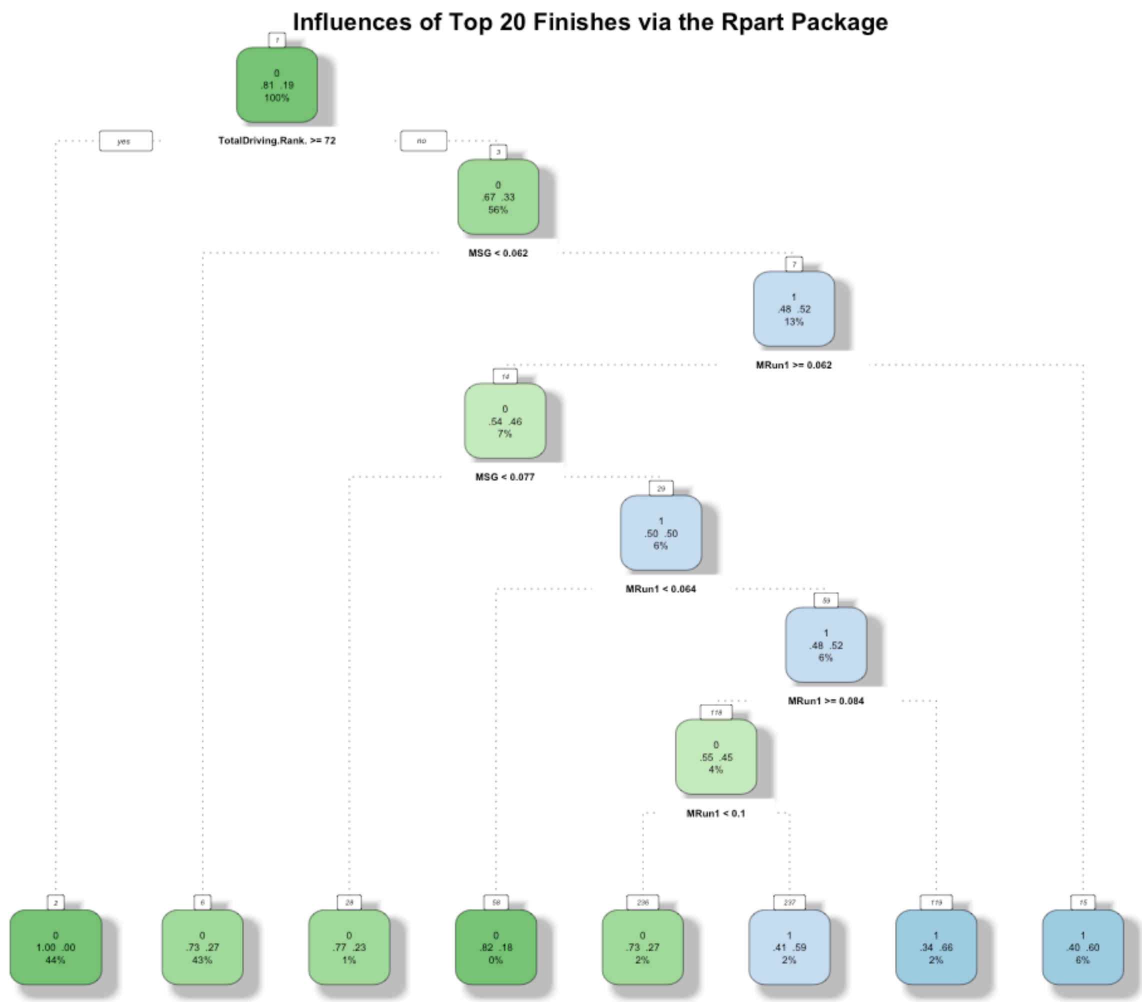

Figure 7: The resulting bagged tree in determing whether a player finishes in the Top 20.

A comparison of the fidelity of the three approaches is seen with the Receiver operating characteristic (ROC) curves[1], with the accompanying area under the curve (AUC) metric. Relatively, the logistic model is the best performing model, which is consistent with the metrics provided in Table 3. While the results indicate all three models have some fidelity, there is

[1] Please note the ROC curve for the random forest model is based on the training data

clearly room for improvement, with the performance of other shot classes, such as putting, and scrambling the obvious candidates.

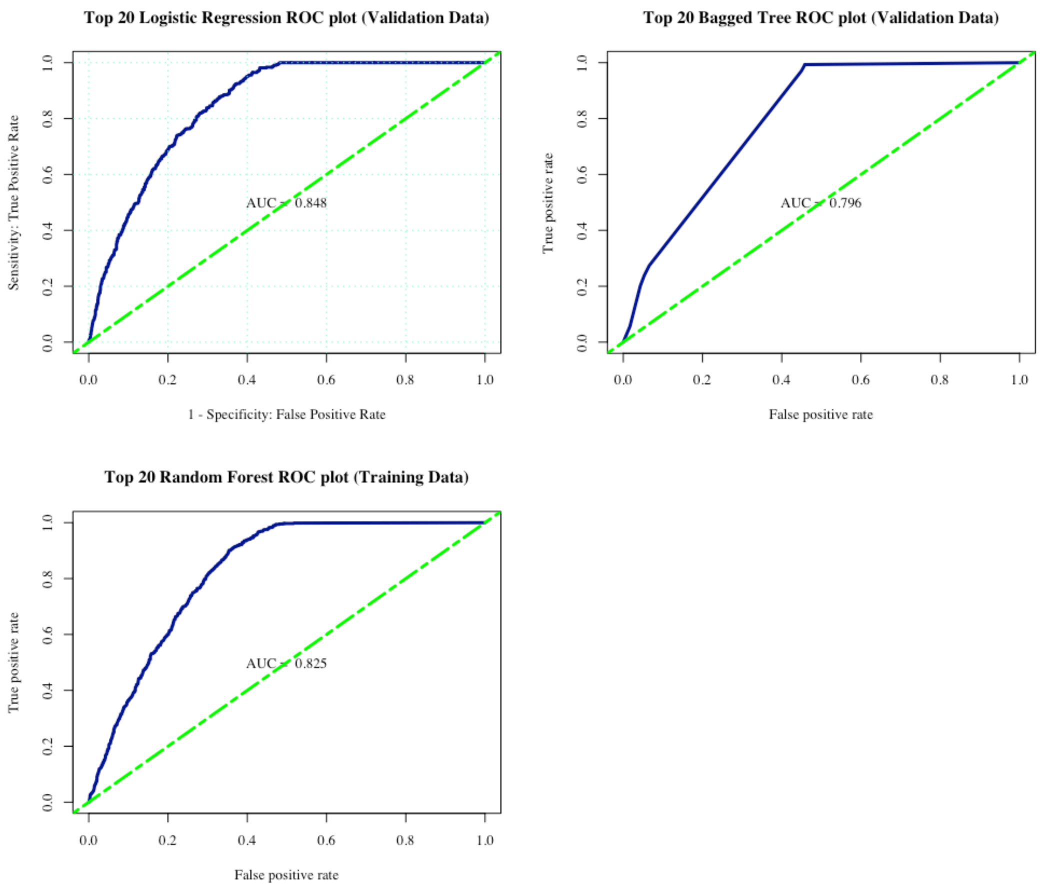

Figure 8: ROC curves and AUC metric for the three models

5 Summary and Conclusion

Utilizing various classification models, some insight was gained into where a player might finish in a PGA tournament by using solely tee-shot data, albeit the greatest insight was that tee-shot performance alone cannot predict whether a player finishes a tournament. Despite this, the tee-shot framework did indicate that relative performance against the field, rather than outright performance is what matters most.

The results pertaining to the detection of a “hot” (or “cold”) hand were more encouraging. While the Wald-Wolfowitz runs test, approach is not without its critics, this approach did provide evidence of a small sample of players performing in streaks. A more encouraging result was the finding that players tended to have longer positive runs of strokes-gained above their season average, meaning that players were likely aware of their relative performance and could adjust their game to either take full advantage of good form to take steps or mitigate poor form.

References

Arkes, J. (2016). The Hot Hand vs. Cold Hand on the PGA Tour. *International Journal of Sport Finance*, *11*(2), 99.

Broadie, M. (2012). Assessing Golfer Performance on the PGA TOUR. *Interfaces*, *42*(2), 146–165. https://doi.org/10.1287/inte.1120.0626

Caeiro, F., & Mateus, A. (2014). *randtests: Testing randomness in R*. Retrieved from https://CRAN.R-project.org/package=randtests

Clark, R. D. (2003a). An Analysis of Streaky Performance On The LPGA Tour. *Perceptual and Motor Skills*, *97*(2), 365–370. https://doi.org/10.2466/pms.2003.97.2.365

Clark, R. D. (2003b). Streakiness Among Professional Golfers: Fact or Fiction? *International Journal of Sport Psychology*.

Clark, R. D. (2005a). An Examination of the “Hot Hand” in Professional Golfers. *Perceptual and Motor Skills*, *101*(3), 935–942. https://doi.org/10.2466/pms.101.3.935-942

Clark, R. D. (2005b). Examination of Hole-to-Hole Streakiness on the PGA Tour. *Perceptual and Motor Skills*, *100*(3), 806–814. https://doi.org/10.2466/pms.100.3.806-814

Gilovich, T., Vallone, R., & Tversky, A. (1985). The Hot Hand in Basketball: On The Misperception of Random Sequences. *Cognitive Psychology*, *17*(3), 295–314. https://doi.org/10.1016/0010-0285(85)90010-6

Levin, A. (2017). Ranking the Skills of Golfers on the PGA TOUR using Gradient Boosting Machines and Network Analysis. Presented at the MIT Sloan Sports Analytics Conference.

Livingston, J. A. (2012). The Hot Hand and the Cold Hand in Professional Golf. *Journal of Economic Behavior & Organization*, *81*(1), 172–184. https://doi.org/10.1016/j.jebo.2011.10.001

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

The Economist. (2015). Why golf is in decline in America. Retrieved from http://www.economist.com/blogs/economist-explains/2015/04/economist-explains-1

Wald, A., & Wolfowitz, J. (1940). On a Test Whether Two Samples are from the Same Population. *The Annals of Mathematical Statistics*, *11*(2), 147–162. https://doi.org/10.1214/aoms/1177731909

6 Appendix

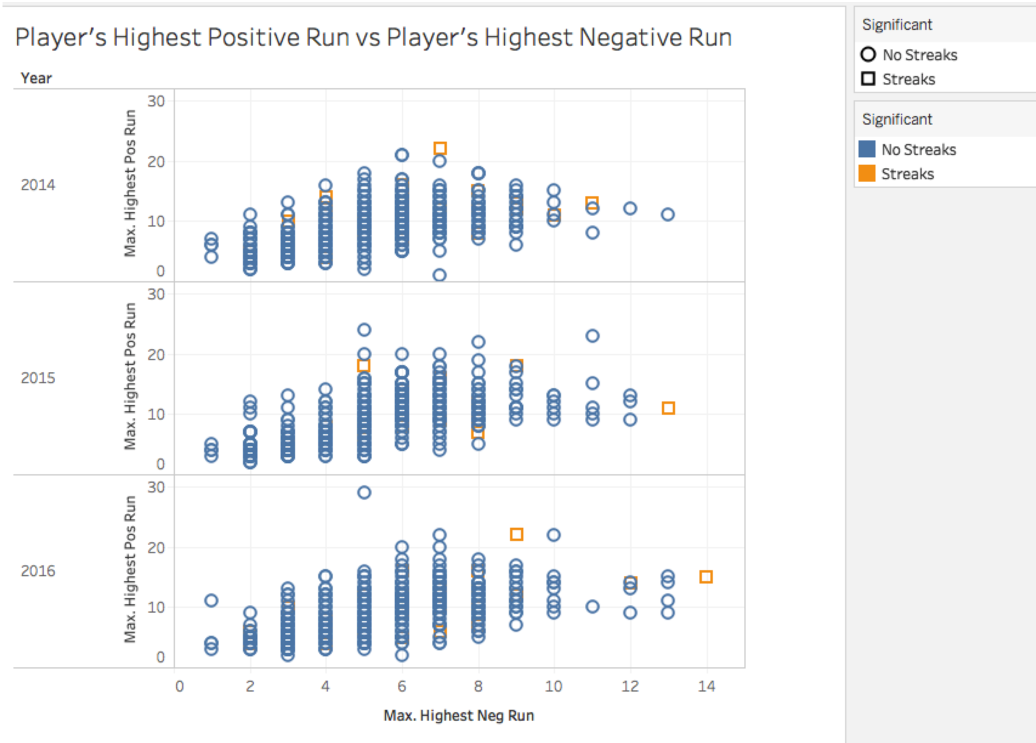

Figure 9: Scatter plots illustrating maximum positive runs vs their negative runs for the PGA players and whether the player recorded a statistically significant streak

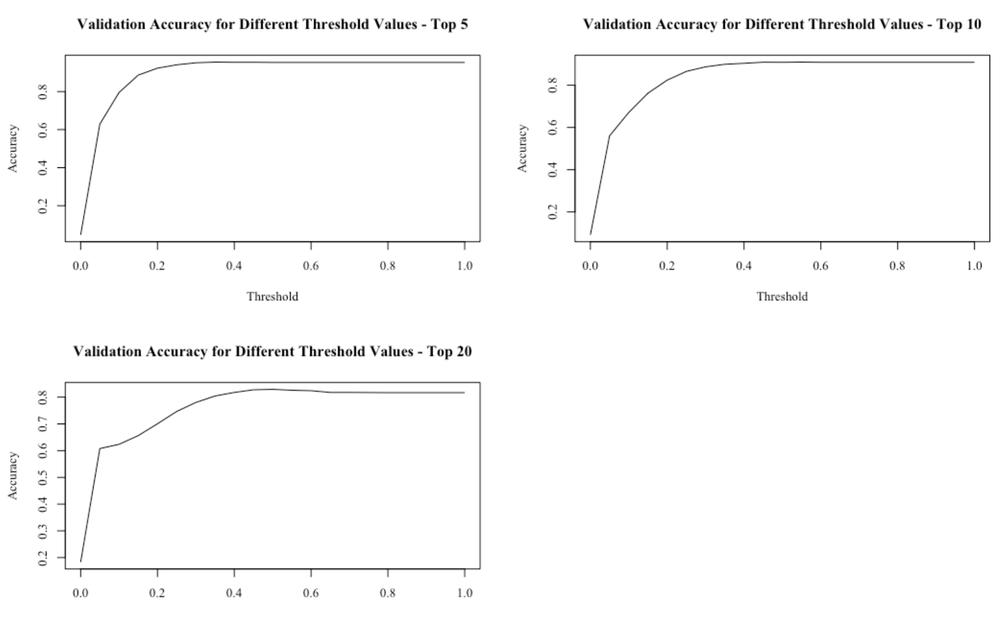

Figure 10: Validation Accuracy curves used to justify the threshold values in the logistic regression models.