Final Project: Multivariate Analysis of Performance in Golf

STAT 497 – Sports Analytics

Dr. Joshua Wyatt Smith

Nikolas Argiropoulos 40044358

Concordia University

1. Introduction

Over the past few decades, data driven decisions have extended beyond the medical and financial world. Most recently, the world of sports has benefits from the innovative developments (Parra et al., 2022). For example, the use of data and mathematical modeling to determine the best possible outcome of a match or strategy, and player performance is a fairly novel application but has yielded fruitful insights (Parra et al., 2022). Organizations like the National Football League (NFL) and Major League Baseball (MLB) have adapted various machine learning techniques and player-based statistics to quantify the players worth and skill level (Parra et al., 2022). Through these models, coaches can aid players in improving various weakness or construct a trade to improve a particular position within the league. Unlike, the NFL and MLB, however, the professional golf association (PGA) has trailed in funding and data collection efforts, which negatively affects their ability to produce high quality and complex solutions for their players. It is only within the last decade that golf adapted training aids like TrackMan or GCQuad that measure various swing indicators like *Smash Factor*, *Spin Rate* and *Angle of Decent* with great accuracy[1]. Furthermore, the complexity of performance analysis is limited to summations, averages, and proportions, like driving accuracy, average carry distance, and greens in regulation (GIR)[1]. Overall, the lack of data guidance within golf is severely lacking.

The work aims to improve data drive practice within golf by using event data gathered between the 2015 and 2019 PGA tour seasons. The goal of this work would be to conduct a performance analysis on the top three professional golfers. More specifically, a multivariate multiple linear regression model will be used to predict three key performance indicators (KPI) and partial η^2 will be used to quantify the proportion of variance accounted for by each regressor in each of the KPIs and the total model. Further, model accuracy will be determined by adj.R^2, and mean absolute percentage error (MAPE)[2] ($MAPE_{i,j,k} = \left(\frac{1}{n}\sum_1^n \frac{|A_{i,j,k}-P_{i,j,k}|}{|A_{i,j,k}|}\right) * 100$, where $MAPE_{i,j,k}$ is the mean absolute percentage error value for the i,j,k KPI and A_{ijk} is the actual value for the i,j,k KPI and P_{ijk} is the predicted value given by the model.

A performance analysis is characterized by defining and analysing how an outcome or metric (e.g., KPI's) are attained (Hughes & Bartlett, 2002). Traditionally, performance indicators were defined and classified based of the shot's distance to the hole, this includes carry distance (Stöckl, Lamb & Lames, 2012). Consequently, these indicators ignore other important and contributing factors/indicators that can lead to a victory. As Arnold Palmer once stated, "Golf is deceptively simple and endlessly complicated […]" due to the endlessly complicated nature of golf, a univariate analysis cannot detect the unique variability of each performance indicator within a multivariate framework. The ability for players to quantify the amount of variability that driving distance compared to driving accuracy is accounted for in GIR, enables the professional to strategize distance over accuracy or vise-versa, possibly leading to lower scores. In addition, since golf is comprised of multiple shots, in which each shot is more difficult and complex than the next due to angle change, terrain factors, etc., a multivariate analysis is needed to account for the variable nature of golf.

2. Literature review

As a golfer, one must incorporate various elements of their environment, skill level and angle to green on every shot. The body of research currently focus on descriptive and univariate analysis of specific aspects of golf. Broadie (2012) conducted a study assessing professional golfers' performance based on strokes gained. The analysis partitioned stokes gained into three main categories: long game (which is classified as shots greater than 100 yards from the pin), short game (shots within 100 yards from the pin) and putting (Broadie, 2012). The results suggest that if the distance to the hole is less than 15 yards or greater than 34 yards, sand shots have larger average stroke than shots from the rough. Broadie (2012) further suggests that one-putts occur 50% of the time within 8-feet and the three-putt probability follows an exponential curve, suggesting that the further the putt, the more likely to three-putt. One limitation of this work, however, is that the analysis is merely descriptive and not predictive in nature. Chamberlain (2017) examined key components of professional golfer games to predict which aspects and skill are needed to succeed on the PGA tour. Using a K Nearest Neighbours (KNN) classifier and a multiple

linear regression (MLR), the author was able to predict 76% of golfers who placed between 31st and last place. Additionally, the MLR model predicted that one golfer would finish 19th, he placed 22nd (Chamberlain, 2017). Chamberlain's (2017) work introduced machine learning and predictive modeling, albeit in univariate form. There is a clear gap in the literature where multivariate analyses are needed. This project aims to add to the current body of literature by implementing a multivariate predictive model and measure the unique variation of each regressor on each KPI to suggest skill and game improvements. Furthermore, a descriptive comparison between the top 3 PGA Tour players and the field will be completed to suggest possible skill improvement for the average PGA Tour player.

3. Methodology

3.1. Data collection

Data for this project was obtained from the PGA Tour website (n.d) and Choi (2020) Kaggle account. The combination of both datasets totals to greater than 1000 metrics varying from off the tee metrics like driving distance, average score on a par 3, 4 and 5, to earning/point totals. The PGA tour website offers data for both the season and per event totals/averages, to increase the power and sample size for the analysis, yearly data will be used. Data from the PGA tour website will be downloaded and imported to Excel and combined with Choi's data. Once the variables of interest are collected into a CSV file, the file will be imported to R-Studio (version 2022.07.2+576) for data wrangling, descriptive, and predictive analyses. Lastly, the analyses will include data from all eligible professional golfers between the years of 2015 and 2019.

3.2. Variables of interest

There are eight variables of interest that are included in the descriptive and inferential analysis. The variables are listed and defined as follows: *greens in regulation (GIR*; the percentage of green's hit that may lead to a birdie or eagle within a season), *bogey average (BA*; the average number of bogey's per round in a season), *scoring average (SA*; the total strokes gained divided by the total rounds played), *driving distance (DD*; the approximate summation of carry and roll distance from the tee-box), *driving accuracy (DA;* the amount of fairways hit divided by the total amount of fairways expressed as a percentage), *approaches from less than 100 yards (APLH;* the average distance between the ball and the hole (in feet) when a player's approach is hit from a distance less than 100 yards), *approaches from greater than 100 yards (APGH*; the average distance between the ball and the hole (in feet) when a player's approach is hit from a distance greater than 100 yards, lastly), and *top 10 finishes* (the summation of all the events a player ranks within the top 10 in a given tournament within a season). The first seven variables are continuous, and last variable is ordinal[1]. (Choi, 2020):

3.3. Data analysis

Prior to descriptive and inferential analysis, data cleaning, normality and assumptions checks will be conducted. If the data is non-normally distributed a logarithm or square-root transform will be completed. Two-tailed significance of 0.05 was adopted for all the statistical tests. During the data cleaning process, PGA professionals with greater than six missing values among the variables of interest were removed. Mean imputation, was used to fill in the remaining two missing data points for a complete dataset. Descriptive statistics will be conducted on all variables. A correlation matrix will measure the association between the variables of interest to measure multicollinearity. The data will have a 70/30 split between training and test, to reduce bias within the model. The construction of the multivariate model will be composed of DD, DA, APLH and APGH regressed on GIR, BA and SA. The multivariate multiple regression formula is calculated as follows $y_{i,j,k} = \beta_{1,i,j,k}DD_1 + \beta_{2,i,j,k}DA_2 + \beta_{3,i,j,k}APLH_3 + \beta_{4,i,j,k}APGH_4$, where $i = GIR$, $j = BA$ *and* $k = SA$. Lastly, partial η^2 of the regression model will quantify the variation accounted for by each regressor on the multivariate dependent variable and will quantify the variation accounted for by each regressor on each unique dependent variable. Results of the partial η^2 can suggest as to which

[1]Pgatour.com. (n.d.). *Golf stat and records: PGA tour*. PGATour. Retrieved October 21, 2022, from https://www.pgatour.com/stats.html

[2] *How to measure the accuracy of predictive models*. Acheron Analytics. (2018, May 28). Retrieved November 20, 2022, from http://www.acheronanalytics.com/acheron-blog/how-to-measure-the-accuracy-of-predictive-models

skills PGA tour professionals should improve to lower their bogey and scoring average, and increase GIR.

4. Results

Prior to data cleaning, 1367 observations and 30 variables were included in the Kaggle dataset (Choi, 2020). Post data cleaning, only 10 variables where kept. Additionally, 31% (427/1367) of PGA professionals had greater than six missing values, thus they were removed from the analysis. Overall, the total sample included 940 complete observations. Mean imputation was completed on two variables for two separate tour players. Residual analysis suggests that most variables of interest were not normally distributed (Shapiro–Wilk Test, $p < 0.05$). However, further analysis of the histograms and QQ-plots suggest that the residuals of all continuous variables were normally distributed. Thus, due to the large sample size and the results from the histograms and QQ-plots, the normality of residual assumption for regression passed. Further, the scatterplots suggest the variables are linearly related and the variance of residual was the same for any value of X variable, thus the assumptions of linearity and homoscedasticity pass. The average (SD) driving distance and driving accuracy percentage for PGA professional between 2015 and 2019 was 293.10(9.07), and 61.37%(5.14), respectively. The mean distance between the ball and the hole for approach shot greater than and less than 100 yards to the centre of the green were 32.65ft(1.62) and 17.14ft(1.93). Furthermore, tour professionals successfully land on average 76.85%(2.84) of greens in regulation and their scoring and bogey average are 70.95(0.71) and 2.61(0.26), respectively.

A multivariate multiple linear regression analysis was performed to examine whether DD, DA, APLH, and APGH significantly predicted GIR, BA and SA. The results suggest that the three regression models were significant [GIR: $F(4,653)= 123.10$, $p < 0.001$, adj-$R^2 = 0.417$; BA: $F(4,653) = 43.71$, $p < 0.001$, adj-$R^2 = 0.222$; and SA $F(4,653) = 88.99$, $p < 0.001$, adj-$R^2 = 0.345$]. Further examination of each model suggests that the four regressors (DD, DA, APLH, and APGH) significantly predicts each regressand ($p < 0.001$). See Table 1 for the partial η^2 for each covariate within the multivariate model and for each individual model, and Table 2 for model accuracy. When descriptively comparing the data from the top three tour professions to the tour average we see that Rory McIlroy and Dustin Johnson are above average in driving distance, GIR, and top 10 finishes. Whereas Rory McIlroy, Dustin Johnson and Jordan Spieth are well below average in approaches from less than and greater than 100 yards, bogey, and scoring average (Figure 1).

5. Discussion

The study found that DD, DA, APLH, and APGH significantly predict GIR, BA and SA. More specifically, a single unit increase in driving distance resulted in a 0.197 increase in GIR, and a 0.009 and 0.044 decrease in bogey and scoring average, respectfully. This result suggests that as a tour professional increase their driving distance they are more likely to land on a green, due to the shorter distance between their second shot and the green, resulting in a lower bogey and scoring average. Likewise, driving accuracy was found to increase GIR by a factor of 0.078 and decrease bogey and scoring average by a factor of 0.021 and 0.053, respectfully. These results suggest that accurately placing the ball in the fairway, away from hazards like water and sand, improves the chances of a tour pro to land on the green, resulting in a lower bogey and scoring average. This result parallels the study completed by Broadie (2012). Furthermore, data suggests that as the distance between the ball and the hole increases from approach shots from less or greater than 100 results in an increase in bogey and scoring average by a factor of 0.022 and 0.072, respectively. Broadie (2012) found similar results, suggesting that as you increase the distance between the ball and hole, you are less likely to make it, resulting in high scores.

Driving distance and driving accuracy accounted for the majority of the variation in the total model, 41.5% and 21.2%, respectively. This suggests that if tour players should focus on improving driving distance and driving accuracy

[1]Pgatour.com. (n.d.). *Golf stat and records: PGA tour*. PGATour. Retrieved October 21, 2022, from https://www.pgatour.com/stats.html

[2] *How to measure the accuracy of predictive models*. Acheron Analytics. (2018, May 28). Retrieved November 20, 2022, from http://www.acheronanalytics.com/acheron-blog/how-to-measure-the-accuracy-of-predictive-models

to improve in all three metrics simultaneously. More specifically, if a tour professional wishes to increase their greens in regulation, they should improve their driving distance and approaches shot greater than 100 yards since these variables account for 29.9% and 8.9% of the total variation in GIR. These results, parallel the top 3 player on tour. Examining Figure 1, Mr. McIlroy and Mr. Johnson outperform the tour average on driving distance and approaches shot greater than 100 yards. One can infer that these qualities supported their winnings. With respect to lowering bogey average, driving distance and driving accuracy account for 7.9% and 15.1%, respectively, of the total variation. Likewise, driving distance and driving accuracy account for 26.8% and 18.2%, respectively, of the total variation in lowering scoring average. Thus, tour players should focus on improving their driving distance and driving accuracy to lower their bogey and scoring average. Lastly, the model accuracy results suggest less than 1% error in predicting future GIR, BA and SA scores. The adj-R2 for GIR, BA, and SA were 41.71%, 22.20%, and 34.51%. These adj-R2 range between weak to median fit, this may be due to the lack of covariates within the model.

6. Conclusion

The data suggests that if tour professional wish to increase their GIR, they should focus on increasing their driving distance and reducing the distance between the ball and hole from approaches shot greater than 100 yards as these variables account for the largest variation in GIR. If tour professionals wish to lower their bogey and scoring average, they should increase the driving distance and driving accuracy.

7. Limitation and Future Research

Due to only event data being available, the study is limited to proportions, averages and percentages. Tracking data will enable future researcher to study classification models on which club or strategy best increase the probability of success with less bias than event data. Furthermore, the current study does not measure any aspects of putting and scrambling. A future study incorporating these metrics can produce a more encompassing result. Lastly, the current model fit (adj-R2) ranges between 22% and 42%, although this is not a great fit, future studies can implement more covariates to account for a larger total variation.

Table 1.
Variation accounted for by each regressor within the total and individual models

	Total Model	*Greens in regulation*	*Bogey Average*	*Scoring Average*
Covariates	*Partial η^2*	*Partial η^2*	*Partial η^2*	*Partial η^2*
1. Driving Distance	0.415	0.299	0.079	0.268
2. Driving Accuracy %	0.212	0.015	0.151	0.182
3. Approaches < 100yds	0.070	0.020	0.031	0.043
4. Approaches > 100yds	0.111	0.089	0.017	0.037

For a percentage, multiply the value by 100.

Table 2.
Model Accuracy

Model	*MAPE%*	*Adj-R^2*
1. Greens in Regulation	0.11%	41.71%
2. Bogey Average	0.81%	22.20%
3. Scoring Average	0.03%	34.51%

[1]Pgatour.com. (n.d.). *Golf stat and records: PGA tour*. PGATour. Retrieved October 21, 2022, from https://www.pgatour.com/stats.html

[2] *How to measure the accuracy of predictive models*. Acheron Analytics. (2018, May 28). Retrieved November 20, 2022, from http://www.acheronanalytics.com/acheron-blog/how-to-measure-the-accuracy-of-predictive-models

Figure 1.

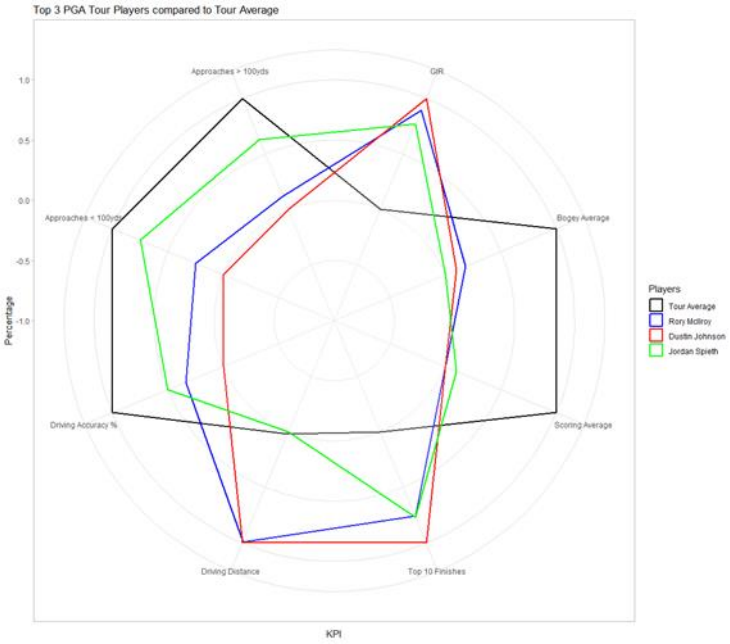
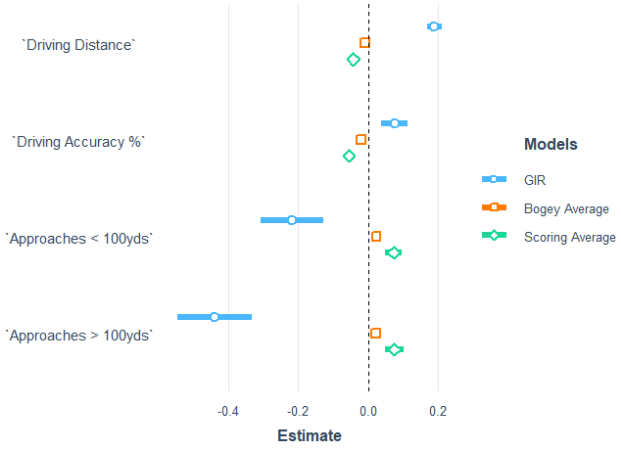


Figure 2.

Beta estimates for each regression in each model



References

Alboukadel. (2020, December 12). *Beautiful radar chart in R using FMSB and GGPlot packages*. Datanovia. Retrieved November 20, 2022, from https://www.datanovia.com/en/blog/beautiful-radar-chart-in-r-using-fmsb-and-ggplot-packages/

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

Chamberlain, Luke (2017) *Assessing Golfer Performance on the PGA Tour.* Undergraduate thesis, Dublin, National College of Ireland.

Choi, J. (2020, April 16). *PGA Tour top 200 player data (2015-2019)*. Kaggle. Retrieved October 21, 2022, from https://www.kaggle.com/datasets/jychoi87/pga-tour-top-200-player-data-20152019

Ford, C. (2017, October 27). *University of Virginia Library Research Data Services + Sciences*. Research Data Services + Sciences. Retrieved November 20, 2022, from https://data.library.virginia.edu/getting-started-with-multivariate-multiple-regression/

Hughes, M.D., & Bartlett, R.M. (2002). The use of performance indicators in performance analysis. Journal of Sports Sciences, 20, 739–754. PubMed doi:10.1080/026404102320675602

Long, J. (2022, April 25). *Tools for summarizing and visualizing regression models*. Tools for summarizing and visualizing Regression Models. Retrieved November 20, 2022, from https://cran.r-project.org/web/packages/jtools/vignettes/summ.html#effect_plot()

Parra, X., Tort-Martorell, X., Alvarez-Gomez, F., & Ruiz-Viñals, C. (2022). Chronological evolution of the information-driven decision-making process (1950–2020). *Journal of the Knowledge Economy*. https://doi.org/10.1007/s13132-022-00917-y

Stöckl, M., Lamb, P. F., & Lames, M. (2012). A model for visualizing difficulty in golf and subsequent performance rankings on the PGA Tour. *International Journal of Golf Science*, *1*(1), 10–24. https://doi.org/10.1123/ijgs.1.1.10