Fantasy Golf Using a Data Driven Team Selection Algorithm

Shane Rooney, B.Sc. Computer Applications

A practicum submitted to University College Dublin in part fulfilment of the requirements of the degree of M.Sc. in Business Analytics

Michael Smurfit Graduate School of Business

University College Dublin

September 2016

Supervisor: Mr Matthias Glowatz, UCD

Head of School: Professor Ciarán Ó'hÓgartaigh

Table of Contents

Table of Contents

List of Figures

List of Tables

List of Equations

List of Important Abbreviations

ANOVA – Analysis of Variance

BIP – Binary Integer Programming

C.I. – Confidence Interval

DFS – Daily Fantasy Sports

FG – Fantasy Golf

FSTA – Fantasy Sports Trade Association

GA – Genetic Algorithm

GIR – Greens in Regulation

NSGA-II – Non-dominating Sorting Genetic Algorithm

PGA – Professional Golfers' Association

SG – Strokes Gained

T2G – Tee-to-Green

YTD – Year-to-Date

Acknowledgements

I would like to extend my heartfelt thanks and gratitude to my academic supervisor Mr Matt Glowatz for his support and guidance. My thanks go to the PGA for the ShotLink data.

Abstract

Fantasy sports is a growing industry, with over $1 billion dollars in prize funds given to participants in 2015. Fantasy golf is a rapidly growing subset of this industry with $30 million in prize funds shared in 2015. Playing fantasy golf professionally is a feasible ambition if one is armed with the right data and the business analytics tools and skills needed to find actionable insights. There is more rich golf data than ever before but there is a lack of understanding to what constitutes a player who is currently on a good run of form. This paper analyses ShotLink data to find explanatory statistics and develop a form indicator. This form indicator is then used as input to an algorithm that selects a team of the most in-form golfers that fit into the defined budget. Both genetic algorithms and binary integer programming solutions were tested with the binary integer programming solution giving the best results. The data driven teams selected in this manner outperform human generated teams. This system can be used to play season-long competitions or used to enter daily fantasy sports leagues where participants finishing in the top 50 percentile of the league collect winnings.

1. Executive Summary

1.1. Fantasy Sports Industry

According to the fantasy sports trade association (FSTA), fantasy sports participants spent over $2 billion in 2015 in entry fees with projected spending to reach $14.4 billion by 2020. Over $1 billion in prize funds were shared by fantasy sports participants in 2015 with $30 million of this shared amongst fantasy golf players. Daily fantasy golf introduces a format of the game where users create and enter a new fantasy team each week with a chance to collect winnings based on that team finishing in the top 50 percentile of participants. This growth in fantasy golf makes playing professionally a viable option.

1.2. Business Opportunity in Fantasy Golf

More and more businesses are looking to gain a competitive edge using actionable insights and predictive patterns found in data. An increasing number of business decisions are being driven by data. If fantasy golf is to be pursued as a business, selecting one's team should also be a data driven decision powered by the wealth of golf data that is now available. New statistics like strokes gained putting and stokes gained tee-to-green have given us a deeper and more accurate understanding of golfer performance.

In this paper, these statistics along with recent tournament history are analysed and interpreted as player form and used to select a winning fantasy golf team. A form indicator is established by applying data mining tools and techniques to golf's ShotLink data. Form is based on YTD data and course history form. A data driven algorithm is then used to automatically select a team of the most in form players that will fit into the defined budget.

1.3. Results and Implication

The most important features driving golfer form are:

- Strokes Gained Tee-to-Green
- Strokes Gained Putting
- Birdie Bogey Ratio
- YTD Average Score
- YTD Average Finish Position
- YTD Fantasy Golf Points
- Course History

Strokes gained is the most predictive indicator for the Majors (the four most important events) but further statistical measures are needed to separate golfers in the regular tour events. These features are combined to create a form indicator which is used as input to an algorithm that automatically selects a fantasy golf team. The teams produced by the algorithm have been shown to outperform the average human competitor in each of the tournaments tested.

Executive Summary

This approach can be used to select a fantasy golf team based on statistical analysis as opposed to subjectively choosing a team based on favourites and familiarity bias.

The form indicators and algorithm outlined in this paper should be used to select teams for the upcoming 2016/2017 season or used for entering teams into upcoming daily fantasy sports competitions. There is significant evidence to suggest fantasy teams selected in this manner could finish in the top 50 percentile of DFS leagues and thus collect winnings.

2. Introduction

2.1. Opening Remarks

Within the realm of business analytics, a growing area of interest and study is sports analytics. Likewise, the fantasy sports industry is growing at an extremely fast pace with participants spending over $2 billion in 2015 and this is projected to grow to $14.4 billion by 2020[1]. Within fantasy sports, daily fantasy sports (DFS) is the catalyst behind this monumental growth. DFS shortens the traditional fantasy sports season to single weeks or single tournaments with new fantasy teams entered every week. In DFS, the participant selects a virtual team based on real-life players in a sport and scores points based on the real players' performance. The better the real players play, the better the fantasy team performs and the more money the participant collects. In daily fantasy golf the participant picks a new team of golfers every week and can collect up to $1 million. In 2015 alone, $30 million was shared in prize funds on a single fantasy golf website[2]. Thanks to the relatively new, rich source of golf data produced by the Professional Golfers' Association (PGA) it is possible to analyse golfer performance in detail. This data can be used to make a data-driven fantasy golf team selection which can then be entered into a fantasy golf game.

2.2. Business Justification

Fantasy sports revenue is growing rapidly as are the prize funds. With bigger prize funds comes more interesting business opportunities and the possibility of playing fantasy sports professionally. Fantasy golf is based on selecting a team of golfers within a specific budget with the aim of scoring as many points as possible. Points are scored based on the actual golfers' performances on the course. In daily fantasy golf, a new team is selected for each tournament and points are scored based on the golfers' performances for that particular tournament. In daily fantasy golf, prize money is collected if the participant's team scores enough points to finish in the top 50 percentile of all participants who submit fantasy golf teams.

The goal of business analytics is to make data driven decisions based on the discovery of patterns in data through machine learning and statistical analysis. Producing a winning fantasy golf team should be a data driven decision based on the discovery of patterns in golf data through machine learning and statistical analysis.

Golf is one of the less mature applications of sports analytics with top class data gathering and statistical analysis very much a recent trend in the sport. Broadie (2008), (2010) led the way with his ground-breaking papers, which not only analysed golfers' performance but also brought new statistical measures to golf

[1] Forbes on Fantasy Sports, http://www.forbes.com/sites/darrenheitner/2015/09/16/the-hyper-growth-of-daily-fantasy-sports-is-going-to-change-our-culture-and-our-laws

[2] DraftKings fantasy golf, https://www.draftkings.com/fantasy-golf

which are widely used today by analysts and golfers alike. The statistic in question is strokes gained which can be simplified to the difference between the expected number of strokes needed to complete a hole versus the actual number of strokes taken.

This data has yet to be used for form interpretation, i.e. who is currently playing well this season or who should do well in the next tournament. This paper will use this new golf statistic as well as previous analysis from (Peters, 2008) and (Callan & Thomas, 2007) who argue measures like driving accuracy, greens in regulation and putting should be used to analyse golfers' performance. Shmanske (2008) uses similar statistics but examines the idea of adjusting scores based on the difficulty of the course. This type of analysis will form the basis of the form indicator developed in this paper.

Once one or more form indicators have been defined these can be used as input or objective functions in an algorithm that will automatically select a fantasy golf team. Genetic algorithms have been used in previous research (Ahmed, Jindal, & Deb, 2011), (Sarda & Sakaria, 2015) to solve the constrained multi-objective optimisation problem of selecting a fantasy sports team but there is no research in selecting a fantasy golf team as of yet. For a constrained single objective, binary integer programming can be used to find the optimum.

2.3. Aim

The aim of this paper is to automatically produce a fantasy golf team using a data driven algorithm. The data driving the algorithm is golfer form, which is deduced from YTD averages of golfer statistics.

2.4. Practical Contribution

This practicum uses the PGA ShotLink dataset to develop YTD moving averages of golfer performance statistics, and course history values based on performance in the same tournament over the previous 5 years. This data is then combined into a form indicator which can be used as input into a binary integer programming (BIP) solution. The BIP algorithm produces a fantasy golf team by maximising form while staying under the allocated budget for the team. A genetic algorithm (GA) is also tested but the BIP produces the best solution. A new fantasy team is produced for each tournament by the BIP and the data driven fantasy team outscores the mean score of human generated teams in each of the 4 tested tournaments. This is extremely promising because fantasy golf teams that finish in the top 50 percentile collect prize money in DFS competitions.

2.5. Document Structure

This paper is organised as follows, Section 3 summarises the development of golf analytics and fantasy sports, as well as previous applications of adaptive learning techniques and BIP. Section 4 describes the methodology used during the data mining process when analysing the golf data. It also describes the experiment settings and choice of objective function(s) for the BIP and GA. Section 5 highlights the best performing features and the form indicator that results. It also delivers the statistical analysis of both the BIP solution and the GA solution. Section 6 discusses some of the findings and what the results mean. Finally, in Section 7 conclusions are made along with future recommendations.

3. Literature Review

3.1. Fantasy Golf

Fantasy sport is a type of online game where users create or draft virtual teams of players from real professional sports. The virtual teams score points based on the performance of the real players and the better the real players plays, the more points the fantasy team scores. Fantasy players can be bought or sold depending on the rules of the fantasy game. Players of fantasy golf pick a team of golfers and these golfers score based on their performance in that week's tournament. Golfers do not enter tournaments every week so the field changes regularly, which means fantasy golf players may change their entire team each week.

Fantasy sports have been around since World War II with one of the earliest recorded examples attributed to Wilfred "Bill" Winkenbach who invented fantasy golf in the 1950's. Fantasy golf (Esser, 1994) has been described as a team of professional golfers that is drafted by participants on a weekly basis and the team with the overall lowest score wins. Winkenbach expanded his idea to American Football and added more rules and ways to score points. E.g. players could score points for a touchdown, a catch, a field goal or an interception. Other sports took note and in the 1980's USA today developed a special paper, Baseball Weekly that contained mainly statistics that could be used in fantasy sports. Regular fans were now starting to use statistics to increase their knowledge and engagement in sports and further publications were released in late 80's and early 90s with participation growing to 1-3 million users[3].

Internet development in the mid 90's led to the initial boom in fantasy sports and when the fantasy sports website www.commissioner.com sold for $31 million, fantasy sports was on its way to becoming big business. More and more fantasy sports sites came on-stream as statistics and up to date news became more accessible. Participation grew and in 1998 the Fantasy Sports Trade Association[3] was formed.

Daily fantasy sport (DFS) is a condensed version of the fantasy sports season and instead of playing over a full season, new competitions can be entered every week or every day where participants create a team to play head-to-head with another participant or enter a cash league where finishing in the top 50 percentile equates to collecting winnings. DFS has accelerated the growth of fantasy sports and the

[3] Fantasy Sports Trade Association, http://http://fsta.org/

FSTA project growth from $2.2 billion in entry fees in 2015 to $14.4 billion in 2020. Developing a statistical form indicator that can be used to produce a fantasy golf team is the goal of this project and teams created using the form indicator can be entered into fantasy golf leagues.

The legality of fantasy sports in America is an ongoing debate. Boswell (2008) illustrates the arguments for and against the legality of fantasy sports in America and indicates the reason why it has been decided that fantasy sports are indeed legal. The predominant reason is that fantasy sports rely on skill rather than chance and do not lead to destructive socio-economical behaviour. Another reason fantasy sports are legal is the fact that match fixing would not affect the outcome of a fantasy player's team as fantasy teams are made up of players from many teams and individual performances are not easy to fix.

Boswell (2008) references the Unlawful Internet Gambling Enforcement Act of 2006 which prohibits the funding of unlawful internet gambling in the United States. Fantasy sports was given an exemption to this law and an Irish man named Nigel Eccles saw this as a gap in the market and developed FanDuel, a company dedicated to fantasy sports. In 2012, DraftKings came to the market and these two companies are the largest fantasy sports companies in the market.

Forbes[4] describes the manner in which FanDuel is turning Fantasy Sports into real money. FanDuel paid out more than $500 million dollars in winnings in 2014. Sports teams have taken note. It is widely accepted that people who have something invested in a sport are more likely to watch the game involved, e.g. players in one's fantasy sports team. This increased viewership can lead to increased revenue for the sport or team involved and sporting organisations are investing in fantasy sports in order to increase their own revenue. The NBA recently signed a four year partnership agreement and have taken a minority share with FanDuel.

What makes FanDuel and DraftKings different is that they have shortened the traditional fantasy sports season which usually matches the real season. Instead players can play weekly or daily competitions. This has led to an exponential growth in fantasy sport participation and competitions. Part of this growth has been due to aggressive marketing and DraftKings spent $250 million dollars on advertising in 2015 in a deal with the sports network ESPN[5].

In terms of golf, DraftKings offer games for as little as $3 to enter and these have 100,000-150,000 participants every week. At the other end of the scale they offered a guaranteed winner's prize of $100,000 to the participant who scored the most points with their fantasy golf team for the Masters tournament in 2014. In 2015, this increased to €1 million dollars and in 2016 there was over €11 million dollars in prize funs distributed throughout the 4 major golf tournaments. In total,

[4] Forbes on Fantasy Sports, http://www.forbes.com/sites/stevenbertoni/2015/01/05/how-fanduel-is-turning-fantasy-sports-into-real-money/2/

[5] Growth of DFS, http://www.forbes.com/sites/darrenheitner/2015/09/16/the-hyper-growth-of-daily-fantasy-sports-is-going-to-change-our-culture-and-our-laws/#d289d695f254

DraftKings paid out €1 billion in prize funds in 2015 with €30 million going to fantasy golf participants.

Fantasy golf is growing each year and is driving the usage and availability of statistical analysis. Golf websites and TV channels understand this synergy and are investing huge sums of money into fantasy golf. This means fantasy golf is a growing market with potential for making substantial profits if players can make sense of the enormous wealth of golf data available.

3.2. Golf Statistics

Statistical analysis is making waves in various sports as teams and players use sports statistics, data mining tools and research methods from business and economics to gain a competitive edge.

Baseball has used statistical analysis for many years. Indeed, Bill James, a pioneer in the field, started writing books about baseball analytics in the 1970's and 1980's (James, 1977-1988). The concept was popularised by Moneyball (Lewis, 2003), which was recently made into a movie about the low-budget Oakland Athletics' use of statistical analysis to field a competitive playoff team in 2002. Now, teams in all major sports, including basketball, American football, soccer, hockey, golf and tennis use these tools.

It has been argued (Gennaro, 2013) that teams should use analytics to evaluate player performance as well as predict future performance. Gennaro suggests statistical analysis can provide better information to assist teams and managers with complex decision making. These decisions could be the upcoming team formation and tactics or business decisions like ticket prices or marketing strategies, or indeed deciding who to play in ones fantasy golf team.

Cutting-edge research is coming from higher education as professors and students in finance, mathematics and computer science apply their expertise to sports. Technology at stadiums, arenas and golf courses are capturing vast amounts of new data that help drive some of the new analysis, particularly in golf.

The earliest known research in golf dates back to the famous book (Cochran & Stobbs, 1968) where authors discuss various probabilities of making putts and where one can save strokes. Later research focussed on what attributes and skills are important for a golfer to perform and win. The game of golf can be broken down into the long game, the short game around the green and putting, and many have argued which of these is the most important. These arguments will help develop the understanding of what should constitute the important factors affecting form. This form indicator will then be used to develop an automated fantasy golf team.

It has been postulated (Alexander & Kern, 2005) putting is the most important factor in determining golfer performance however the authors admit driving is becoming more important. They are one of the first to highlight the fact putting averages and GIR are contaminated stats due to the fact the difficulty of the putt is not taken into account. Similarly, the quality of the tee-shot taken before the approach shot taken is not taken into account. However, they do not delve deep enough into this contamination to try and resolve it statistically.

Callan and Thomas (2007) model explanatory variables that can predict golfer earnings which is an attempt to understand seasonal form and which skills are most important. They also examine the possibility of experience being an explanatory variable although the results for the experience factors are weak. The explanatory skill variables evoked are:

- Driving Distance
- Driving Accuracy
- GIR
- Sand Saves
- Putting Average

These variables are in keeping with earlier research (Nero, 2001) which tried to relate particular skills to a golfer's salary.

Peters (2008) establishes the best performing attributes when predicting scoring averages by using regression. Peters uses regression analysis to predict final year earnings while looking at driving distance, driving accuracy, putting average, greens in regulation and sand saves. When examining the coefficients he found putting averages to be the most important whereas driving stats were the least important. This is an early attempt at a form indicator but it does not take into account the fact the statistics being used are contaminated. In other words, shots leading up to the putt affect the difficulty of the putt but are not taken into account when measuring putting averages.

For Scoring Average ~ Skills, the explanatory variables are:

- Driving Distance
- Driving Accuracy
- GIR
- Average Putts per GIR
- Sand Saves
- Experience

The author seems to find more correlation with his version of experience than previous research and he uses a slightly different putting statistic. However, there appears to be collinearity between certain features and again he does not take into account the contamination of the stats. E.g. GIR does not take into account the shot that put the golfer in such a good position that it made it easier to get to the green in regulation. A bad tee shot that goes into the trees can lead to a player missing the green with the next shot whereas a long and straight tee-shot leaves a much easier approach than the first one. Player A did not hit a green in regulation but player B did. In truth it was the quality of the tee-shot that lead to player B to making the green in regulation as opposed to the actual shot into the green. Similarly, putting averages are negatively affected due to the difficulty of getting to the green brought on by poor tee shots. Callan and Thomas (2007) found putting averages to have the highest regression coefficient values and hence declare putting to be the most important factor affecting earnings. The second most important factor is GIR but as I have illustrated above, these statistics are contaminated and should be analysed in conjunction with strength off the tee.

Another way golfing statistics can be contaminated is via the difficulty of the course, strength of the field or the quality of player that is involved. More difficult courses lead to higher scores, less birdies and less players under par. Less birdies means lower putting averages and lower GIR. Shmanske (2008) investigates this idea of different difficulty levels by introducing adjusted data into his model. Depending on the difficulty of the scoring or the variance, he adjusted statistics like driving accuracy, putting and GIR. Regression models were adjusted at tournament level because 60% GIR on a very difficult course might actually position a player better than 80% GIR in an easier tournament so these raw numbers may not be true reflections of the skill levels involved. The author found that adjusted data based on field averages improves the R^2 or variation explained in the model. Interestingly, driving accuracy became more significant when adjusted data was used. This is an indication that driving is more important than previous research had noted but does not yet statistically isolate the true advantage of excellent driving skills.

Golf statistics on the PGA tour are now powered by the ShotLink system[6]. ShotLink technology has been used on tour since 2003 to track and record every golf shot in near real time. ShotLink uses lasers that pinpoint the ball's location as well as hundreds of volunteers armed with wireless handheld devices that enter added data manually like number of strokes taken. Both the GPS co-ordinates combined with manually entered data are fed back to a data centre housed inside a nearby digital hub, where a team of tour staff manage the entire operation. Each golf course is digitally mapped and this acts as a map to convert ShotLink co-ordinates into distances between shots and distances to the hole. ShotLink tracks millions of shots in this way and delivers a dataset with over 500 data points. This data is immediately delivered to TV broadcasters, onsite LED scoreboards and sports websites, and after each tournament, the data is added to the historical database at its headquarters.

By 2008, ShotLink had been in use on the PGA tour for 5 years and this set of data was the most rich and substantial source of golf data that had ever existed. By using this dataset along with his GolfMetrics software, Mark Broadie (2008) revolutionised the way we analyse golfer performance. Broadie developed the strokes gained concept (initially called shot value) by using ShotLink data and his

[6] ShotLink, http://www.shotlink.com/

own mathematical equations. Broadie wanted to understand what separates the elite golfers from the rest and sought to measure the quality of their individual shots. Instead of looking at putting averages he looked at the value of each shot by comparing the potential gain to that of the average.

$$\text{Shot Value} = \text{Strokes-to-go}_{\text{before}} - \text{Strokes-to-go}_{\text{after}} - 1$$

Equation 1: Shot Value

This equation means you are calculating the expected number of strokes to go before the current shot and the expected number of strokes to go after the current shot. This means if you are in a good position for your next shot, the shot value of the previous shot will be high whereas the shot value will be low if the shot has not put you in a good position. This helps remove the contamination of stats like GIR which does not measure how good or bad the tee shot was. This new angle on golf statistics brought golf into the 21st century of analysis as now putting stats could be measured based on the difficulty of the putt and the strength of the shots preceding it.

In 2010 and 2011 a team of MIT researchers led by Professor Stephen Graves added to the strokes gained concept by using Broadie's mathematical formula to rank putters on the tour, culminating in (Fearing, Acimovic, & Graves, 2011). Using shot value and average field values for each shot the authors could then calculate probabilities of holing putts based on logistic regression and gamma regression analysis.

The strokes gained statistic (Broadie M. , 2010) was refined and simplified further by accounting for the field average that week as opposed to just the overall expected value and the analysis showed that it is actually driving that contributes most to the golfer’s score. This debunked many years of previous research although it could be argued that certain tournament winners may still rely more on their putting ability depending on the course.

The following year, the TOUR worked with Broadie and Graves to fine-tune the strokes gained statistic, and in 2014, the organization introduced it to the golf community. It was the first time in 15 years that the tour modified its core set of statistics. Today, strokes gained putting and strokes gained tee-to-green are considered the most accurate way to measure overall performance. The new putting statistic calculates the number of putts a player takes to reach the hole and compares it with his opponents, while taking the distance of the putts into account which is something previous putting stats had ignored. These new statistics form the foundation of my form indicator analysis.

The top PGA stars and coaches actively use these statistics to understand their own performance but golf fans are now seeing stats as a new and interesting dimension to the enjoyment of golf. MoneyGolf (Agger, 2010) chronicles the evolution of golf statistics since the inception of ShotLink data. In the article, Agger describes the development of Broadie’s Shot Value which led to the strokes gained stats. He recognises that these statistics are much more informative as they illustrate the value of each shot and expected score as opposed to a more general stat like putting average. However there is still a lack of data mining and performance prediction or form analysis. The strokes gained statistics are fantastic for evaluating performance but they have not been investigated as predictors or indicators of form and future performance. This will be addressed in this paper when strokes gained are used to derive a form indicator.

Form and confidence can change in golf and Rosenqvista and Skans (2015) illustrate the confidence that good form can bring and how this extends the run of form. I will use this idea when I analyse the year to date form and course history. Golfers perform well repeatedly in certain tournaments and there does not seem to be much research into this area yet. Previous performance in a tournament is considered as an input to the data driven team selection.

3.3. Combinatorial Optimisation

Genetic Algorithms

A Genetic Algorithm (GA) is a population-based optimisation algorithm that uses techniques inspired by evolutionary biology such as selection, inheritance, mutation, and recombination. GAs manipulate a population of candidate solutions (or individuals), traditionally represented by binary strings, that evolve towards better solutions. Some of the best individuals can be kept (elites) and used in the new population. This prevents the best solutions being lost during evolution. To find out more about GAs and examples of various applications, see (Davis, 1991). A typical GA implementation is shown in Figure 1.

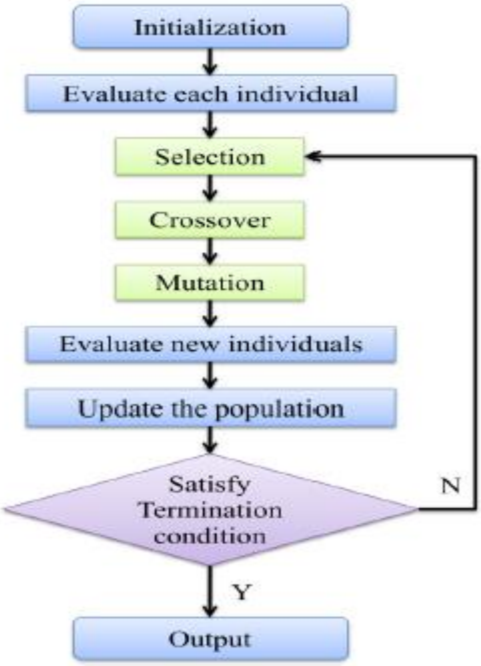


Figure 1: Standard GA implementation (Xu, 2012)

Genetic Algorithms were first developed in the 1960's and on into the 1970's by John Holland, culminating in (Holland, 1975). At the First National Conference on Genetic Algorithms in 1985, it was shown that GAs could be used to solve combinatorial problems with a particular focus on engineering problems. E.g. (Fourman, 1985). As research and applications increased, GAs were used to solve problems that were too difficult to solve using enumerative methods. GAs have since been used to solve many combinatorial problems from the world of engineering to the world of finance, and beyond.

The selection of a fantasy team can be seen as a combinatorial problem with one or more objective functions and at least one constraint (budget). In (Matthews, Ramchurn, & Chalkiadakis, 2012) the authors use past performance to derive Bayesian probabilities that the fantasy football player will play, score or assist or keep a clean sheet. They then try to maximise this value while observing game rule constraints. They run many simulations using team based probabilities and player based probabilities to ascertain expected results. Using the Bayesian Q-learning heuristic they maximise expected points by observing the fantasy football game rules for scoring points. I will also use fantasy golf points when calculating my form indicator.

Cricket team selection was solved as a multi-objective problem using NSGA-II in (Ahmed, Jindal, & Deb, 2011) by selecting a team based on maximising batting and maximising bowling while remaining under budget. This is a similar goal to what I am looking for in fantasy golf although they do not perform statistical analysis on their results to convey significance which is something I do when comparing the data driven team with the human generated teams. More recently (Sarda & Sakaria, 2015) also used NSGA-II to select a football team using a variety of features depending on the strategy of the proposed team, i.e. defensive or offensive. They propose using weights for different features which can then alter the strategy and team produced. It is not clear if this is necessarily a multi-objective problem or a single objective that could be solved with linear programming. I will use NSGA-II to solve the team selection problem when using two objective functions.

Binary Integer Programming

If the team selection can be modelled as a constrained (budget) single-objective (single form indicator) optimisation problem then BIP can be used to solve the problem by using each eligible golfer as a decision variable that can be selected or not selected (binary). Examples of problems that can be solved using BIP can be found in (Chinneck, 2015) and (Nering & Tucker, 1993).

3.4. Summary

Fantasy sports has become a multi-billion dollar industry, with millions of dollars flowing through the industry each day thanks to the introduction of DFS. Fantasy golf has grown in popularity in recent years but it is still in its infancy in terms of monetary winnings. Golf analytics is also relatively new in terms of quality statistics. Golf statistics have been around for many years but many of the old adages like "*Drive for show and putt for dough*" (Alexander & Kern, 2005) are being questioned. It seems now thanks to the strokes gained statistics, driving is more important than putting in terms of gaining strokes on the field and that one should actually "drive for dough". GA's have been used to solve team selection problems and these team selections have often been formulated as multi-objective problems. However, the team selection in (Sarda & Sakaria, 2015) seems as though it could be solved as a single objective problem with the correct use of weights which is a method I explore.

A lot of golf research focusses on the analysis of what makes a successful player over a longer period but there is a lack of research over shorter periods. There is an old saying in sport that form is temporary but class is permanent and I would like to investigate and expand the existing research that studies longer term class (years) and see if it translates to form over a single season. I will use the existing analysis and some of my own analysis and data mining techniques to fuel the objective function of the GA. Broadie (2010) signals the over-emphasis on putting stats so I will look at drives, approaches and strokes gained in order to build a form indicator. I shall also look at course history and see if past performances have an impact in future performances in the same tournament. I shall investigate this problem as both a multi-objective formulation and a single objective formulation of the team selection problem with a budget constraint.

The gap in the research is to analyse year-to-date form using the available statistics while also automating team selection for fantasy golf. The output of the algorithm will be a fantasy golf team that can be entered into an online fantasy golf league.

4. Methodology

4.1. Overview

The first part of the practicum is to analyse golf data from ShotLink to establish the best performing features and develop a numeric interpretation of form.

The second part of the practicum is an implementation of an algorithm that takes a set of golfers, their cost and their form as input. The output of the algorithm is a team of 10 golfers that can be used as a winning fantasy golf team. A GA is used for multi-objective optimisation (maximise YTD form, maximise course history) but when the form indicator is combined into a single objective (combining YTD with course history) binary integer programming (BIP) is used as this will guarantee an optimum solution.

Combining both parts of this practicum, the aim is to produce a data driven team of golfers that can be entered into online fantasy golf competitions and finish in the top 50 percentile.

Shotlink data as well as freely available data from the online fantasy golf game[7] is used and the CRISP-DM process is followed to establish form indicating variables. The CRISP-DM process is described in more detail in 4.3.

Following on from the previous research highlighted in 3.2, initial features analysed were strokes gained tee-to-green, strokes gained putting, driving accuracy, GIR and scrambling. There are over one hundred and fifty other available measures and these were also analysed. Aggregations and combinations of features were used to uncover further measures. In order to use data in year-to-date (YTD) form, moving averages were developed before each of the viable tournaments. These moving averages indicate seasonal form up until the start of the upcoming tournament and this way the GA or BIP can be run before every tournament and use YTD form that is updated weekly.

GA uses NSGA-II and is detailed further in 4.4. The BIP solution was produced using Excel solver and both solutions were evaluated by calculating the points scored by the BIP/GA team compared to points scored by human generated teams.

4.2. Tools and Software

Tools and software used were as follows:

- Shotlink Dataset
- Microsoft Excel used for data pre-processing and LP solver
- SolveXL used for NSGA-II
- R and RStudio
- Java 1.8 and Eclipse
- HP Elitebook, Intel Core Duo CPU @2.53GHz, Windows 7 x64

4.3. ShotLink Golf Data Set using CRISP-DM

The CRISP-DM model is used for data analysis and the data mining part of the practicum. The CRISP-DM model helps focus the mind and iteratively step through the various stages of the process, looping back to previous stages if necessary. It gives structure to a process which involves a large amount of exploring the unknown intricacies of a new data source.

[7] https://fantasygolf.irishtimes.com

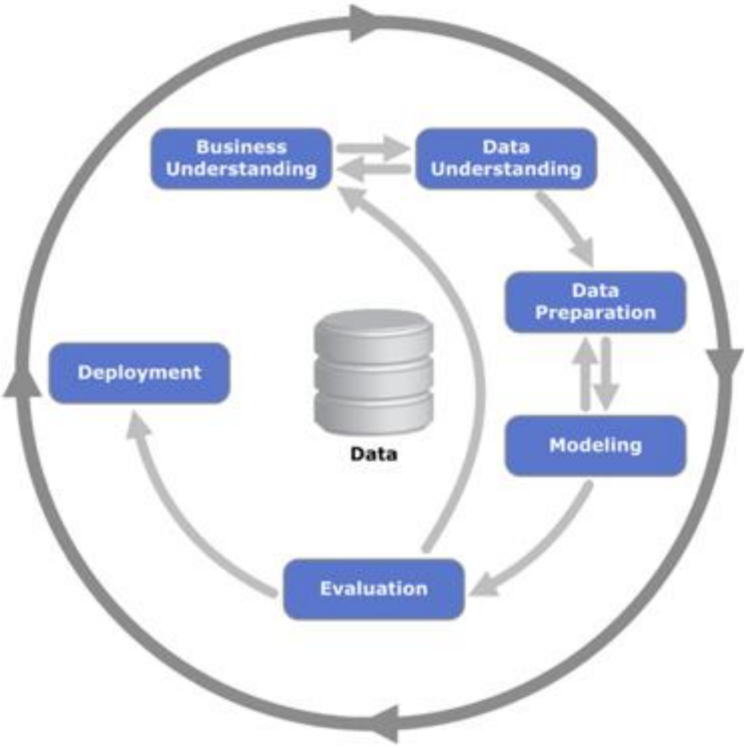


Figure 2: CRISP-DM

4.3.1. Business Understanding

The business is fantasy golf and the aim is to produce a data driven team of golfers using a data driven algorithm that outperforms human competitors. Form indicators must be established using ShotLink data and the seasonal form should be a moving average that is updated before every upcoming tournament. Course history form is established by analysing results from the 5 previous seasons in the same tournament. The form indicators and algorithm should be flexible, reusable and easily modifiable. Summary reports should also be made available for a weekly illustration of in-form golfers.

4.3.2. Data Understanding

Individual Tournaments

Tukey (1977) identifies a number of processes when exploring data and these are used in this practicum. These include: explain data, explore data, visualise, identify best performing features, identify outliers, use appropriate tools, techniques (aggregation, decision trees, correlation) and hypothesise. Tukey also explains the importance of fitting the model to the data and not trying to fit the data to your favourite model. This principle is followed throughout the paper.

There are several levels of data granularity in the ShotLink dataset: stroke level, hole level, round level, event level, course level, radar launch and radar trajectory. For this practicum, event level was chosen as the other levels are unnecessarily granular and event level contains statistics on strokes gained which are aggregated to cover an entire event. This is a fairer indication of form as strokes gained at hole-level or even round-level is too narrow a sample size. (See Appendix for list of event level columns).

Boxplots, scatterplots, aggregation, categorisation and classification trees are used. The goal of the fantasy golf team is not to predict the tournament winner per se but rather select a team of 10 golfers with as many top 25 finishers as possible. Because of this reason the data was categorised by fitting players into 3 categories based on their finish position:

- Top25
- 26-CutLine
- MC (Missed Cut)

Top 25 refers to players finishing in the top 25 and this was chosen as a category as these players score the most points in fantasy golf. 26-CutLine refers to players who finished outside the top 25 but made the cut. MC refers to players who missed the cut and a player misses the cut if he is outside the top 70 after the first 2 days. Players outside the cut do not play the second half of a tournament and they do not score points in fantasy golf.

With this understanding of the business goal and data layout, it was decided that classification trees could be used as they are easy to use, well tested and powerful methods for examining the relationships between numeric variables and categorical variables/classes. Correlation analysis and linear regression were used to examine numeric relationships as these are good tools for examining such relationships. These are examples of fitting models to the data and not the other way round.

There were some challenges and limitations of the data that were discovered during the data understanding phase.

- ShotLink data does not have strokes gained data for all the tournaments
- ShotLink data only covers PGA tour events
- ShotLink has data on many players that are not included in fantasy golf
- Short sided/obstructed/plugged/player stance

Some tournaments do not include strokes gained data as the tournament organiser has not yet agreed to include it. However, unless a player is drastically better or significantly worse than their average, the overall average should not be affected greatly. This is one of the benefits of using seasonal YTD averages. ShotLink does not include data for European Tour events so players from the PGA tour who decide to play some events in Europe cannot have their statistics from the European events included in their YTD averages. ShotLink cannot yet give indications of a shot being obstructed or being more difficult than the distance suggests. Strokes gained expected values are based on the expected number of strokes taken for shots of similar distances but this does not take into account a shot that may be blocked by a tree or plugged deep in a bunker. This is a potential area of future research. Fortunately, these situations are rare in the grand scheme of golf tournament play so it does not greatly affect YTD averages.

4.3.3. Data Preparation

YTD Moving Averages

Pre-processing needs to be performed so analysis can be completed and moving averages calculated. Steps for YTD moving averages:

- Categorise data such that players fall into the categories highlighted above
- Select individual tournaments for initial exploration
- Aggregate data on player to introduce YTD sums and use this to produce moving averages
- Add user defined columns

Column Name	Column Description
ID	Unique identifier
AvgScorePerRound	Average = total score / number of rounds played
BirdieBogeyRatio	Number of birdies / number of bogeys
FG_Points	YTD fantasy golf points scored
Course History	Performance over past 5 seasons in same tournament
GA Form	Form indicator based on combination of best performing features

Table 1: User defined columns

4.3.4. Data Modelling

Decision tree learning and regression used to help determine best performing features and explanatory variables. Data modelling tools used were decision trees and regression with statistical significance tests performed on correlations and results.

4.3.5. Model Evaluation

The model in this case is not a predictive model but a data driven form indicator based on perceived explanatory variables. As part of individual tournament analysis, features that are highly correlated when compared to average round score are important and will be used in the YTD moving averages.

The highly correlated features found during analysis of individual tournaments were then used in YTD moving averages and compared to known results (Top25, 26-CutLine, MC) to see if correlation still holds.

4.3.6. Deploy

Deploy by using the form indicators as input to the team selection program. The form indicators are varied and tested until the program produces the best results. The algorithm is trained on random tournaments throughout the year.

The team selection program was then tested on the following tournaments. The results of each program is a team of 10 golfers that can be entered into fantasy golf competitions corresponding to that particular tournament.

Tournament Number	Tournament Name
1	The Open Championship
2	Canadian Open
3	PGA Championship
4	Travelers Championship

Table 2: PGA tournaments used in experiments

4.4. Genetic Algorithm Implementation

There are over three hundred golfers to choose from in the game and the user must pick a team of 10. This means there are more than 3x10^{18} possible combinations. Not all golfers take part in every tournament and the combination of golfers must remain under €100 million in value. We want to maximise the YTD form of the team and also maximise the course history form of the team. Both of these factors combined makes systematically choosing a team a difficult multi-objective combinatorial problem. For this reason, a GA is a good choice as it can evolve a solution from the vast solution space and converge to an optimal, close to optimal or a set of Pareto optimal solutions depending on the fitness function.

The goal of the GA is to produce a chromosome containing 10 genes that represents a fantasy golf team of 10 golfers and this was carried out using NSGA-II. Each team consisted of a chromosome of 10 integers. There were over three hundred golfers used, each of whom had an ID number, value (€ million), Form Score and Course History Form. A valid team must consist of 10 golfers, none of whom can be repeated and the team value must be less than or equal to €100 million. The algorithm evolved an initial random population over a thousand generations. Tournament selection was used to pick parents for reproduction. A one-point crossover operator was used in conjunction with standard integer mutation which replaces a randomly selected gene with a random golfer not already in the team. Thirty trials were carried out for each experiment and the experiment was run independently for 4 official PGA tournaments: The Open Championship, Canadian Open, PGA Championship and Travelers Championship. Teams produced by the GA were compared against other teams in the fantasy golf game by calculating the total number of points scored and comparing this to the mean.

Fitness Function

Two fitness functions were investigated, a multi-objective fitness function and a single objective fitness function.

The multi-objective fitness function for this GA is a mixture of YTD form and course history form found during the data analysis and data mining phase of the practicum.

$$Max \sum_{i=1}^{10} YTD\ form(i)$$

$$Max \sum_{i=1}^{10} YTD\ course\ history(i)$$

Subject to:

$$\sum_{i=1}^{10} cost(i) \leq €100$$

Equation 2: Multi-objective fitness function

The single objective fitness function combines YTD form with course history into a single form indicator.

$$Max \sum_{i=1}^{10} YTD\ form(i)$$

Subject to:

$$\sum_{i=1}^{10} cost(i) \leq €100$$

Equation 3: Single-objective fitness function

Crossover

There has been much written about various implementations of crossover; see (Holland, 1975), (De Jong, 1975), (Booker, 1987) and (Davis, 1991), but 1 point crossover used as this has been proven to be a simple but effective use of crossover. Teams are randomly selected for crossover at a rate of 75%.

Mutation

Likewise there has been much written about mutation design and rates with varying implementations. The goal of mutation is to prevent a loss of diversity and this is achieved by randomly selecting a golfer from a team and replacing it with another randomly selected golfer not already in the team. Teams are randomly selected for mutation at a rate of 10%.

Random Number Generator

Pseudo-random number generator used and a new seed set for each implementation.

4.5. Binary Integer Programming Implementation

As Equation 3 has a discrete solution, BIP can be used to solve the equation. Microsoft Excel Solver was used and the BIP solution can also be used to compare results with that of the GA for the single objective function. A GA is not guaranteed to converge to the optimum solution if one exists so when generating a fantasy golf team using the single objective solution the BIP solution is the preferred solution. The GA can however be used if it converges to the BIP optimum.

4.5.1. Experiments

Once the form indicators have been produced, they are used to run the GA or BIP for the 4 tournaments indicated in Table 3. The GA is run 30 times for each tournament and the fittest team from the set of 30 is chosen as the Fantasy golf team for that particular tournament. The BIP need only be run once because it will find the optimum for a constrained single-objective optimisation problem. The results of each of the 4 experiments are calculated based on the official fantasy golf points scored that week.

Tournament Number	Tournament Name
1	The Open Championship
2	Canadian Open
3	PGA Championship
4	Travelers Championship

Table 3: PGA tournaments used in experiments

4.5.2. Measure Success

Using Students t-tests, the points scored by the data driven teams are compared against the mean weekly scores of human participants to test the statistical significance of the results. The data driven teams generated by the GA or BIP solver will be successful if they outperform human generated teams.

5. Results and Analysis

5.1. Feature Reduction

Correlation was performed on:

- Individual Tournaments
- YTD Averages

The best performing features are:

- Strokes Gained Tee-to-Green
- Strokes Gained Putting
- Birdie:Bogey (Ratio)
- Adjusted Average Score
- Average Position (user defined column)
- YTD Fantasy Points (user defined column)
- Course History (user defined column)

Individual Tournaments

Correlation tests were performed on attributes based on findings outlined in the literature review in section 3.2, starting with strokes gained statistics and then trying statistics referenced in earlier literature (driving accuracy, putts per green in regulation, GIR, scrambling, sand saves etc.). Correlation tests were also performed on many other event level statistics like “proximity to the hole” or “putts from 5 feet or less”. A full list of event level statistics can be found in the Appendix. Figure 3 highlights the strong correlation between strokes gained statistics and Average Score per Round as well as Birdie:Bogey ratio and Average Score per Round. A low Average Score per Round emphasises a good finishing position and we can see clearly here that the average score per round decreases as strokes gained and Birdie:Bogey ratio increases.

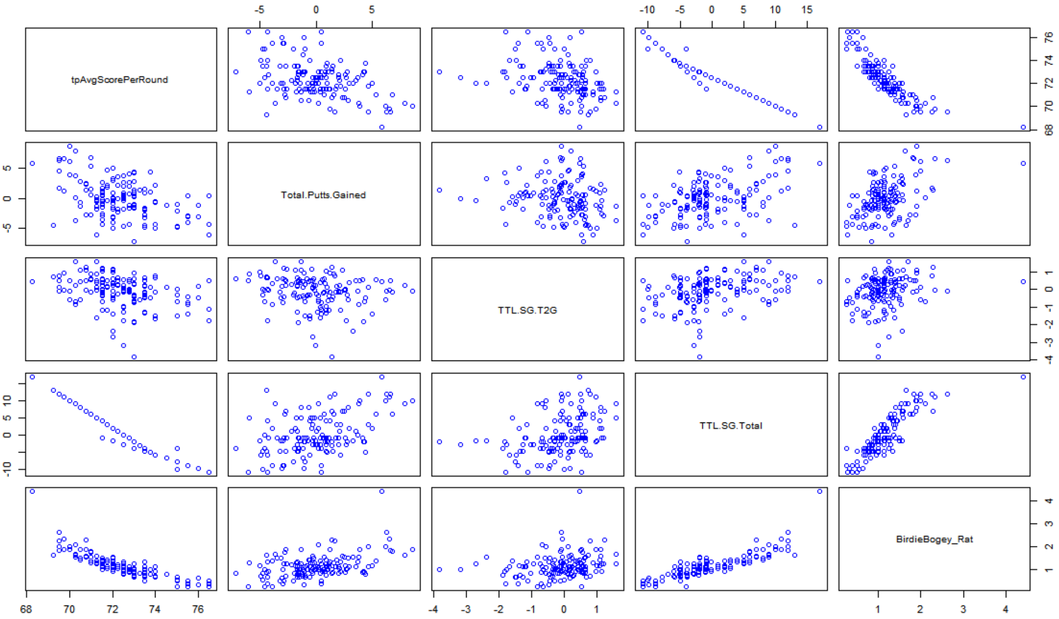


Figure 3: Correlations for the Players Championship

Table 4 shows the correlation between the strokes statistics and Average Score per Round as well as Birdie:Bogey ratio and Average Score per Round. As expected we have a negative relationship which is particularly strong between SG Total and Average Score per Round

	SG Putting	SG T2G	SG Total	Birdie:Bogey
AverageScorePerRound	-0.44	-0.32	-0.94	-0.84

Table 4: Average Score per Round correlation for the Players Championship

This process was repeated for many tournaments and the results are shown in Table 5. Strokes gained statistics along with Birdie:Bogey ratio were consistently the best performing variables.

Average Score Per Round Tournament Name	SG Putting	SG T2G	SG Total	Birdie:Bogey
The Players Championship	-0.44	-0.32	-0.94	-0.84
Pebble Beach	-0.53	-0.22	-0.76	-0.86
The Memorial	-0.58	-0.48	-0.96	-0.89
The Sony Open	-0.45	-0.53	-0.97	-0.76
Dean and Deluca	-0.60	-0.29	-0.95	-0.87
Sanderson Farm	-0.56	-0.48	-0.92	-0.85

Table 5: Correlation matrix for individual tournaments

Before the strokes gained statistics were in use it was commonly believed that Driving Accuracy, GIR and GIR Putts were leading factors in analysing or predicting golfer performance. The analysis here gives some credence to those theories but strokes gained has a much stronger correlation. Indeed, there is collinearity at play when linear regression is used and the correlation can primarily be explained by the strokes gained statistic. Figure 4 illustrates the correlation between Driving Accuracy and Average Score per Round, as well as Putting Avg. GIR Putts and Average Score per Round. However, as we can see below, strokes gained statistics are more closely correlated to Average Score per Round. This is confirmed in more detail in (Broadie M. , 2010).

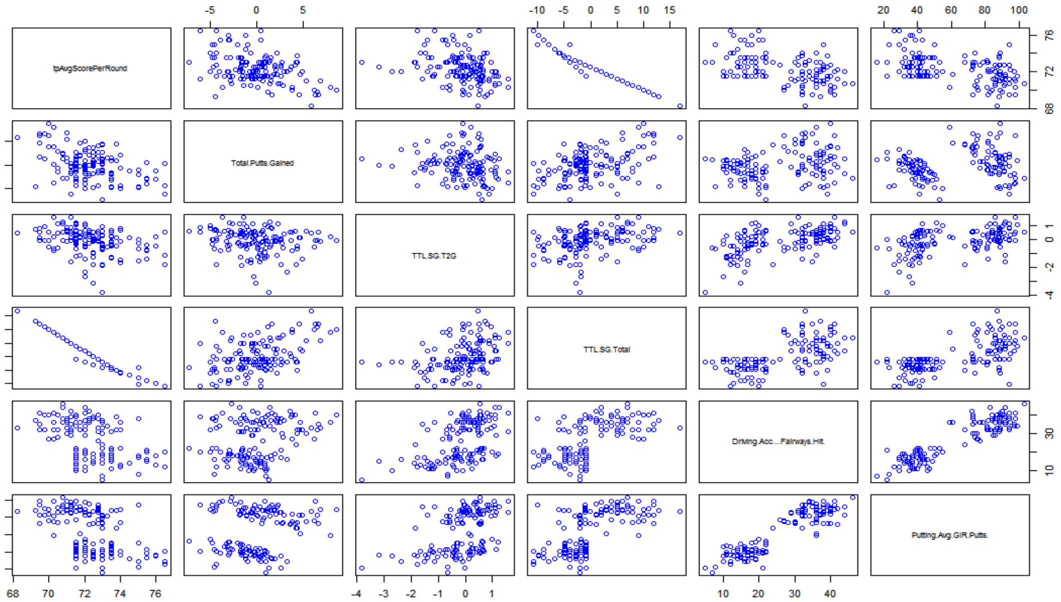


Figure 4: Collinearity for the Players Championship

Linear regression was used to further explain the relationships between the explanatory variables above and the dependent variable (Average Score per Round). Low p-values intimate the explanatory variables in the model are significant and could be used as part of the form indicators.

Tournament	**Linear Regression Explanatory Variables**	**p-values(s)**
The Players Championship	SG T2G + SG Putting	$2.07\text{x}10^{-8}$, $6.85\text{x}10^{-12}$
The Players Championship	SG Total	$2\text{x}10^{-16}$
The Players Championship	SG Total + Birdie:Bogey	$2\text{x}10^{-16}$, 0.00854
Pebble Beach	SG T2G + SG Putting	$6.16\text{x}10^{-5}$, $2.63\text{x}10^{-14}$
Pebble Beach	SG Total	$2\text{x}10^{-16}$
Pebble Beach	SG Total + Birdie:Bogey	$5.6\text{x}10^{-8}$, $2\text{x}10^{-16}$
The Memorial	SG T2G + SG Putting	$1.46\text{x}10^{-9}$, $2.24\text{x}10^{-13}$
The Memorial	SG Total	$2\text{x}10^{-16}$
The Memorial	SG Total + Birdie:Bogey	$2\text{x}10^{-16}$, 0.206
Sony Open	SG T2G + SG Putting	$2\text{x}10^{-16}$, $1.02\text{x}10^{-15}$
Sony Open	SG Total	$2\text{x}10^{-16}$
Sony Open	SG Total + Birdie:Bogey	$2\text{x}10^{-16}$, 0.00182
Dean and Deluca	SG T2G + SG Putting	$5.75\text{x}10^{-8}$, $2\text{x}10^{-16}$
Dean and Deluca	SG Total	$2\text{x}10^{-16}$
Dean and Deluca	SG Total + Birdie:Bogey	$2\text{x}10^{-16}$, 0.767
Sanderson Farm	SG T2G + SG Putting	$3.35\text{x}10^{-15}$, $2\text{x}10^{-16}$
Sanderson Farm	SG Total	$2\text{x}10^{-16}$
Sanderson Farm	SG Total + Birdie:Bogey	$2\text{x}10^{-16}$, 0.445

Table 6: Linear regression results for individual tournaments

As Table 6 highlights, strokes gained statistics are good predictors and although Birdie:Bogey ratio on its own can be a good predictor it is not always significant when combined with strokes gained statistics. This implies there may be some collinearity between these factors. This will be taken into account when producing the form indicators. Regression analysis was performed on other variables but the variables in Table 6 were the most significant.

Using the categorisation described in the data preparation phase and the features pertained above, decision trees were used to see if these features could be used as classifying features. Figure 5 is an example of a decision tree classifier using strokes gained statistics and Birdie:Bogey ratio. The results of the best performing decision trees are specified in Table 7.

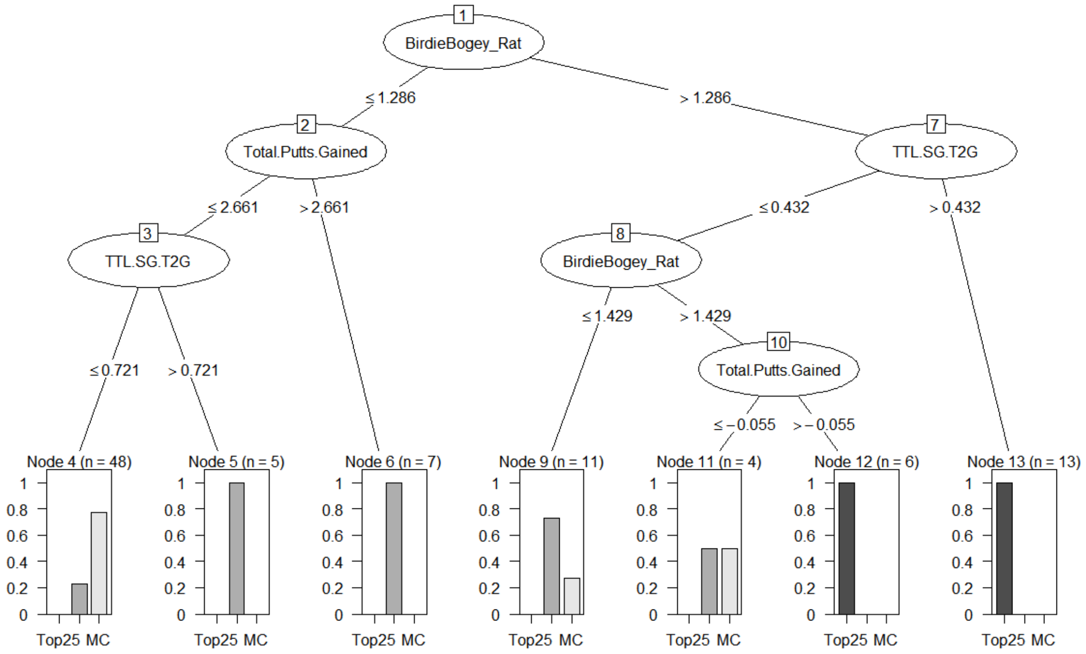


Figure 5: Decision tree for the Players Championship

Tournament	**Decision Tree Classification Variables**	**Classification Success on Test Data %**
The Players Championship	SG-T2G + SG-Putting	61%
The Players Championship	SG-Total	69%
The Players Championship	SG-T2G + SG-Putting + Birdie:Bogey	87%
Pebble Beach	SG-T2G + SG-Putting	
Pebble Beach	SG-Total	
Pebble Beach	SG-T2G + SG-Putting + Birdie:Bogey	
The Memorial	SG-T2G + SG-Putting	40%
The Memorial	SG-Total	85%
The Memorial	SG-T2G + SG-Putting + Birdie:Bogey	56%
Sony Open	SG-T2G + SG-Putting	57%
Sony Open	SG-Total	93%
Sony Open	SG-T2G + SG-Putting + Birdie:Bogey	90%
Dean and Deluca	SG-T2G + SG-Putting	52%
Dean and Deluca	SG-Total	85%

Dean and Deluca	SG-T2G + SG-Putting + Birdie:Bogey	61%
Sanderson Farm	SG-T2G + SG-Putting	58%
Sanderson Farm	SG-Total	93%
Sanderson Farm	SG-T2G + SG-Putting + Birdie:Bogey	74%

Table 7: Decision Tree Results on Individual tournaments

Table 7 indicates the best performing decision trees using the best performing variables. Many other decision trees were run using different combinations and variables, such as GIR and scrambling but these were eliminated due to poor classification and collinearity.

YTD Moving Averages

After performing correlation analysis, regression analysis and decision tree learning on individual tournaments, the strongest features were used to create moving averages for YTD data (Table 8). Several new columns were also added based on existing data (Table 9).

Column Name	**Description**	**High or Low Preferred?**
SGT2G	Average strokes gained tee to green (YTD)	High
SGPutting	Average strokes gained putting (YTD)	High
FGAvg_score	Average adjusted score (YTD)	Low
FGBirdieBogey	Average ratio of birdies to bogeys (YTD)	High

Table 8: Best performing features

User Defined Column Name	**Description**	**High or Low Preferred?**
FGAvg_Position	Average finish position (YTD)	Low
FG_Points	Points the golfer has score in the fantasy golf game (YTD)	High
Course_History	Points scored based on results in the same tournament over previous 5 seasons	High

Table 9: Best performing features (user defined columns)

Adjusted score used as this is the PGA’s attempt to adjust scores based on the strength of the field. Some courses are easier than others and so the average score for those tournaments is lower than the seasonal average. Conversely, the Majors tend to be more difficult and so the average score is higher than the seasonal average. As a further adjustment and to take into account finishing position as well as score, a column was added for average position. FG_Points added to give credit to European Tour players whose stats are not included in the ShotLink database. Course History added based on performances in the same tournament over the previous 5 years.

Figure 5 shows the correlation between YTD moving averages and the actual scores during The Masters tournament. This tournament and the Heritage tournament formed the basis of the training data as these occurred earlier in the season. Testing of the team selection program took place in the summer.

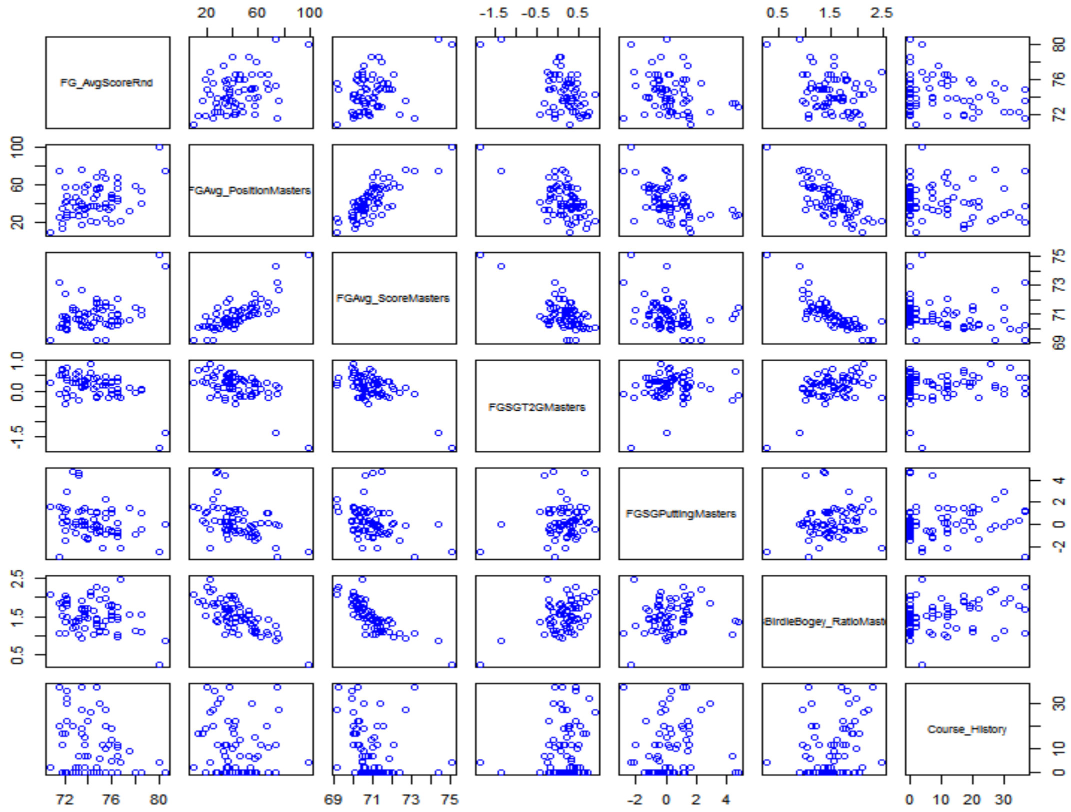


Figure 6: Correlation between YTD averages and avg score in The Masters

Table 10 shows the correlation between the best YTD moving averages and Average Score per Round. As expected we have positive relationships for AvgPosition and AvgScore because high scores and high positions are bad whereas a low score is good. When it comes to the other statistics like SG and Course History, we have a negative correlation as the score gets lower (good) when players SG statistics or course history increases (good). Correlations are not as strong for YTD aggregates as they are for individual tournaments because form fluctuates throughout the season so correlations will not be as strong over a longer period of time. These results are satisfying nonetheless as they can be used together to form one or more form indicators.

Avg Score Per Round Tournament Name	**SG T2G**	**SG Putting**	**Birdie: Bogey (ratio)**	**Avg Position**	**Avg Score**	**Course History**
The Masters	-0.52	-0.27	-0.30	0.41	0.39	-0.27
The Heritage	-0.1	-0.04	-0.16	0.22	0.15	-0.21

Table 10: Avg Score Per Round correlation for YTD data

Decision trees were run on the YTD moving averages to evaluate the relationship between chosen features and the categorical classes created as part of data preparation, i.e. “Top25”, “26-CutLine” and “MC”. The results can be seen in Table 11.

Tournament	**Decision Tree Classification Variables**	**Classification Success on Test Data %**
The Masters	SG-T2G + SG-Putting	39%
The Masters	SG-Total	52%
The Masters	SG-T2G + SG-Putting + Birdie:Bogey	39%
The Masters	SG-T2G + SG-Putting + Birdie:Bogey + Avg_Poisition + Avg_Score + FG_Points + Course_History	39%
The Heritage	SG-T2G + SG-Putting	38%
The Heritage	SG-Total	38%
The Heritage	SG-T2G + SG-Putting + Birdie:Bogey	41%
The Heritage	SG-T2G + SG-Putting + Birdie:Bogey + Avg_Poisition + Avg_Score + FG_Points + Course_History	35%

Table 11: Decision tree results on YTD data

The results in Table 11 are mixed and strokes gained statistics seem the most predictive. After completing the exploration, correlation analysis and decision tree learning, the best performing features are:

- Adjusted Average Score
- SG Tee-to-Green
- SG Putting
- Birdie:Bogey
- Average Position
- YTD Fantasy Points
- Course History

The features above were combined and experimental tests run using the GA in order to get the best combination of YTD Form and Course History. These two form measures could then be used in the GA which could solve the multi-objective optimisation problem of selecting a team of 10 golfers.

Occam’s razor recognises the simplest solution is often the best so initial attempts to define the form indicator were based on SG statistics only. Other weighted moving averages were added to the form indicator and tested.

5.2. Form Indicator

Based on empirical results using the GA to solve the problem as a constrained multi-objective optimisation problem and a constrained single-objective optimisation problem it was discovered that a single-objective formulation gave the best results. By combining YTD form with course history, a single form indicator was created. In the constrained single-objective optimisation problem the formulation is similar to Equation 3.

$$Max \sum_{i=1}^{10} (YTD\ form + course\ history)(i)$$

Subject to:

$$\sum_{i=1}^{10} cost(i) \leq €100$$

Equation 4: Single-objective fitness function

Subsequently, based on empirical results using the constrained single-objective formulation it was concluded that the strokes gained statistics were sufficient for indicating form in the Majors while added features were needed for regular tour events. The reason for this is the strength of the field. The Majors have stronger fields and thus a larger range and standard deviation of strokes gained stats which means strokes gained provides a good differentiator. In the lesser tournaments the players are more closely matched so strokes gained alone does not differentiate them enough in terms of form. For this reason, more features are needed.

Final form indicators deduced are highlighted below, based on empirical data and experimentation of weights.

Major Tournaments:

$$f(x) = SGT2G + SGPutt$$

Equation 5: Form indicator for Majors

Regular PGA events:

$$f(x) = \left[\frac{AvgScore - BirdieBogey - SGT2G - SGPutt}{AvgPos}\right] + \omega_1 FGPoints + \omega_2 CourseHistory$$

Equation 6: Form indicator for non-Major events

The best weightings found empirically are as follows:

ω1 = 0.005

ω2 = 0.1

Using a genetic algorithm provides flexibility, adaptability and versatility because it can be used to solve constrained multi-objective optimisation problems. It can also be adapted and updated to take into account different rules of different fantasy sports games. However now that the problem has become a constrained single-objective optimisation problem it can be formulated and solved using Integer Programming. Indeed, BIP should be selected over the GA as BIP is guaranteed to find an optimum whereas a GA is not guaranteed to find the optimum.

5.3. Binary Integer Programming Formulation

As we are using a single objective function, the problem can be formulated as an Integer Programming problem with binary decision variables. The constrained optimisation formulation is

$$Max \sum_{i=1}^{n} (FormIndicator)_i g_i$$

Subject to:

$$\sum_{i=1}^{n} cost(i) \leq €100$$

$$\sum_{i=1}^{n} g_i = 10$$

$$g_i = 0 - 1$$

$$N = Total\ number\ of\ eligible\ golfers$$

Where G is the set of eligible golfers and “$(FormIndicator)_i$” is the form indicator for golfer g_i,

Equation 7: Single-objective fitness function

This problem was solved for the 4 test tournaments to give the optimum data driven team selection for each.

5.4. Binary Integer Programming Results

The BIP was run for each tournament and the result of the BIP is a team of 10 golfers per tournament. The problem was solved using Excel solver and a solution was also found using the GA. The GA was run 30 times with 30 different random number seeds for each tournament and converged to the same result every time. The GA results matched those of the BIP optimum solution.

To understand the relevance of the BIP results I first calculated a confidence interval for the mean score of the human generated fantasy teams. To do this, a random sample of 100 participants were chosen and a sample mean score was calculated for each tournament. I then built a confidence interval using the Students t-distribution which gives us a confidence interval that the mean will fit into 95% of the time. This confidence interval can then be used as a comparison to the BIP score that week.

Results for The Open Championship can be seen in Table 12.

- BIP team points for The Open Championship: 1018
- Sample mean team points for The Open Championship in FG game: 903
- BIP outperforms the mean

The Open Championship

Player Name	Value	Form	Actual Finish	FG Points
Jason Day	15	3.39	Top25	46
Jordan Spieth	15	2.57	26-CutLine	7
Phil Mickelson	13.5	2.25	Top25	517
Steve Stricker	5.5	2.24	Top25	442
Jamie Donaldson	10.5	2.31	26-CutLine	7
William McGirt	6.5	1.65	MC	-5
Daniel Summerhays	7	1.56	26-CutLine	7
Patton Kizzire	8.5	1.52	26-CutLine	7
Kiradech Aphibarnrat	11	2.12	MC	-5
Paul Dunne	7	1.80	MC	-5

Table 12: BIP team and results for The Open Championship

Results for the Canadian Open can be seen Table 13

- BIP team points for the Canadian Open: 854
- Sample mean team points for the Canadian Open in FG game: 374
- BIP outperforms the mean

Canadian Open

Player Name	Value	Form	Actual Finish	FG Points
Dustin Johnson	14.5	23.63	Top25	345
Jason Day	15	14.99	Top25	95
Jim Furyk	14	11.67	Top25	115
Matt Kuchar	12.5	15.14	Top25	195
William McGirt	6.5	8.95	26-CutLine	5
Colt Knost	5.5	6.79	MC	-3
Daniel Summerhays	7	6.73	MC	-3
Roberto Castro	5	6.15	26-CutLine	5
Chris Kirk	10	6.01	Top25	95
Ryan Palmer	9	5.78	26-CutLine	5

Table 13: BIP team and results for Canadian Open

Results for the PGA Championship can be seen in Table 14.

- BIP team points for the PGA Championship: 1453
- Sample mean team points for the PGA Championship in FG game: 775
- BIP outperforms the mean

PGA Championship

Player Name	Value	Form	Actual Finish	FG Points
Jason Day	15	3.06	Top25	517
Jordan Spieth	15	2.40	Top25	172
Phil Mickelson	13.5	2.11	26-CutLine	7
Jamie Donaldson	10.5	2.08	26-CutLine	7
Steve Stricker	5.5	2.04	26-CutLine	7
Kiradech Aphibarnrat	11	1.89	26-CutLine	7
William McGirt	6.5	1.59	Top25	262
Bryce Molder	5.5	1.49	MC	-5
Patton Kizzire	8.5	1.48	26-CutLine	7
Daniel Summerhays	7	1.40	Top25	472

Table 14: BIP team and results for PGA Championship

Results for the Travelers Championship can be seen in Table 15.

Total Week Points for Travelers Championship: 774

Travelers Championship

Player Name	Value	Form	Actual Finish	FG Points
Kevin Chappell	7	7.99	MC	-3
Jim Furyk	14	9.15	Top25	275
Branden Grace	14.5	15.74	MC	-3
Tyrrell Hatton	7	12.77	Top25	65
Charley Hoffman	10	8.20	Top25	15
Brooks Koepka	13.5	12.99	Top25	195
Matt Kuchar	12.5	13.87	Top25	65
Bryce Molder	5.5	7.47	26-CutLine	5
Webb Simpson	8.5	8.44	26-CutLine	5
Daniel Summerhays	7	12.16	Top25	155

Table 15: BIP team and results for Travelers Championship

5.5. Statistical Analysis

BIP Results (examine weekly points scored by fantasy team)

To put the BIP results into perspective and test for statistical significance, Students t-tests were performed using R. One sample t-tests were performed to present a 95% confidence interval of the means for each tournament (excluding Travelers Championship as data not available). Two-sample t-tests were performed to compare the means of real scores with scores derived via the BIP team.

The null hypothesis states that the difference between the human participants' weekly mean score and the BIP derived weekly score is zero. The alternative hypothesis is that the BIP derived team score is greater than the real participants' mean.

Table 16 shows the confidence interval for the true mean and the BIP derived optimum. The p-value is resulting from the two-sample t-tests.

Tournament	Sample Mean	95% C.I. for Mean	BIP Mean	p-value	Null Hypothesis
Open Championship	903	[783,1024]	1018	0.03128	Rejected
Canadian Open	374	[292,455]	854	$2.2\text{x}10^{-16}$	Rejected
PGA Championship	775	[671,880]	1453	$2.2\text{x}10^{-16}$	Rejected

Table 16: 95% Confidence interval for means

With 95% confidence, the null hypothesis is rejected for all available tournaments at 0.05 significance level, illustrating there is significant evidence to believe the BIP performed better than real life participants. It is worth noting that t-tests were performed to see if the BIP team score was less than the participants' mean and the null hypothesis is not rejected.

Full season scores were extrapolated using the tournament scores above and these were compared to end of season official rankings. Using the official game scores the quartiles and median can be calculated (See Figure 1Figure 7).

- Quartile 1 ≈ 0-7500 points
- Median ≈ 10000 points
- Quartile 3 ≈ 10000-13000 points
- Quartile 4 ≈ 13000-24500 points
- Top 10 Percentile ≈ 17500-24500

Extrapolate sample mean to calculate full season mean score:

$$\frac{(903 + 374 + 775)}{3} * 17 = \ 11628$$

Extrapolate BIP mean to calculate full season score:

$$\frac{1018 + 854 + 1453}{3} * 17 = 18841$$

Using this extrapolation, the full season score will be in the top 10 percentile and outperform the mean. The sample size here is not large enough and so this is not a stringent formulation of full season scores. This can however act as a baseline with which the BIP could be measured against when running for an entire season.

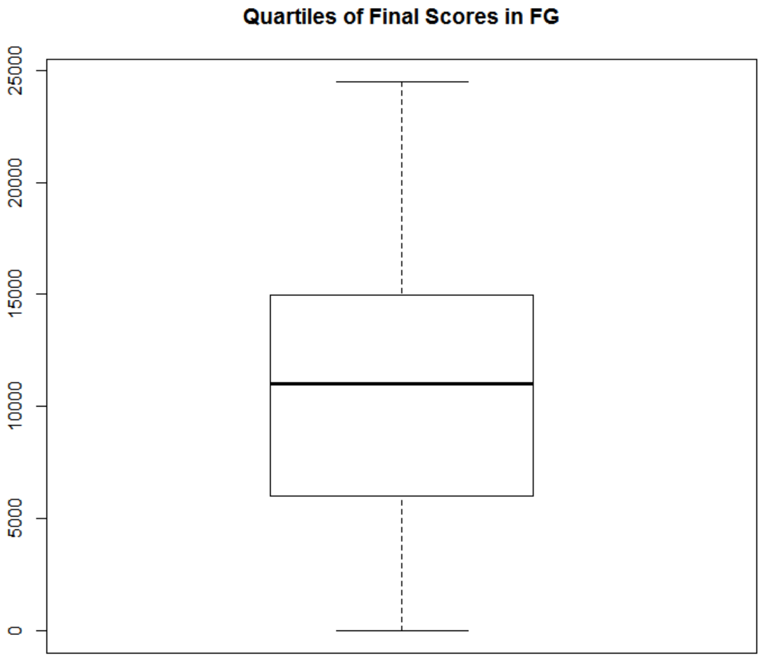


Figure 7: Boxplot of estimated final scores in FG game

GA Results (examine GA parameters)

Although the BIP is used to calculate the definite optimum, the GA did indeed produce the same optimum. To confirm the validity of the GA the parameters of the GA were altered and tested. It was also tested against a random selection to confirm its superiority over random search.

Parameter settings for GA:

- Standard (Crossover = 0.75/Mutation = 0.1)
- Crossover Only (Crossover = 0.75/Mutation = 0.0)
- Mutation Only (Crossover = 0.0/Mutation = 0.1)
- Random (Crossover = 0.0/Mutation = 0.0)

As can be seen in Figure 8, using crossover converges to a single solution every time whereas random search or mutation only do not.

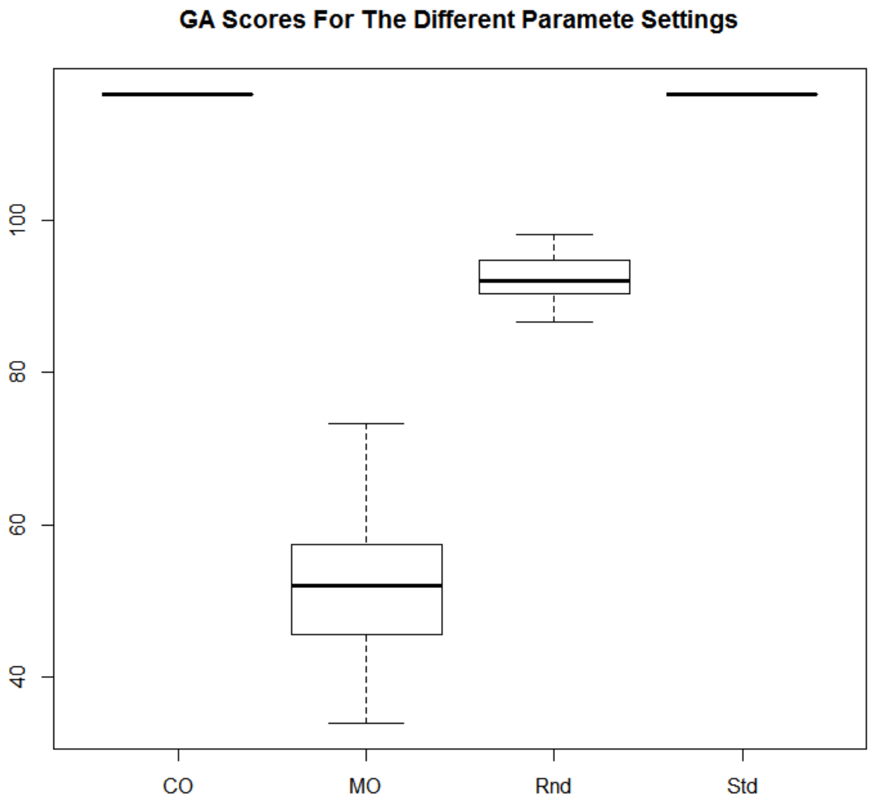
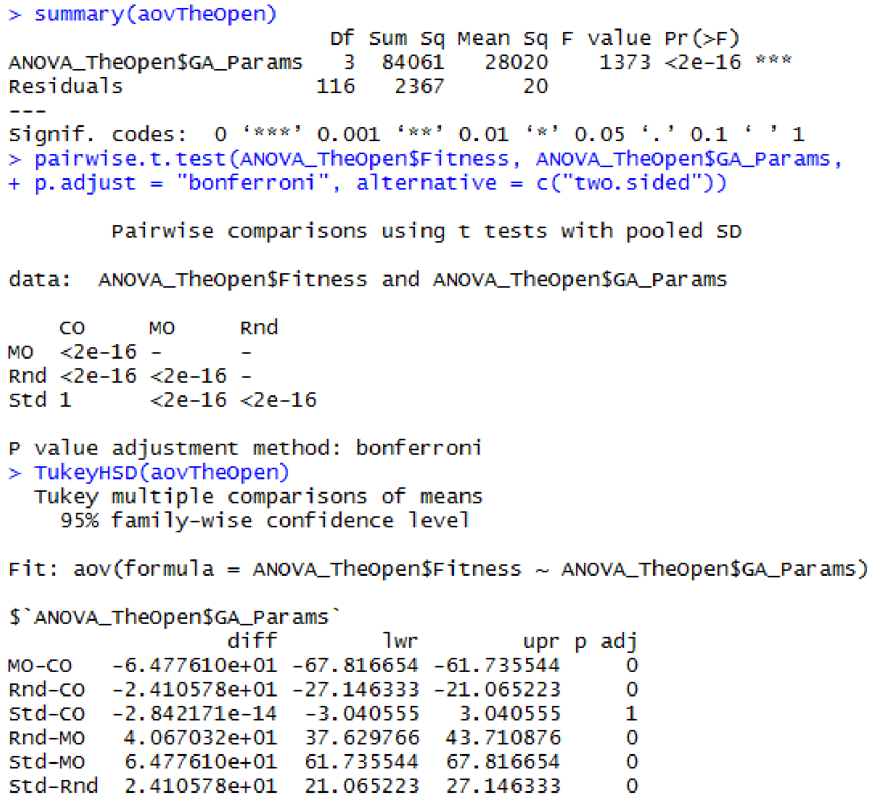


Figure 8: GA scores for different settings



```
> summary(aovTheOpen)
                         Df Sum Sq Mean Sq F value Pr(>F)    
ANOVA_TheOpen$GA_Params   3  84061   28020    1373 <2e-16 ***
Residuals               116   2367      20                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> pairwise.t.test(ANOVA_TheOpen$Fitness, ANOVA_TheOpen$GA_Params,
+ p.adjust = "bonferroni", alternative = c("two.sided"))

        Pairwise comparisons using t tests with pooled SD 

data:  ANOVA_TheOpen$Fitness and ANOVA_TheOpen$GA_Params 

    CO     MO     Rnd   
MO  <2e-16 -      -     
Rnd <2e-16 <2e-16 -     
Std 1      <2e-16 <2e-16

P value adjustment method: bonferroni 
> TukeyHSD(aovTheOpen)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = ANOVA_TheOpen$Fitness ~ ANOVA_TheOpen$GA_Params)

$`ANOVA_TheOpen$GA_Params`
                 diff        lwr        upr p adj
MO-CO   -6.477610e+01 -67.816654 -61.735544     0
Rnd-CO  -2.410578e+01 -27.146333 -21.065223     0
Std-CO  -2.842171e-14  -3.040555   3.040555     1
Rnd-MO   4.067032e+01  37.629766  43.710876     0
Std-MO   6.477610e+01  61.735544  67.816654     0
Std-Rnd  2.410578e+01  21.065223  27.146333     0
```

Figure 9: ANOVA results for variation between parameter settings in GA

The 1-way ANOVA tests in Figure 9 indicates there is no statistically significant difference between the standard use of crossover and mutation, and the use of crossover only. Mutation only and random search are significantly weaker than the standard or crossover only.

5.6. Summary

The best performing features for golf analysis are:

- Strokes Gained Tee-to-Green
- Strokes Gained Putting
- Birdie:Bogey (Ratio)
- Adjusted Average Score
- Average Position (user defined column)
- YTD Fantasy Points (user defined column)
- Course History (user defined column)

A single form indicator used in the constrained single-objective optimisation problem provided better results than the multi-objective formulation. As the optimisation problem had now become a single-objective as opposed to a multi-objective problem, BIP was used as it guarantees an optimum solution. In order to future proof the problem for possible additions of objective functions, the GA was maintained and tested with the single objective. The GA performed equally well and gave the same (optimum) solution for each tournament.

Although only one form indicator was used per tournaments, two options emerged as form indicators depending on which type of tournament the golfer was due to play in next. Equation 8 is the form indicator for the Majors and uses the strokes gained statistics. Equation 9 gives the form indicator for non-Major tournaments and the extra complexity is needed as the players involved are too alike to separate by strokes gained alone.

Major Tournaments:

$$f(x) = SGT2G + SGPutt$$

Equation 8: Form indicator for Majors

Regular PGA events:

$$f(x) = \left[\frac{AvgScore - BirdieBogey - SGT2G - SGPutt}{AvgPos}\right] + \omega_1 FGPoints + \omega_2 CourseHistory$$

Equation 9: Form indicator for non-Major events

Using the form indicator illustrated above, the BIP produced very strong teams for 3 out of the 4 tournaments. The other team matched the mean performance so it was not a weak team. The 3 stronger teams were significantly better than the human generated mean. If the 4 tournaments were extrapolated out to a full season score the team would have finished in the top 10 percentile. This suggests this formula could be used as is in next year's game. It also suggests the BIP should be used to create teams that could be used in Daily Fantasy Golf where a new team is created each week and teams finishing in the top 50 percentile of all teams entered collect winnings. With the BIP team performing significantly better than the mean there is great potential to beat many individual players but this would need to be tested further and verified.

6. Discussion

As the latest research has eluded to, the strokes gained statistics are the most accurate measure of a golfer's performance and this is evident in the data understanding section. Strokes gained also acts as a good predictor when interpreting form. Both the form indicators (Equation 5: Form indicator for Majors and Equation 6: Form indicator for non-Major events) are successful in predicting players who will perform well in the next tournament. In terms of fantasy golf, making the cut is vital and making the top 25 is important for high scoring and almost all of the golfers selected met one or both of these criteria. This suggests that players who are consistent over the season perform to their expected standard more often than not. The accurate form indicator illustrates golfers who are playing well throughout the season continue to play well and golfers who are performing poorly continue to struggle.

There exists a lot of familiarity bias in golf prediction as broadcasters dictate who we see playing well. So much golf is played simultaneously that we cannot actually watch all the players and see who is playing well. An example of this bias came in the 2016 US Masters, arguably the most prestigious tournament in the world. This was won by English man Danny Willet at odds of 145-1. This came as a complete shock to American viewers, many of whom had never heard of him[8]. However, coming into the Masters, Danny Willet was in the top 10 using my form indicator so he was clearly in good form (see Appendix). He was not shown much on television so many American viewers were not aware of him before The Masters. Using the form indicator helps analyse players without bias and is necessary in understanding the form of players fantasy golf participants rarely see live.

The ShotLink data source is an excellent source of data but there are no strokes gained data for some of the tournaments. This is not a fault of ShotLink, it is a matter for some of the tournament organisers who have not agreed to use ShotLink yet. This means the strokes gained moving averages are not 100% accurate reflections of the golfers' seasonal strokes gained numbers. However, the number of tournaments affected is small and because it is a seasonal average it is believed the overall average is not affected greatly unless a golfer performs drastically worse than their mean or significantly better than their mean in the non-ShotLink tournaments.

Generation and training of the form indicator and course history resulted in a single form indicator (single-objective function) performing better than the separation of YTD form and course history (multi-objective function). Clearly, too much weight is given to course history if you try to maximise it. Course history is important but not as important as first thought.

BIP favoured over a GA as it can converge to the optimum solution whereas the GA is not guaranteed. For testing purposes, the GA was used as well as BIP and it did indeed reach the same optimum. BIP used as the solution here because it finds a guaranteed optimum but a GA may potentially be used in future iterations when new data becomes available, e.g. minimising standard deviation over previous 10 weeks could be added as an objective function that tries to observe short-term consistency in the model.

The weekly statistical significance tests are the most important results as the aim of the fantasy golf player in business terms is to finish in the top 50 percentile of DFS cash games. Winning over a full season is difficult and means one must finish in the

[8] Danny Willet shock win, http://www.newser.com/story/223374/in-major-shock-willett-wins-the-masters.html

top fraction of a percentile. Earning a weekly profit is more attainable and is the long term goal of this solution.

7. Conclusions

7.1. Summary

The form indicators deduced are good indicators of past success and future performance. There is significant statistical evidence to suggest that the form indicators discovered can be used in conjunction with binary integer programming to succeed at fantasy golf.

The binary integer programming solution used was simpler than the originally envisaged multi-objective function but the combination of YTD form and course history into a single objective function provided stronger results. The use of a single objective function provides the opportunity for weighting the various features differently if the fantasy golf participant so desires.

7.2. Contribution

Academic Contribution

Strokes gained and other golfing skills like driving and putting have been researched and evaluated to predict end of season earnings or assess previous performance. This paper has introduced the idea of moving averages and combinations of the previously researched statistics to develop a form indicator that can be used to select in-form golfers that are expected to perform well in the upcoming tournament.

Practical Contribution

Using the available ShotLink data and moving averages, the form indicators described here can be used by fantasy golf players to select their teams on a weekly basis. Teams created using the BIP solution could be entered weekly with a significant chance of beating the average player and making a profit.

The form indicator could be used to rank players ignoring budget. Fantasy golf players could use this to select or ignore golfers they do or do not want to include in certain teams. It is common practice for fantasy golf players to enter more than one team weekly and a player may choose to keep or ignore certain golfers in their various teams.

7.3. Future Work

One could investigate the use of other variables as part of the form indicator, e.g. weather, course length, performance on specific grass and form over last “n” weeks.

Future experimentation could try different weights and variables, and proceed to use multiple binary integer programs to produce multiple teams. The results of these teams could then be analysed to investigate the possibility of different form indicators being used under specific circumstances.

In the immediate future, the form indicators described here should be used in conjunction with binary integer programming to enter upcoming fantasy golf tournaments.

Bibliography

Agger, M. (2010). Moneygolf, Will new statistics unlock the secrets of golf? *Slate*, Available at: http://www.slate.com/articles/sports/moneygolf/2010/09/moneygolf.html.

Ahmed, F., Jindal, A., & Deb, K. (2011). Cricket team selection using evolutionary multi-objective optimization. *International Conference on Swarm, Evolutionary, and Memetic Computing*, 71-78.

Alexander, D. L., & Kern, W. (2005). Drive for Show and Putt for Dough? An Analysis of the Earnings of PGA Tour Golfers. *Journal of sports economics, 6*(1), 46-60.

Booker, L. (1987). Improving search in genetic algorithms. *Genetic algorithms and simulated annealing*, 61-73.

Borycki, D. (2011). A strategy in sports betting with the nearest neighbours search and genetic algorithms. *Annales UMCS, Informatica, 11*(1), 7-13.

Boswell, J. (2008). Fantasy sports: A game of skill that is implicitly legal under state law, and now explicitly legal under federal law. *Cardoza Arts & Entertainment Law Journal*, 1257-1278.

Broadie, M. (2008). Assessing Golfer Performance Golfmetrics. *Science and Golf V: Proceedings of the World Scientific Congress of Golf*.

Broadie, M. (2010). *Assessing Golfer Performance on the PGA Tour.*

Byrne, J., O'Neill, M., & Brabazon, A. (2010). *Optimising offensive moves in toribash using a genetic algorithm.*

Callan, S. J., & Thomas, J. M. (2007). Modeling the Determinants of a Professional Golfer's Tournament Earnings A Multiequation Approach. *Journal of sports economics, 8*, 394–411.

Chang, T., Yang, S., & and Chang, K. (2009). Portfolio optimization problems in different risk measures using genetic algorithm. *Expert Systems with Applications, 36*(7), 10529-10537.

Chinneck, J. W. (2015). *Practical Optimization.*

Cochran, A., & Stobbs, J. (1968). *Search for the Perfect Swing: The Proven Scientific Approach to Fundamentally Improving Your Game.* Chicago: Triumph Books.

Davis, L. D. (1991). *Handbook of Genetic Algorithms.*

De Jong, K. (1975). *An Analysis of the Behaviour of a class of genetic Adaptive Systems. Doctorat dissertation.* Dept. of Computer and Communication Sciences, University of Michigan.

Esser, L. (1994). The birth of fantasy football. *Fantasy Football Index*.

Fearing, D., Acimovic, J., & Graves, S. C. (2011). How to Catch a Tiger: Understanding Putting Performance on the PGA TOUR. *Journal of Quantitative Analysis in Sports, 7*(1).

Fernandez, A. J. (2008). *Generating emergent team strategies in football simulation video games via genetic algorithms.*

Fourman, M. P. (1985). Compaction of Symbolic Layout using Genetic Algorithms. *Proceedings of an International Conference on Genetic Algorithms and their Applications.*

Gennaro, V. (2013). *Diamond dollars: The economics of winning in baseball.*

Harville, D. (1980). Predictions for National Football League Games via Linear-Model Methodology. *Journal of the American Statistical Association, 75*(371).

Holland, J. (1975). *Adaptation in Natural and Artificial Systems.*

IrishTimes. (2016). *Fantasy Golf*. Retrieved from Irish Times: https://fantasygolf.irishtimes.com/

James, B. (1977-1988). *Baseball Abstract.*

Lewis, M. (2003). *Moneyball – The Art of Winning an Unfair Game.*

Matthews, T., Ramchurn, S., & Chalkiadakis, G. (2012). *Competing with Humans at Fantasy Football: Team Formation in Large Partially-Observable Domains.*

Nering, E. D., & Tucker, A. W. (1993). *Linear Programs and Related Problems.*

Nero, P. (2001). Relative salary efficiency of PGA tour golfers. *American Economist, 45*(2), 51–56.

Oh, K. J., Kim, T. Y., & Min, S. (2005). Using genetic algorithm to support portfolio optimization for index fund management. *Expert Systems with Applications, 28*, 371–379.

Peters, A. (2008). Determinants of performance on the PGA tour. *Issues in Political Economy*.

Pope, D. G., & Schweitzer, M. E. (2011). Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes. *American Economic Review, 101*(1), 129-157.

Rosenqvista, O., & Skans, O. N. (2015). Confidence enhanced performance? – The causal effects of success on future performance in professional golf tournaments. *Journal of Economic Behavior & Organization, 117*, 281-295.

Sarda, V., & Sakaria, P. (2015). *Football Team Selection Using Genetic Algorithm.*

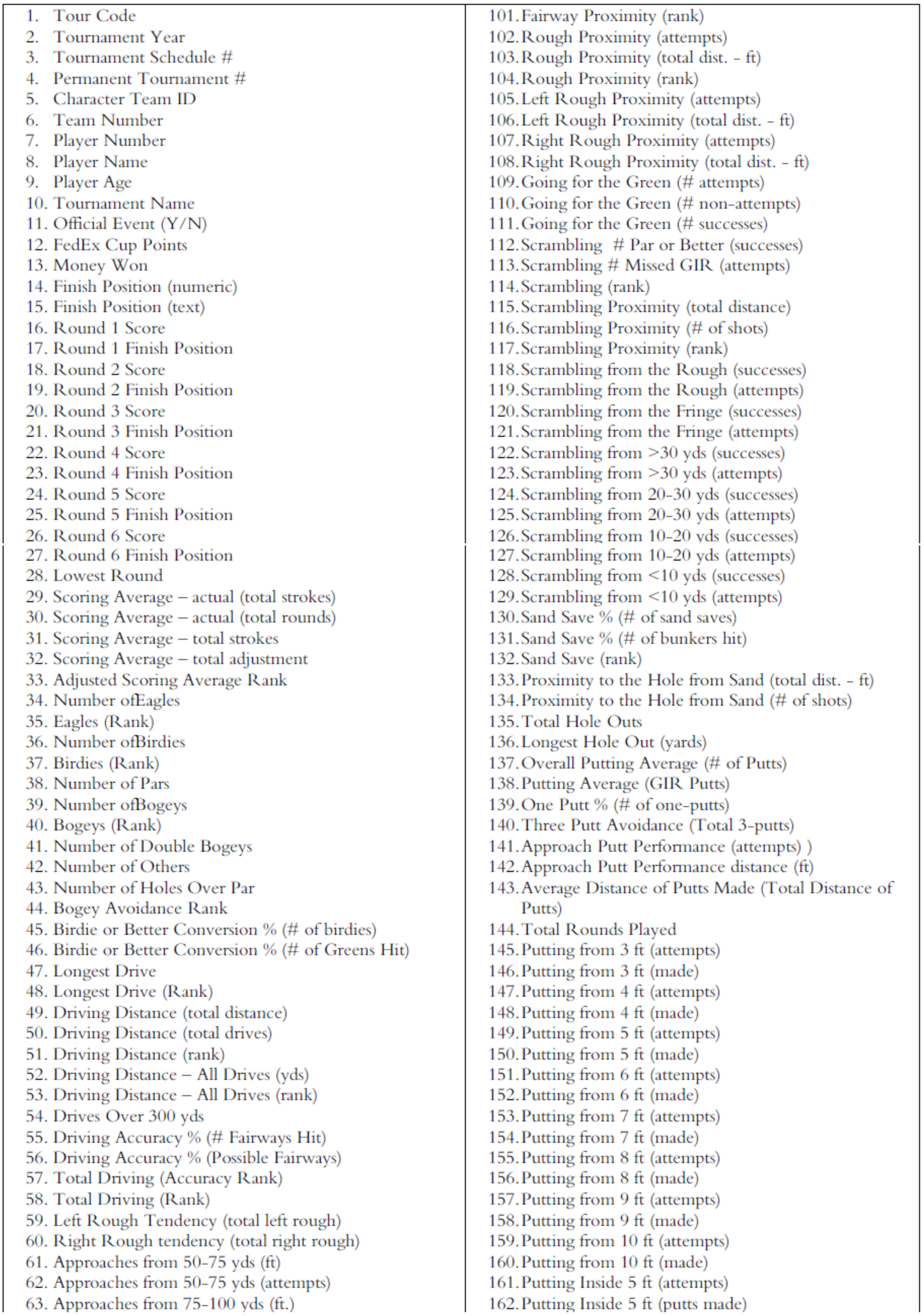
Shmanske, S. (2008). Skills, performance, and earnings in the tournament compensation model: Evidence from PGA Tour microdata. *Journal of Sports Economics*.

Tukey, J. W. (1977). *Exploratory data analysis.*

Xu, B. (2012). *Prediction of Sports Performance based on Genetic Algorithm and Artificial Neural Network.*

Appendix

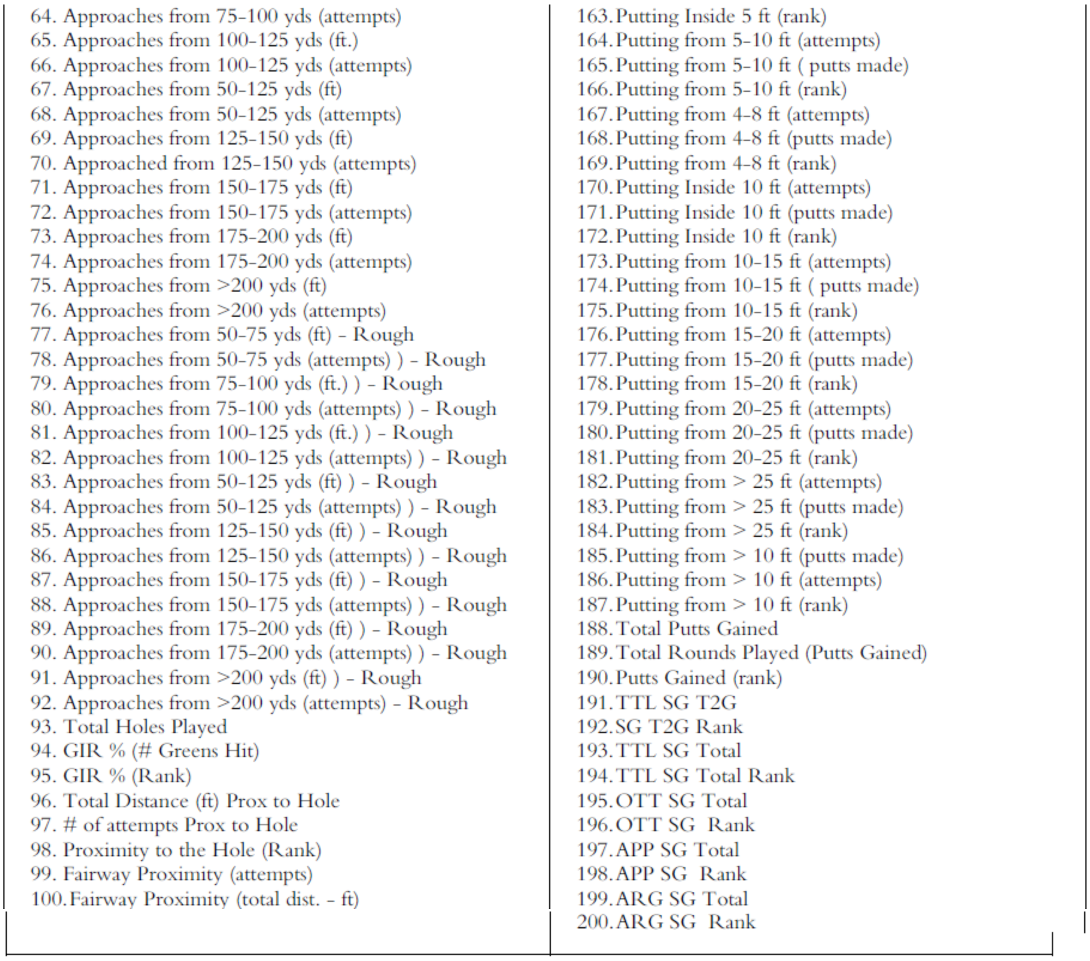
ShotLink Data – Event Level Columns



1. Tour Code
2. Tournament Year
3. Tournament Schedule #
4. Permanent Tournament #
5. Character Team ID
6. Team Number
7. Player Number
8. Player Name
9. Player Age
10. Tournament Name
11. Official Event (Y/N)
12. FedEx Cup Points
13. Money Won
14. Finish Position (numeric)
15. Finish Position (text)
16. Round 1 Score
17. Round 1 Finish Position
18. Round 2 Score
19. Round 2 Finish Position
20. Round 3 Score
21. Round 3 Finish Position
22. Round 4 Score
23. Round 4 Finish Position
24. Round 5 Score
25. Round 5 Finish Position
26. Round 6 Score
27. Round 6 Finish Position
28. Lowest Round
29. Scoring Average – actual (total strokes)
30. Scoring Average – actual (total rounds)
31. Scoring Average – total strokes
32. Scoring Average – total adjustment
33. Adjusted Scoring Average Rank
34. Number ofEagles
35. Eagles (Rank)
36. Number ofBirdies
37. Birdies (Rank)
38. Number of Pars
39. Number ofBogeys
40. Bogeys (Rank)
41. Number of Double Bogeys
42. Number of Others
43. Number of Holes Over Par
44. Bogey Avoidance Rank
45. Birdie or Better Conversion % (# of birdies)
46. Birdie or Better Conversion % (# of Greens Hit)
47. Longest Drive
48. Longest Drive (Rank)
49. Driving Distance (total distance)
50. Driving Distance (total drives)
51. Driving Distance (rank)
52. Driving Distance – All Drives (yds)
53. Driving Distance – All Drives (rank)
54. Drives Over 300 yds
55. Driving Accuracy % (# Fairways Hit)
56. Driving Accuracy % (Possible Fairways)
57. Total Driving (Accuracy Rank)
58. Total Driving (Rank)
59. Left Rough Tendency (total left rough)
60. Right Rough tendency (total right rough)
61. Approaches from 50-75 yds (ft)
62. Approaches from 50-75 yds (attempts)
63. Approaches from 75-100 yds (ft.)

101. Fairway Proximity (rank)
102. Rough Proximity (attempts)
103. Rough Proximity (total dist. - ft)
104. Rough Proximity (rank)
105. Left Rough Proximity (attempts)
106. Left Rough Proximity (total dist. - ft)
107. Right Rough Proximity (attempts)
108. Right Rough Proximity (total dist. - ft)
109. Going for the Green (# attempts)
110. Going for the Green (# non-attempts)
111. Going for the Green (# successes)
112. Scrambling # Par or Better (successes)
113. Scrambling # Missed GIR (attempts)
114. Scrambling (rank)
115. Scrambling Proximity (total distance)
116. Scrambling Proximity (# of shots)
117. Scrambling Proximity (rank)
118. Scrambling from the Rough (successes)
119. Scrambling from the Rough (attempts)
120. Scrambling from the Fringe (successes)
121. Scrambling from the Fringe (attempts)
122. Scrambling from >30 yds (successes)
123. Scrambling from >30 yds (attempts)
124. Scrambling from 20-30 yds (successes)
125. Scrambling from 20-30 yds (attempts)
126. Scrambling from 10-20 yds (successes)
127. Scrambling from 10-20 yds (attempts)
128. Scrambling from <10 yds (successes)
129. Scrambling from <10 yds (attempts)
130. Sand Save % (# of sand saves)
131. Sand Save % (# of bunkers hit)
132. Sand Save (rank)
133. Proximity to the Hole from Sand (total dist. - ft)
134. Proximity to the Hole from Sand (# of shots)
135. Total Hole Outs
136. Longest Hole Out (yards)
137. Overall Putting Average (# of Putts)
138. Putting Average (GIR Putts)
139. One Putt % (# of one-putts)
140. Three Putt Avoidance (Total 3-putts)
141. Approach Putt Performance (attempts))
142. Approach Putt Performance distance (ft)
143. Average Distance of Putts Made (Total Distance of Putts)
144. Total Rounds Played
145. Putting from 3 ft (attempts)
146. Putting from 3 ft (made)
147. Putting from 4 ft (attempts)
148. Putting from 4 ft (made)
149. Putting from 5 ft (attempts)
150. Putting from 5 ft (made)
151. Putting from 6 ft (attempts)
152. Putting from 6 ft (made)
153. Putting from 7 ft (attempts)
154. Putting from 7 ft (made)
155. Putting from 8 ft (attempts)
156. Putting from 8 ft (made)
157. Putting from 9 ft (attempts)
158. Putting from 9 ft (made)
159. Putting from 10 ft (attempts)
160. Putting from 10 ft (made)
161. Putting Inside 5 ft (attempts)
162. Putting Inside 5 ft (putts made)

Appendix



64. Approaches from 75-100 yds (attempts)
65. Approaches from 100-125 yds (ft.)
66. Approaches from 100-125 yds (attempts)
67. Approaches from 50-125 yds (ft)
68. Approaches from 50-125 yds (attempts)
69. Approaches from 125-150 yds (ft)
70. Approached from 125-150 yds (attempts)
71. Approaches from 150-175 yds (ft)
72. Approaches from 150-175 yds (attempts)
73. Approaches from 175-200 yds (ft)
74. Approaches from 175-200 yds (attempts)
75. Approaches from >200 yds (ft)
76. Approaches from >200 yds (attempts)
77. Approaches from 50-75 yds (ft) - Rough
78. Approaches from 50-75 yds (attempts)) - Rough
79. Approaches from 75-100 yds (ft.)) - Rough
80. Approaches from 75-100 yds (attempts)) - Rough
81. Approaches from 100-125 yds (ft.)) - Rough
82. Approaches from 100-125 yds (attempts)) - Rough
83. Approaches from 50-125 yds (ft)) - Rough
84. Approaches from 50-125 yds (attempts)) - Rough
85. Approaches from 125-150 yds (ft)) - Rough
86. Approaches from 125-150 yds (attempts)) - Rough
87. Approaches from 150-175 yds (ft)) - Rough
88. Approaches from 150-175 yds (attempts)) - Rough
89. Approaches from 175-200 yds (ft)) - Rough
90. Approaches from 175-200 yds (attempts)) - Rough
91. Approaches from >200 yds (ft)) - Rough
92. Approaches from >200 yds (attempts) - Rough
93. Total Holes Played
94. GIR % (# Greens Hit)
95. GIR % (Rank)
96. Total Distance (ft) Prox to Hole
97. # of attempts Prox to Hole
98. Proximity to the Hole (Rank)
99. Fairway Proximity (attempts)
100. Fairway Proximity (total dist. - ft)
163. Putting Inside 5 ft (rank)
164. Putting from 5-10 ft (attempts)
165. Putting from 5-10 ft (putts made)
166. Putting from 5-10 ft (rank)
167. Putting from 4-8 ft (attempts)
168. Putting from 4-8 ft (putts made)
169. Putting from 4-8 ft (rank)
170. Putting Inside 10 ft (attempts)
171. Putting Inside 10 ft (putts made)
172. Putting Inside 10 ft (rank)
173. Putting from 10-15 ft (attempts)
174. Putting from 10-15 ft (putts made)
175. Putting from 10-15 ft (rank)
176. Putting from 15-20 ft (attempts)
177. Putting from 15-20 ft (putts made)
178. Putting from 15-20 ft (rank)
179. Putting from 20-25 ft (attempts)
180. Putting from 20-25 ft (putts made)
181. Putting from 20-25 ft (rank)
182. Putting from > 25 ft (attempts)
183. Putting from > 25 ft (putts made)
184. Putting from > 25 ft (rank)
185. Putting from > 10 ft (putts made)
186. Putting from > 10 ft (attempts)
187. Putting from > 10 ft (rank)
188. Total Putts Gained
189. Total Rounds Played (Putts Gained)
190. Putts Gained (rank)
191. TTL SG T2G
192. SG T2G Rank
193. TTL SG Total
194. TTL SG Total Rank
195. OTT SG Total
196. OTT SG Rank
197. APP SG Total
198. APP SG Rank
199. ARG SG Total
200. ARG SG Rank

Appendix

Summary Reports of Top 10 Form Golfers

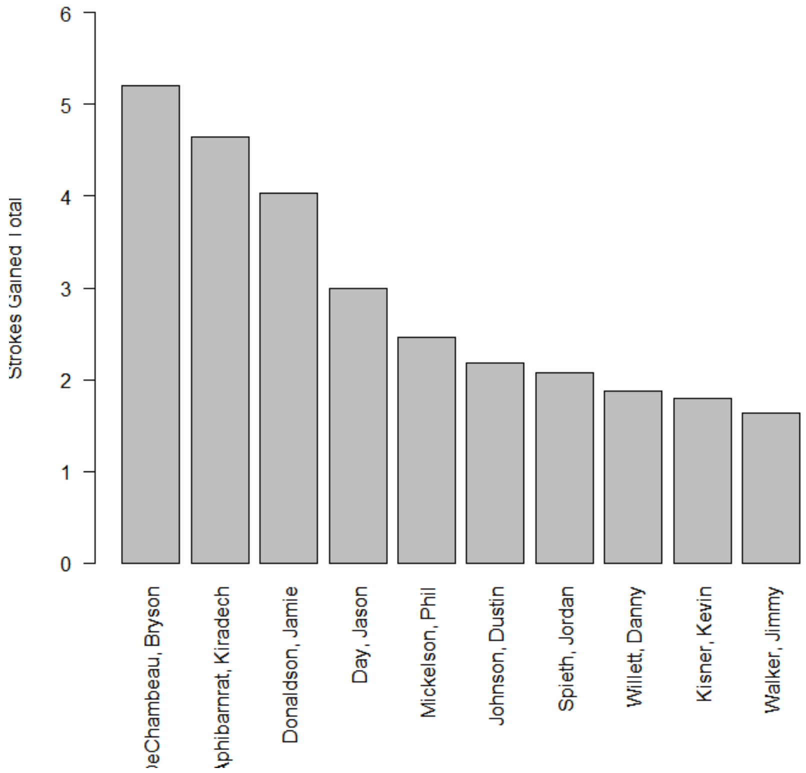


Figure 10: Top 10 form players pre Masters

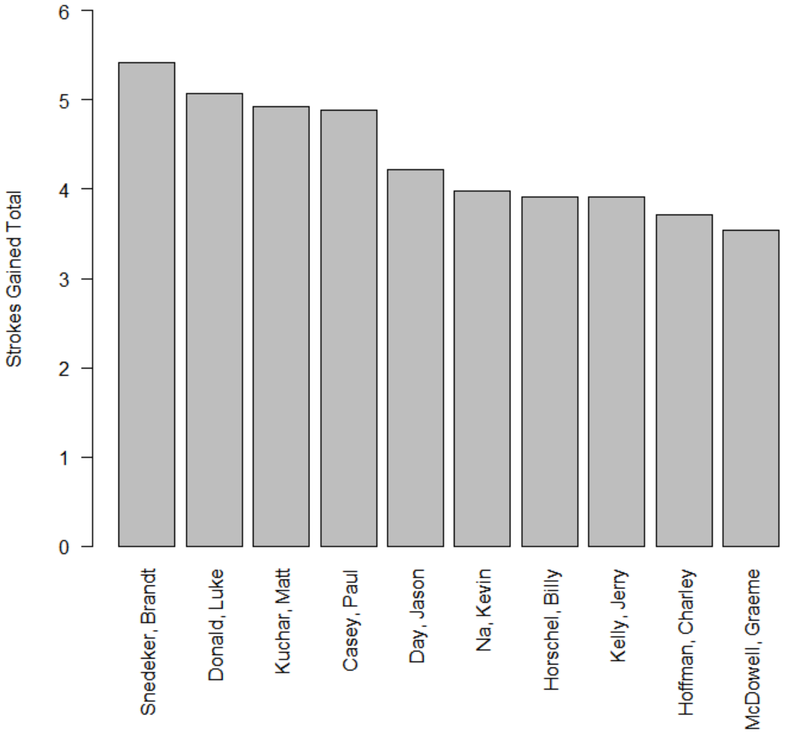


Figure 11: Top 10 form players pre Heritage

Appendix

GitHub Link – Code and Data

https://github.com/ShaneRooney