PREDICTING PGA TOUR WINS USING BINOMIAL LOGISTIC REGRESSION

BY NATHAN CARNEY

1. INTRODUCTION

Golf has become a sport which is increasingly statistics-driven in recent years (Arastey, 2020). The implementation of data collection systems such as the modern 'ShotLink' database in 1999 has significantly expanded the availability of golf-related statistics across the globe (ShotLink, 2021). The data recorded by the Shotlink system predominantly includes distance and accuracy figures.

Despite the emergence of the Shotlink platform, which James has posited stands as "the future of performance analysis in golf" (James, 2007), research and studies involving golf performance and the impact of specific performance indicators on player success is sparse (James, 2007).

The reason for this sparsity in research can be attributed to the nature of golf as a sport. Stockl and Lamb found that individual golfer performance in the PGA Tour is "chaotic", in that unlike other sports it has proven particularly difficult in golf to maintain a consistent quality of play (Stockl & Lamb, 2018). This in turn makes the prediction of player performance inherently difficult.

Of the few studies published surrounding player performance, many examine how player performance can be improved in areas such as exercise programs (Evans & Tuttle, 2015) and the technical aspects of their golf swing (Parker, et al., 2019) (Myers, et al., 2008), Very few focus on how performance can be predicted based on their quality of play in distinct areas in previous tournaments. As there is a lack of research into this particular aspect of golf, this paper will attempt to present a method to use a golfer's Shotlink statistics to predict their likelihood of achieving a win in a PGA Tour Season.

1.1. BACKGROUND

The PGA Tour comprises of 150 professional golfers (Golf News Net, 2020) who play in a selection of 50 different events every year. Each event is played on a unique course and is made up of 4 rounds, each 18 holes long, played over a period of 4 days. In order to win a tournament or event, a golfer must finish their event with the lowest cumulative shots taken to complete the course.

The dataset being used in this paper contains performance-related statistics for all golfers on the PGA Tour from 2010 until 2018 using the Shotlink data made available by the PGA.

A prior knowledge of golf terminology is required to understand several of the statistics being recorded in this dataset. The most important of these as it relates to this study is the concept of Strokes Gained (SG). SG records a player's performance relative to the rest of the PGA Tour at each stage of play. There are four primary stages of play in golf:

- Off The Tee (OTT) - A golfer's first shot on a hole.
- On Approach (APR) – Usually a golfer's second shot from the fairway or rough towards the green.
- Around the Green (ARG) - A golfer's shots taken within 30 yards of the edge of the green
- Putting - A golfer's shots taken on the putting green.

For instance, if a golfer takes their first shot and the ball lands in a particularly favourable position compared with the same shots taken by other golfers, that player may gain half a stroke on that hole. Total Average SG is a combination of all four of these components. A high SG value is desirable for all players. A higher value is also desired for all other statistics apart from Average Putts and Average Score, where a low value is desirable, and Player Name and Year, as these are ID variables. All variables, excluding Player Name and Year, are continuous.

2. EXPLORATORY ANALYSIS

2.1. DATASET DESCRIPTION

The dataset being used in this regression analysis is a PGA Tour Shotlink Dataset which contains a record of 18 statistics for every PGA Tour golfer over nine seasons (2010 – 2018). The outcome variable in the case of this analysis is Player Wins.

2.2. DATA CLEANING AND FORMATTING

Firstly, Null values in the Wins and Top 10's columns were converted to 0's. Secondly, all rows with NA values were removed with the 'na.omit' function in R. This reduced the dataset from 2312 observations to 1678 observations. Thirdly, the Points and Money variables were formatted correctly as initially they were not formatted as integers. Due to the fact that Money, Points and Top 10's are directly associated with a player's Wins in a given season, these were removed from analysis. Both the Name and Year variables were removed, as they do not contribute to the outcome variable and act as ID columns. Additionally, the Total Average Strokes Gained and Average Score columns were removed. Total Average SG is essentially an aggregate of all other SG statistics and for the purpose of this analysis, emphasis was placed on examining how individual aspects of a golfer's game influence whether they will achieve a win. The same reasoning was applied to the removal of Average Score, as this statistic is inherently linked to Total Average SG. Hence, 11

columns were selected for Exploratory Analysis. Slight changes were made to the names of the columns to maintain a consistent format and for improved readability.

It is highly unlikely for a player to achieve more than one tournament win in a single PGA Tour Season. Of the 1678 remaining observations, 45 had more than a single win in a season. Hence, observations with more than 1 win were considered outliers and removed from the dataset. Subsequent to this removal of outliers, 1633 observations remained in the dataset.

Due to the fact that at this point there were two outcome variables, Win and No Win, with all predictor variables being continuous, it was decided that a binomial logistic regression model would be created. Following the creation of the below Correlation Heatmap in Figure 1, the Win variable was set as a factor with 2 levels: 0 (No Win) and 1 (Win), in order to maintain the correct format for a binomial logistic regression model.

2.3. CLEANED DATASET CHARACTERISTICS

In order to better understand the relationship between variables prior to beginning regression analysis, a correlation heatmap was created.

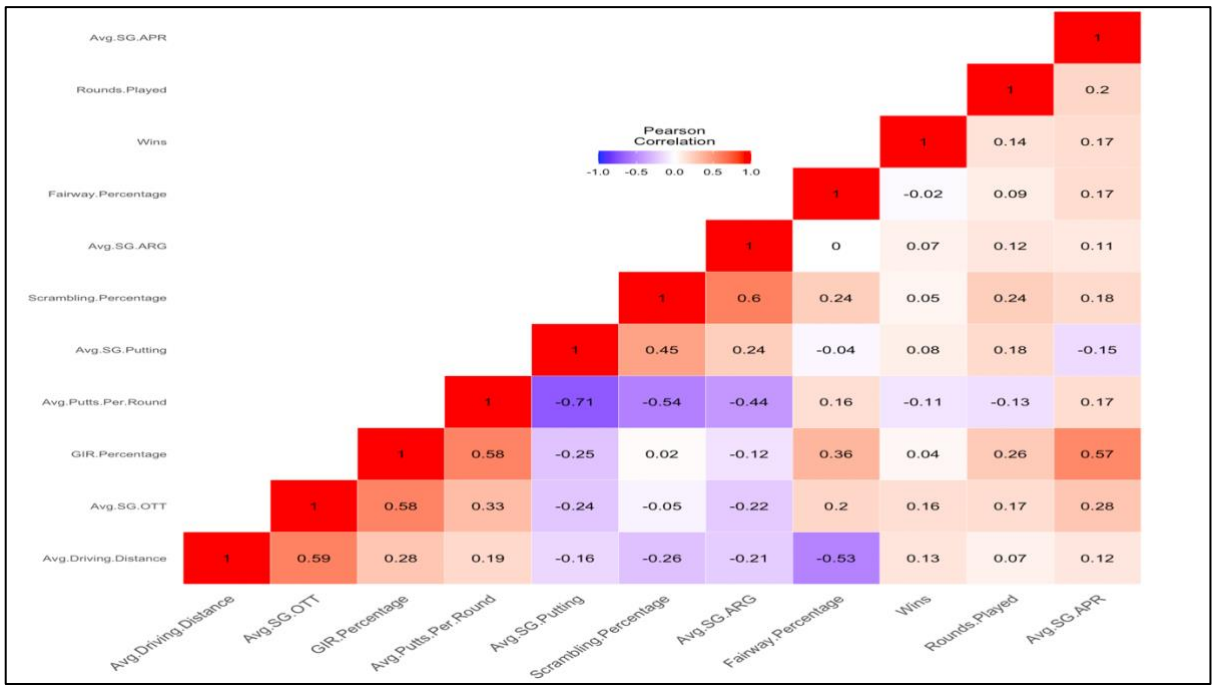

Figure 1: Correlation Heatmap

Unsurprisingly, no single variable had a significant impact on the value of Wins, which is evident in the fact that there were no Pearson coefficient values greater than 0.2. This was to be expected, as a golfer needs to blend different parts of their playing style to be successful. SG on Approach and SG Off the Tee had the highest positive correlation with

Wins, albeit by a small margin. Overall, there was little to substantively take away from the above correlations due to their relatively small magnitude.

Upon analysing the differences in characteristics of players with and without a win in terms of spread, results remained relatively consistent to those of the correlation heatmap. Rounds Played and Average Driving Distance varied in that players with a win displayed a particularly high value in both, which was not in agreement with their corresponding Pearson Coefficient values.

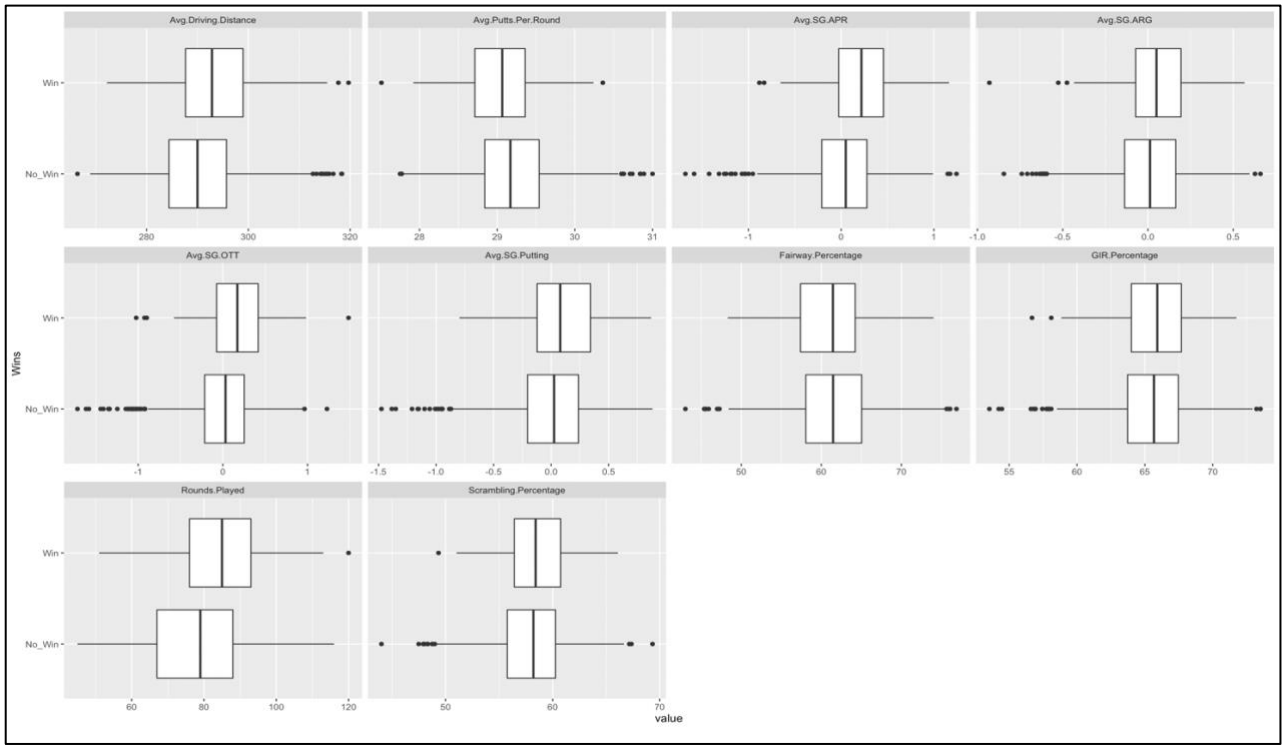

Figure 2: Boxplots by Wins

3. METHODS

3.1. VARIABLE SELECTION

Two methods of variable selection were employed to determine the optimal mixture of variables which would create the most effective model: the R 'regsubsets' function and the coefficients of the Least Absolute Shrinkage and Selection Operator (LASSO) model.

3.1.1. REGSUBSETS

The 'regsubsets' function in the 'leaps' package was used to identify the best models with different numbers of predictor variables. In the case of this analysis, an exhaustive search of the model space was performed to determine variables which contribute to the lowest

complexity of the final model in terms of its Bayesian Information Criterion (BIC). BIC indicates a model's cost-complexity trade-off. It was decided to incorporate the results of this search in the latter stage of analysis.

3.1.2. LASSO COEFFICIENTS

Lasso regression penalises the coefficients of variables with comparatively less contribution to the final model based on the sum of their absolute values, in a process commonly called regularisation. Through building a lasso model using all variables and examining which coefficients were converted to zero insight was gained into which variables were best to include.

3.2. BUILDING REGRESSION MODELS

Six regression models were built in total: two ridge regression models, two lasso regression models and two elastic net regression models using the 'glmnet' package in R . A 70%, 15%, 15% split was adopted for training, validation and test data respectively for all models.

3.2.1. RIDGE REGRESSION MODEL

Ridge Regression introduces a parameter 'lambda' to a standard regression model which penalises variables that are not contributing to the model or contribute very little by converting their coefficient to a value close to zero, based on the sum of the squared coefficients. Lambda was calculated using k-fold cross validation with the Area Under Curve figure selected as a type measure. AUC was chosen as two-class logistic regression is being performed.

3.2.2. LASSO REGRESSION MODEL

Similar to Lasso regression penalises variables that do not contribute to a model by reducing their value or by converting them to zero. A lasso model was built using the same training, validation and test datasets as above and lambda was tuned using k-fold cross validation with AUC chosen as the type measure.

3.2.3. ELASTIC NET REGRESSION MODEL

Elastic Net Regression blends both Lasso and Ridge Regression with a mixing parameter alpha that mitigates the extreme penalisation of variables. Where a Lasso model might reduce a variable to zero, an Elastic Net model may keep this variable and reduce it so that it is close to zero.

3.3. CHOICE OF THRESHOLDS – ROC CURVES

A Receiver Operating Characteristic Curve displays the performance of a binary classifier as its discrimination threshold changes. In the case of this analysis, a plot was created for Ridge, Lasso and Elastic Net models in order to determine the best threshold to use for each. With the aim of the model being to determine what aspects of player performance will result in a win, a true positive rate was prioritised without trying to minimise the false positive rate, as golf wins are inherently difficult to predict.

3.3.1. COORDS FUNCTION

The 'coords' function in R as part of the ROCR package returns a point in a ROC curve corresponding with the optimal threshold. Two methods were chosen to determine the best thresholds: Youden's J statistic and the 'closest top left' point. Using the youden method, the optimal threshold yields the maximum sum of sensitivities and specificities (true positive observations and true negative observations). Closest top left chooses the threshold located closest to the top left position of the ROC curve yielding perfect specificities and sensitivities.

These thresholds were then validated by plotting a graph of the accuracy of each model based on their thresholds or 'cut-offs' and identifying the 'kink' or elbow of the plot where accuracy seems to plateau.

3.4. ASSESSMENT OF MODELS

To assess model performance, the Area Under Curve (AUC) value for each model's corresponding ROC curve was used along with each model's Akaike Information Criterion (AIC) and accuracy.

4. RESULTS

4.1. VARIABLE SELECTION

Regsubsets output indicated that in the lowest cost-complexity trade off model Fairway Percentage, Driving Distance, GIR Percentage, SG Putting and SG Around the Green should be removed. However, the removal of all five of these variables was considered excessive. Fairway Percentage and Driving Distance were not included in any of the optimal models in this output, hence these were removed after the initial Ridge, LASSO and Elastic Net models were created.

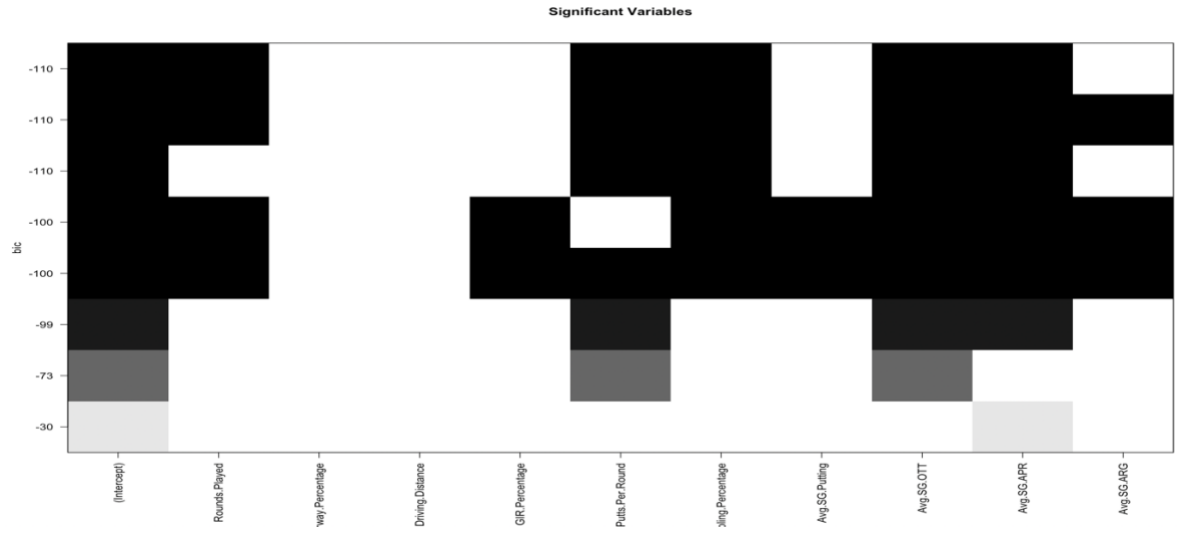

Figure 3: RegSubsets Results

Upon completing the initial LASSO model with all variables included, Fairway Percentage and Driving Distance again did not appear in the final model. This reinforced the results of the search of the model space.

	Ridge	Lasso	Elastic Net
(Intercept)	6.755761360	23.60453741	25.649127749
Rounds.Played	0.016634768	0.02442496	0.027547271
Fairway.Percentage	0.001852144	.	0.001966177
Avg.Driving.Distance	0.013985083	.	.
GIR.Percentage	-0.029278284	-0.10026541	-0.141131890
Avg.Putts.Per.Round	-0.358305678	-0.55728941	-0.493324708
Scrambling.Percentage	-0.031523105	-0.08324777	-0.111996179
Avg.SG.Putting	0.260817572	0.34480104	0.568896635
Avg.SG.OTT	0.723813318	1.51244949	1.717146834
Avg.SG.APR	0.824736653	1.45455030	1.658005306
Avg.SG.ARG	0.505528374	0.81056901	1.148053714

Figure 4: Model Coefficients

4.2. THRESHOLDS

Through examination of the ROC plot shown in Figure 5, the optimum threshold for the LASSO model was chosen as 0.19. Both the youden and closest left methods yielded a threshold of 0.1923. Upon examining the accuracy plot for each model's ROC plot, it was decided that the closest top left thresholds would be used. For the Ridge and Elastic Net models, these thresholds were 0.1618 and 0.1565 respectively.

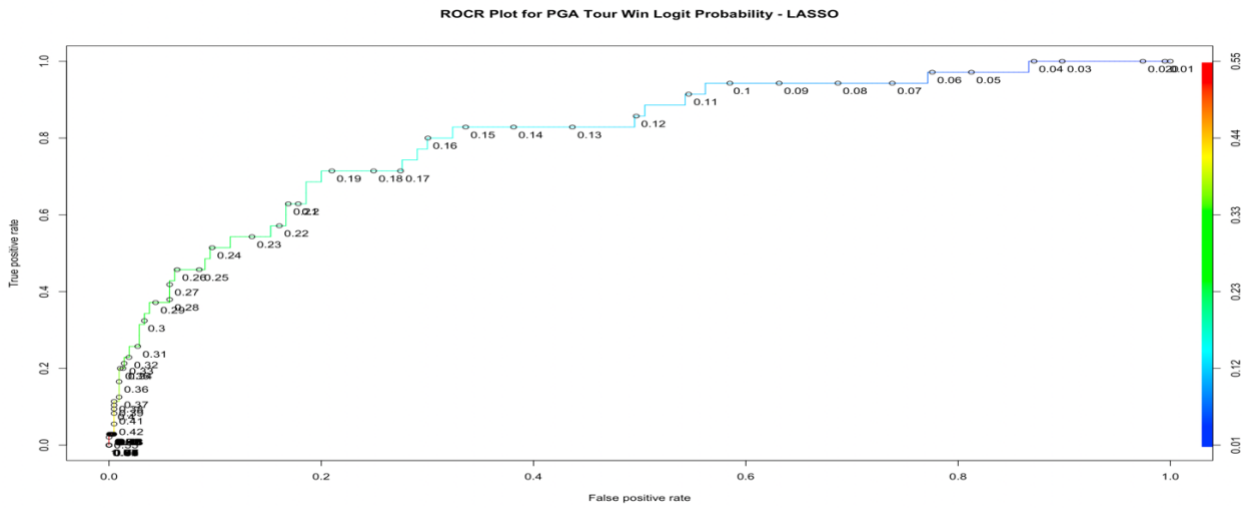

Figure 5: ROC Plot for LASSO Model

4.3. MODEL PERFORMANCE

All of the initial models achieved an AUC of over 0.8, with the highest being that of the Ridge Model with an AUC of 0.8186. This result supported the viability of all three models with all variables included. Each model achieved a negative AIC. The Ridge Model's AIC was -68.17, LASSO achieved an AIC of -95.97 and the Elastic Net model's AIC was -98.77. This result indicated that the LASSO and Elastic Net models were favourable over the Ridge Model in terms of model complexity. In terms of accuracy, each model had a similar performance. Elastic Net had the lowest accuracy with 68.16% of classifications correctly predicted, Ridge displayed an accuracy of 69.8% and LASSO performed the best with 74.69% accuracy. Due to LASSO displaying a relatively high AIC value, a low AIC value and the highest accuracy, it was decided on as the best model to predict golfer's PGA Tour Wins.

Upon removing both the Driving Distance and Fairway Percentage variables, the same model creation process was repeated and a Ridge, LASSO and Elastic Net model were built. However, these three models all performed worse than when all variables were included in terms of AUC, with each model's AUC decreasing to below 0.8. The accuracy of Elastic Net increased to 70.2%, however both Ridge and LASSO models' accuracy decreased to 66.53% and 71.84% respectively. Additionally, the AIC of all models increased. Hence, none of these models were favourable to the initial models with all variables.

As is apparent in Figure 4 above, high Average SG OTT and Average SG APR values seem to be the best predictors of success for PGA Golfers. Variables such as Rounds Played and Scrambling Percentage were surprisingly minimal in their influence on the final model. However, the value for Intercept was notably much higher than all other variables, which means that the true effects of these variables individually may in fact be minimal.

5. DISCUSSION

Results from this analysis agree with the assessments of previous research, in that golfer performance is highly chaotic, and identifying individual aspects of their game which most contribute to winning on the PGA Tour is inherently difficult. However, the models created displayed reasonable performance given the complexity of the task. Key insights gained include the indication that the first two shots on a hole represent the most pivotal part of a golfer's game according to the created model. Average SG OTT, Average SG APR, Average SG ARG and Average Putts taken appear to be the statistics that may be worth putting more time into by golfers hoping to improve their chances of winning as these were the most significant variables in the final model.

Due to timing constraints, there were certain avenues of research which could not be pursued and will hopefully be pursued in the future. These include the use of a Precision-Recall Curve Plot (Brownlee, 2020) in place of a ROC plot, as the data being used in this analysis is highly imbalanced, as very few golfers achieve a win in a PGA Tour Season. An expansion of the original dataset with data from more PGA seasons would also undoubtedly add to the reliability and accuracy of the models created.

BIBLIOGRAPHY

Arastey, G. M., 2020. *SPORT PERFORMANCE ANALYSIS.* [Online]
Available at: https://www.sportperformanceanalysis.com/article/increasing-presence-of-data-analytics-in-golf
[Accessed May 2021].

Brownlee, J., 2020. *Machine Learning Mastery.* [Online]
Available at: https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/
[Accessed May 2021].

Evans, K. & Tuttle, N., 2015. Improving performance in golf: current research and implications from a clinical perspective. *Brazilian Journal of Physical Therapy,* 19(5), pp. 381-389.

Golf News Net, 2020. *GNN.* [Online]
Available at: https://thegolfnewsnet.com/golfnewsnetteam/2020/08/16/the-125-pga-tour-players-who-got-their-2020-21-cards-qualified-for-fedex-cup-playoffs-120122/
[Accessed May 2021].

James, N., 2007. The Statistical Analysis of Golf Performance. *International Journal of Sports Science and Coaching ,* 2(1).

Myers, J. et al., 2008. The role of upper torso and pelvis rotation in driving performance during the golf swing. *Journal of Sports Science,* 26(2), pp. 181-188.

Parker, J., Hellstrom, J. & Olson, M. C., 2019. Differences in kinematics and driver performance in elite female and male golfers. *Sports Biomechanics.*

ShotLink, 2021. *shotlink.com.* [Online]
Available at:
http://www.shotlink.com/about/history#:~:text=The%20TOUR%20implemented%20its%20first,device%20and%20two%20mini%2Dcomputers.&text=This%20system%20would%20become%20known%20as%20ShotLink.
[Accessed May 2021].

Stockl, M. & Lamb, P. F., 2018. The variable and chaotic nature of professional golf performance. *Journal of Sports Sciences,* 36(9).