2014-07-15

Predicting Professional Golfer Performance Using Proprietary PGA Tour “Shotlink” Data

Brian Leahy
Technological University Dublin

Recommended Citation

Leahy, B. (2014) *Predicting Professional Golfer Performance Using Proprietary PGA Tour “Shotlink” Data*. A dissertation submitted in partial fulfilment of the requirements of Technological University Dublin for the degree of M.Sc. in Computing (Data Analytics) July 2014.

Predicting Professional Golfer Performance Using Proprietary PGA Tour “Shotlink” Data

Brian Leahy

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Data Analytics)

July 2014

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: ______________________________

Date: ***14th July 2014***

1 ABSTRACT

It can cost a minimum of $110,000 a year for a professional golfer to compete on the PGA Tour. For the successful golfers who earn millions every year, this is not a problem. For those lower ranked golfers, it is a problem. This is due to the fact that almost half the golfers who compete in any one PGA Tour tournament will not get paid, because they have missed the dreaded cut. When a golfer begins to consistently miss the cut, they can come under financial pressure which may manifest itself into poor further tournament performances.

This dissertation attempts to aid these less successful golfers by developing models which will predict the likelihood of a professional golfer missing the cut or not. By using the prediction provided by these models, a golfer could then decide not compete in a tournament and so save money on travel expenses and support staff. They could then practice their golf game, working on the aspects of their skills that the model has suggested.

Additionally, the dissertation will attempt to answer a number of questions surrounding the influence of external factors on a golfer’s performance using statistical inference.

Key words: Predictive Analytics, Golf, Shotlink, Caddy Changes, Golfer Similarity

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my project supervisor, Dr. Michael Collins. Without his continual guidance, patience, advice and motivation, this dissertation would not exist. Thanks also to Aoife D'Arcy and Brendan Tierney for their help and advice.

For the creation of the vast and rich dataset that is ShotLink^{TM}, and for granting me the privilege of using it for this project, my appreciation and thanks go to Steve Evans (Head of IS), Adam Mersereau (Legal), Ty Votaw (CMO), Ken Lovell & Royce Thompson, all from the PGA Tour.

Thanks to Jamie for being a listening post in addition to proof reading the document and also to Paul for the thorough proof reading.

Finally, special thanks have to be given to my wife Louisa whose constant support has allowed me to partake in this M.Sc. in the first place. A further special mention also has to go to my son Diarmuid, whose unbridled laughter has kept me sane.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

TABLE OF ACRONYMNS

Term	Acronym
Analysis of Variance	ANOVA
Greens in Regulation	GIR
Graphical User Interface	GUI
Professional Golfers Association	PGA
Receiver Operating Characteristic	ROC
Royal & Ancient	R&A
United States Golf Association	USGA

1. INTRODUCTION

"We'd like to know a little bit
about you for our files"
Simon & Garfunkel 1968[1]

1.1 Background

To some, the world of professional golf could be seen as belonging to the elite – those of certain privilege. The stereotype of a professional golfer is that of someone who had considerable financial backing on their journey to the top of the game. However, when the surface is scratched, it can be seen that in many cases up-and-coming golfers enter the ranks of professionalism with considerable financial burden, expectation and pressure.

In return for financial backing to play in college tournaments and the lower tier competitions, the golfer typically has to give a portion of their future earnings to those backers for an agreed period of time. In essence, what the public views or is predominately aware of is the wealthy pinnacle of the sport, where the winners of tournaments can earn millions of dollars and even those in 70th place can earn substantial amounts.

One of the problems of a professional golfer (especially lower ranked) is that if they miss the cut for a given tournament, they do not get paid. Exceptions to this rule are for those fortunate golfers who obtain an appearance fee, just for turning up, so missing the cut brings no financial risk.

According to (Noer, 2012), it costs a minimum of $110,000 to compete for a year on the PGA Tour. Even if a player misses the cut, they still incur travel expenses and caddie fees.

[1] From the song "Mrs. Robinson" written by Paul Simon

Table 1.1 presents an excerpt from the “Money Per Event Leaders” statistics for year-end 2012, on the PGA Tour’s website.

Player	Events	Average Earnings per event (2012)	Total earning for 2012
Biershenk, Tommy	27	3,972	107,266
Loar, Edward	23	4,258	97,946
Gangluff, Stephen	23	2,552	58,702
Thompson, Kyle	22	2,066	45,460
Dawson, Marco	22	3,623	79,727
Bertsch, Shane	17	4,340	73,795
Duval, David	17	1,937	32,936
Lovemark, Jamie	16	6,848	109,571
Glover, Lucas	16	4,194	67,111
Willis, Garrett	15	6,334	95,014

Table 1.1: Sample of professional golfers who earned less than $110,000 in 2012

This table represents just a small sample of the players who earned under the $110,000 mark, so they didn’t break even for the year. None of these players could expect to keep their Tour Card for 2013 so the financial implications are even harsher as the Web.com (second tier tour) pays typically only 11% of the prize money of the main tour.

In 2003, the United States PGA Tour (hereafter referred to as the “PGA Tour”) started collecting data in earnest regarding the minutest details of every golf shot taken by every player on all of their official tours. This dataset is known as “Shotlink” data and was opened up to the academic community with the intention / hope of producing insight along the ilk of what SABERMETRICS[2] has achieved so famously in baseball.

The provision of this Shotlink data represents an immense opportunity to provide or create real value, not just from an academic standpoint but also from the view of the professional golfer who could benefit from an intelligent decision making tool.

[2] SABERMETRICS is the term commonly used to describe the analysis of baseball statistics

1.2 Research problem

In the years since the Shotlink dataset has been made available to the academic community, it appears that the scope of research carried out using the data has been of a narrow focus. Examples include the creation of new statistics, or creating new ranking lists for events that have already happened in previous seasons. Some regression studies have been undertaken but these did not use the Shotlink dataset.

While this research is interesting and has merit, none of it has attempted to use Data Analytics or Machine Learning techniques to perform predictive analytics. The key research problem of this dissertation is to assess if the application of Data Analytical techniques to the Shotlink data can predict the performance of professional golfers. A secondary problem revolves around the suitability of the Shotlink data for statistical analysis to answer specific claims of golfer comparisons and performances.

1.3 Intellectual challenge

There are a number of challenges that this project will face:

"The Curse of Dimensionality"

The number of columns for each detail level available from the Shotlink dataset, in addition to the number of rows for one year's worth of data is outlined in Table 1.2.

Detail Level	No of Columns	No of Rows for 2013
Stroke	38	942,438
Hole	50	288,343
Round	173	153,96
Event	190	5,310
Radar Launch	50	13,729
Radar Trajectory	56	532,319

Table 1.2: Shotlink Dataset Levels (Shotlink, 2014)

As can be seen, there is a high level of detail, especially at "Round" and "Event" levels. This can lead to what is known as "The Curse of Dimensionality". In essence this means that the more columns you have, the more rows of data you will need in order to for a machine learning model to be effective. In Table 1.2 it can be seen that the levels with the most number of rows have the least number of columns, and vice versa.

To account for this high level of dimensionality, in order to create effective, flexible models, it will be necessary to employ some dimension reduction techniques. These include Principal Components Analysis (PCA) and correlation analysis to determine which columns (also known as features or attributes) of the data are important to the particular question posed, at an Event and Round level.

Performance Prediction

Part of the challenge of this project will be to create suitable metrics that accurately describe an association which is pertinent to the problem to be solved, for example predicting if a golfer will miss the cut or not. This will be a significant part of the experiment and the identification of key metrics, either derived or transformed from the source data, or the original data items, will be the key to success of the experiments.

Previous Research

The use of previous research, both directly and indirectly related to the Shotlink dataset, to form opinion and intelligently influence the direction of experimentation, will have a substantial bearing on the project. As will be outlined in coming chapters, most of the research based on Shotlink data does not use data analytics to aid in its conclusions. Data Mining is referenced in some of the research but not in its correct context, that of using Machine Learning algorithms to extract potentially useful information from data.

Research into the use of data analysis in baseball will be undertaken in order to supplement the thought process and the direction of the project. Diversification of research into another sport will ensure that lessons learned in other sporting disciplines will be considered for their usefulness with respect to the project. They will then be scrutinised from a golfing perspective, in line with the data available to determine if they influence any models.

Psychology

Another of the main challenges will be the creation of successful models that may not be able to take into account any psychological aspects of performance. It is theorised that by measuring average past performance, it will smooth out any “bad days” (outliers) in a golfer’s performance. However, it will not be possible to incorporate the likelihood of a golfer having once-off issues, or possibly, the beginning of ongoing psychological / off-the-course pressures such as those experienced by Rory McIlroy in the 2013 season.

Interpretability

The final challenge to this project will be its ability to explain the decision for each prediction question posed, for example it is determined that a golfer will miss the cut at a given tournament. In the real world, it would be necessary to outline to the golfer the reasons why the model came to the specific conclusion. Depending on the most accurate model, it may not be possible to easily explain why a golfer may miss the cut. This would be the case with a Neural Network model, for example.

1.4 Research objectives

The objectives of this project are twofold. The first objective is to derive value and insight from the use of Machine Learning techniques in order to predict professional golfer performance - specifically whether or not a golfer will miss the cut.

If achieved, this objective would allow golfers to make an informed decision as to which tournaments or courses to play. This may help alleviate the financial burden of golfers such as Tommy Biershenk who missed the cut 17 times out of 27 events in 2012 (63% of events played).

The second objective is to determine what data if any, can be used from the Shotlink dataset to test various hypotheses on golfer performance using statistical analysis.

These objectives will be achieved via completion of the following milestones:

- Review of previous research from a broad range of golf related areas.
- Prepare and transform the Shotlink data for use.
- Generate “Golfer Analytical Records” from which models will be trained.

- Design and build predictive models.
- Design and implement statistical experiments.
- Evaluate the success or otherwise of the models and experiments deriving insight from the results.

1.5 Contribution to the body of knowledge

This project will contribute to the body of knowledge by examining the use of Data Analysis techniques to test the hypothesis that the performance of a professional golfer can be accurately predicted, using both the collective and individual past performances of all golfers.

A further contribution will be analysing the performance of particular individual golfer's for comparative purposes, and to determine if external influences on performance can be measured and tracked.

Specifically, the research presented in this project will contribute to the following topics:

1. Cut Line Prediction.

This research will determine whether it is possible to predict the cut line for a given professional golf tournament. Results from this research could be used to inform a golfer if they should enter the tournament or just go and practice for the next one.

2. Determination of the influence of a caddy on a golfer's performance.

Caddies often have an understated reputation for the abilities, value and service they provide to their employers, professional golfers. This research will examine the Shotlink data to determine if there is any evidence to suggest that caddies do indeed have a significant influence on their employer's performance.

3. Comparison of Rory McIlroy & Jordan Spieth.

During the course of the 2013 and 2014 seasons Jordan Spieth has been regularly compared to Rory McIlroy due to the similarity in performances when McIlroy was the same age. This research will focus on the these two golfers to ascertain if there is enough evidence to truly warrant the comparison between McIlory, who has won two majors and Spieth who has won just one regular PGA Tour tournament.

4. Analysis of the age profile of a professional golfer and their ability to make the cut.

Unlike other professional sports such as basketball, soccer, or tennis, golf is not as physically demanding. However, age does have an impact on a golfer's ability to compete to win, but does it also affect a golfer's ability to make the cut? This research will analyse the cut line with respect to the age of golfers in order to discover if there is a statistical significance or not.

1.6 Research Methodology

The research methodology includes:

- Review of previous research to guide project direction
- Develop an understanding of the data, including analysis of data quality and statistical analysis
- Data Cleansing & Transformation. Dealing with missing values and deriving valuable metrics
- Development of a "business" understanding. Translation of the more unfamiliar golfing terms and statistics to determine their real merit.
- Evaluation of all the previous steps undertaken in the experiment and wider project to determine success or otherwise of the various stages.

1.7 Resources

In order to achieve the goals of this dissertation, the following technical and non-technical resources were identified and acquired:

Technical Resources

- Dell XPS Workstation: Intel Core i7-4770 CPU @ 3.4Ghz, 24GB RAM, running Windows 8.1
- Shotlink Dataset
- SAS Enterprise Guide 5.1

- SAS Enterprise Miner 12.1
- MySQL Server 5.6.16
- Teradata SQL Assistant 14.10.0.02
- MS Word / Excel 2010
- HP Officejet 6500A Plus Inkjet Printer
- Multiple Back-up Devices
- Internet Access

Non-Technical Resource Requirements

- Library Access
- Project Supervisor
- PGA Tour point of contact

1.8 Scope and Limitations

One of the limitations to this project is that there is limited research published that relates specifically to the use of Shotlink data. There also appears to be very little research published in the realm of golfing with relation to value-added statistics. There are some websites which offer "Golf Betting Systems" or golf prediction (golfprediction.com). These sites do not outline their methodologies publically but if these were truly successful the bookmakers' profits would be markedly different.

The vast number of dimensions in the data may also prove to be a limitation to this project. As mentioned previously, the "curse of dimensionality" could have an adverse impact on any models ability to correctly predict golfer performance. Intelligent reduction of the number of dimensions by employing various techniques such as multi-collinearity and Principal Components Analysis may need to be used in order to mitigate or remove this limitation.

Professional golfers with dual tour membership, i.e. those who compete on both the PGA Tour & the European Tour could pose a problem. For this project, data is not available to the same level of detail for European Tour events. This means that any models built may miss vital data if a golfer has been playing a number of European Tour events prior to competing in a PGA Tour event.

The scope of this project will be confined to the prediction of golfer performance and what variables (derived or otherwise) most influence the predictability of this performance. It will examine the factors which possibly affect a golfer's chance of making the cut and identification of streaks or dips in form which can influence their ability to make the cut. It will not attempt to predict specific issues such as which golfer is likely to win a tournament.

Psychological influences on golfers will be investigated to determine if there is any potential to incorporate findings from these studies with the available data in Shotlink. It is expected however that given the time scale of the project it may only be possible to incorporate the findings at a basic level.

External datasets will not be appended to the Shotlink dataset but limited external data such as the details pertaining to the timings of caddy changes will be used.

1.9 Organisation of the Dissertation

Chapter Two will provide an introduction to the sport of Golf and the Shotlink Dataset. This will include an explanation of any terminology which is used throughout this project.

Chapter Three will review some of the previous research undertaken in the study of the game of Golf and discuss how it is pertinent to the project. Chapter Four will detail the classification techniques and statistical methods used in the project.

Chapter Five will outline the experimentation design of this project, while Chapter Six will present the results of these experiments. Chapter Seven will evaluate the project and this will include an evaluation of the experiment design and an interpretation of the results, with emphasis on the stated goals of the project.

Finally, Chapter Eight will discuss the conclusions of the project, while summarising areas where future research could be continued.

2. AN INTRODUCTION TO GOLF

This chapter will provide an introduction to the game of golf, its rules, terminology, its governing bodies and how professionalism is organised. It will also provide an introduction to the Shotlink dataset, detailing the various levels of detail available for analysis.

This information is intended to provide the reader with a level of familiarity with the game of golf and its intricacies, which in turn will give the reader a thorough comprehension of terms used throughout the remainder of the project.

2.1 A brief overview of Golf, its lexicon and its organisations

According to the Royal & Ancient (R&A), the governing body which sets the rules of golf (outside of the USA), golf is *"... a game in which a player, using a club, tries to hit a small, round ball into a small, round hole (also known as the "cup") in as few shots as possible."*[3]

Golf is played outdoors on a golf course. A golf course typically consists of 18 holes, though 9 hole courses exist where each hole is played twice. A player has completed a round of golf when they have played all 18 holes.

Each hole is given a notional score called "Par" which can be considered to be the typical number of strokes (the number of times a golfer hits the ball with a club) that a good player should take in order to complete the hole. If a player takes more strokes than par for the hole, they have completed the hole "over par". Similarly, if they complete the hole in less stokes than par they have completed the hole "under par".

All holes on a golf course will have par scores of 3, 4 or 5 strokes, which typically indicate the length of the hole - the lower the par score, the shorter the distance between the tee box and the front of the green.

[3] http://www.randa.org/en/Playing-Golf.aspx

A player’s score for each hole is aggregated for the total round score. So, on a golf course with a total par of 72, if a player has an aggregated round score of 70, they played the course “two under par”.

2.1.1 Playing formats

There are a number of different formats for playing golf with Stroke Play and Match Play being two of the more popular and well known.

- Stroke Play

Stroke play consists of a field of players who compete over a number of rounds. The winner is the player with lowest aggregate score after all rounds have been played. The focus of this project will be based on stroke play competitions.

- Match Play

Match Play is when players compete directly against each other, in a knockout style competition. Players play to win each hole with the winner being the player who has won the most number of holes in the round.

2.1.2 Scoring Terms of golf

- Birdie

When a player achieves a score of one under par on a hole, it is called a “Birdie”. For example, a golfer takes 3 strokes to complete a par 4 hole.

- Eagle

When a player achieves a score of two under par on a hole, it is called an “Eagle”. For example, a golfer takes 3 strokes to complete a par 5 hole.

- Bogie

When a player achieves a score of one over par on a hole, it is called a “Bogie”. For example, a golfer takes 5 strokes to complete a par 4 hole.

- Other scoring terms

An albatross (also known as a double eagle in the USA) is a score of 3 under par on a given hole, and is quite a rare score. An example would be when a golfer takes 2 strokes to complete a par 5 hole. A famous example of an albatross was in the final round of the 2012 US Masters when Louis Oosthuizen achieved the score on the par 5 2nd hole (Lamport-Stokes, 2012).

A double bogie is a score of 2 over par on a given hole and is not as rare as an Albatross. An example would be when a golfer takes 7 strokes to complete a par 5 hole.

2.1.3 Other terms

- Greens in Regulation

One of the statistics used to measure the performance of a golfer is called "Greens in Regulation". A player hits a green in regulation if the golf ball reaches the green in two strokes on a par 4 hole, three strokes on a par 5 hole, or 1 stroke on a par 3 hole (two strokes under par).

- Scrambling

Scrambling is when a player misses the green in regulation but still completes the hole with a score of par or better.

- Approach Shot

This is a shot from the fairway towards the green.

- "Holing Out"

A golfer "holes out" when their shot commences off the green and the ball finishes in the cup.

- Adjusted Weighted Score

A measure of a golfer's performance relative to the competition.

- Caddie

A caddie is a person who carries the golfer's bag and clubs. Part of their job is to provide insight and guidance to the golfer so that amongst others things, the correct club is chosen for any given shot. This is based on the both the caddie's and player's knowledge of the player's particular ability to hit the club to a given distance. As (Huguenin, 2013) notes, caddies can also be the voice of reason and help to calm a golfer's nerves if the pressure of a situation is being to get the better of them.

- The Cut

In most 72-hole tournaments, after the first two rounds are completed, there is a "cut" that reduces the number of golfers who will continue to play in the remaining rounds of the tournament. A cut rule is used to determined which golfers will continue to play and can vary from one tournament to the next. An example of a cut rule would be that the Top 65 golfers including those tied at 65th position after 2 rounds will continue to play the remaining rounds.

The cut is quite significant, especially to professional golfers. If a professional golfer misses the cut so that they fail to progress to rounds 3 and 4, they get no share of the prize money. They still incur the cost of travel to the tournament and the expenses of paying their caddie and other support staff.

- The Handicap System

In amateur golf, players are given a "handicap" based on their level of ability. This allows players of all levels of ability to compete at the same level. For example, if a player has a handicap of 13, this means that they can deduct 13 strokes from their final gross score (in a stroke play event) in order to derive their net score. The net score is the determinate for whom the winner of an amateur completion is. As a player's game improves, or deteriorates, their handicap will fall or rise accordingly. Professional golfers do not have handicaps.

2.1.4 Governing Bodies & Professional Tours

There are two governing bodies which set the rules for golf. The United States Golf Association (USGA) who govern the game in the United States and Mexico, and the Royal & Ancient (R&A), who govern the game elsewhere in the world. The two bodies have been jointly issuing the rules of golf since 1952[4].

The game of golf has two groups of players: Amateur & Professional. Amateur golfers do not make a living from playing golf, though they can win prizes below a certain financial value, for example in club competitions. Professional golfers try to make their living from golf, and if they are even moderately successful, the financial reward can be immense. Professional golfers will be the subject matter for this project.

There are a number of professional golfer organisations around the world which organise tournaments for professional golfers. The two wealthiest – i.e. those whose tournaments pay the most money in prizes are the PGA Tour & the European Tour. The most prestigious tournaments in golf, known as the "majors", are not organised by either tour, but winnings from these tournaments count towards the money lists (described below) on each tour.

- **The PGA Tour**

The US PGA Tour is the organisation which runs most of the male professional golf tournament events in the United States. It organises 3 tiers of events for professionals. Its flagship series of events is simply known as the "PGA Tour" and for the 2013 season, had 36 tournaments organised (excluding the major championships). If a golfer has membership of the PGA Tour (known as the card), they are considered to be at the pinnacle of professional golf.

The Web.com tour is a "feeder" tour to the PGA Tour. Golfers who play on the Web.com tour would be those of lower rank and are competing for one of fifty PGA Tour cards available to the best players on the tour at the end of a season. Members of

[4] http://www.randa.org/en/RandA.aspx

the Web.com tour can still earn a decent living, though the average prize money per event is typically only 11% of that on offer for a PGA Tour tournament.

The Champions tour is for golfers of 50 years of age and over. Tournaments are typically played over 3 rounds rather than 4, reflecting the physical ability and age profile of the competitors. For 2013, the average prize money per event was 30% of that on offer for a PGA Tour tournament. Table 2.1 shows the average prize money for each of the tours in 2013.

Tour Name	Number of Tournaments (2013)	Average Prize Fund (per tournament)
PGA Tour	36	$6,401,017
Champions Tour	26	$1,952,077
Web.com Tour	25	$715,948

Table 2.1: Average prize money by PGA sanctioned tour (Shotlink, 2014)

- **The European Tour**

The European Tour has a similar structure to the PGA Tour's. It has 3 tiers of tournaments, the main European Tour, which is also the wealthiest of its tours, the Senior Tour which is the equivalent of the Champions Tour and the Challenge Tour which is the equivalent of the Web.com, the developmental tour for lower ranked golfers aspiring to join the main European Tour.

Prize money for European Tour tournaments is on average 47% of the money on offer when compared to the PGA Tour average. Table 2.2 outlines prize money on offer for European Tour events in more detail.

Tour Name	Number of Tournaments (2013)	Average Prize Fund (per tournament)
European Tour	37	$2,976,504
European Senior Tour	13	$481,636
Challenge Tour	25	$311,715

Table 2.2: Average prize money by European Tour sanctioned tours (europeantour.com)

- **The Money Lists**

To many, if not all professional golfers, their main objective is to win tournaments and earn substantial reward from prize money and appearance fees. As a by-product of their success, they will hope to gain sponsorship from rich corporations who will pay them to use their golfing equipment and wear their clothing ranges.

Throughout each season, on both sides of the Atlantic, as tournaments are played and prizes are won, professional golfers are ranked on the "money list", with the golfer who has won the most money ranked 1st.

As the season draws to a close, there are number of tournaments where the prize fund is substantially higher than those on regular tour events. In the US, these are known as the FedEx Cup Playoffs, a series of four tournaments where the entry list is reduced in each competition, based on the number of FedEx Cup points a golfer has accumulated over the season. The first of these playoff tournaments has 125 players, the last has only 30.

At the end of the season, once the final playoff tournament has concluded, the player ranked first is awarded the "Fedex Cup" and $10 million in the US. In Europe, it is called the "Race to Dubai" with $1.5 million awarded to its winner.

The Race to Dubai has a similar progressive cut over its final playoff tournaments, but it has a field of 60 for its season concluding event.

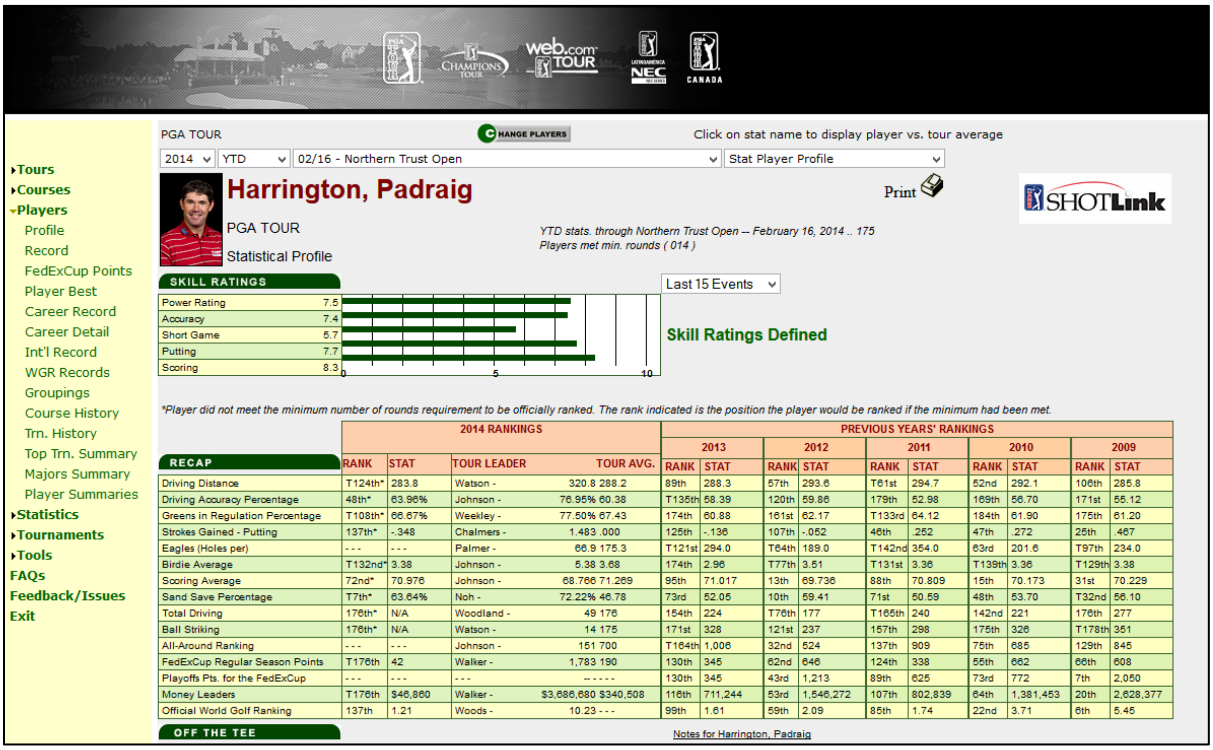
2.2 Shotlink

Since 2003, the PGA Tour has collected detailed information on every shot played by every player in every tournament sanctioned by the PGA Tour. This data is known as Shotlink data and is the primary source for all statistics used by the PGA Tour and affiliated TV Networks.

In 2005, the PGA Tour made the Shotlink data available for academic research use and they claim that more than 65 institutions have partaken in research using the Shotlink data. However, only 14 papers have resulted from this access, which are published on the PGA's website.

One of the successful outcomes of this research was the launch of a new "Strokes Gained - Putting" statistic. The aim of this new statistic was to remove bias found in other statistics such as "putts per round" which are influenced by previous shots, to give a clearer indication of which golfer is outperforming the field average with their putting.

The Shotlink dataset is accessed via a secure website. There are a number of Graphical User Interface (GUI) based tools available to query the dataset, which present the data pre-formatted, as can be seen in Figure 2.1.



PGA TOUR

Click on stat name to display player vs. tour average

2014 | YTD | 02/16 - Northern Trust Open | Stat Player Profile

Harrington, Padraig

PGA TOUR

Statistical Profile

YTD stats. through Northern Trust Open -- February 16, 2014 .. 175 Players met min. rounds (014)

Last 15 Events

SKILL RATINGS	
Power Rating	7.5
Accuracy	7.4
Short Game	5.7
Putting	7.7
Scoring	8.3

*Player did not meet the minimum number of rounds requirement to be officially ranked. The rank indicated is the position the player would be ranked if the minimum had been met.

	2014 RANKINGS			PREVIOUS YEARS' RANKINGS									
				2013		2012		2011		2010		2009	
RECAP	RANK	STAT	TOUR LEADER TOUR AVG.	RANK	STAT	RANK	STAT	RANK	STAT	RANK	STAT	RANK	STAT
Driving Distance	T124th*	283.8	Watson - 320.8 288.2	89th	288.3	57th	293.6	T61st	294.7	52nd	292.1	106th	285.8
Driving Accuracy Percentage	48th*	63.96%	Johnson - 76.95% 60.38	T135th	58.39	120th	59.86	179th	52.98	169th	56.70	171st	55.12
Greens in Regulation Percentage	T108th*	66.67%	Weekley - 77.50% 67.43	174th	60.88	161st	62.17	T133rd	64.12	184th	61.90	175th	61.20
Strokes Gained - Putting	137th*	-.348	Chalmers - 1.483 .000	125th	-.136	107th	-.052	46th	.252	47th	.272	25th	.467
Eagles (Holes per)	- - -	- - -	Palmer - 66.9 175.3	T121st	294.0	T64th	189.0	T142nd	354.0	63rd	201.6	T97th	234.0
Birdie Average	T132nd*	3.38	Johnson - 5.38 3.68	174th	2.96	T77th	3.51	T131st	3.36	T139th	3.36	T129th	3.38
Scoring Average	72nd*	70.976	Johnson - 68.766 71.269	95th	71.017	13th	69.736	88th	70.809	15th	70.173	31st	70.229
Sand Save Percentage	T7th*	63.64%	Noh - 72.22% 46.78	73rd	52.05	10th	59.41	71st	50.59	48th	53.70	T32nd	56.10
Total Driving	176th*	N/A	Woodland - 49 176	154th	224	T76th	177	T165th	240	142nd	221	176th	277
Ball Striking	176th*	N/A	Watson - 14 175	171st	328	121st	237	157th	298	175th	326	T178th	351
All-Around Ranking	- - -	- - -	Johnson - 151 700	T164th	1,006	32nd	524	137th	909	75th	685	129th	845
FedExCup Regular Season Points	T176th	42	Walker - 1,783 190	130th	345	62nd	646	124th	338	55th	662	66th	608
Playoffs Pts. for the FedExCup	- - -	- - -	- - - -- - - - -	130th	345	43rd	1,213	89th	625	73rd	772	7th	2,050
Money Leaders	T176th	$46,860	Walker - $3,686,680 $340,508	116th	711,244	53rd	1,546,272	107th	802,839	64th	1,381,453	20th	2,628,377
Official World Golf Ranking	137th	1.21	Woods - 10.23 - - -	99th	1.61	59th	2.09	85th	1.74	22nd	3.71	6th	5.45

OFF THE TEE

Figure 2.1: Screenshot of the Shotlink Graphical User Interface

Six levels of detail are provided for export, Stroke Level, Hole Level, Round Level, Event Level, Radar Launch and Radar Trajectory

- **Event Detail**

Provides 190 separate data points on each golfer, aggregated for the four rounds of each tournament. Examples of data points include: Player Age, FedEx Cup Points, Finish Position, Round scores, as well as a multitude of variations on scrambling, driving and putting distances.

- **Hole Detail**

Provides 50 separate data points on each hole in a tournament for each golfer, for each round. Examples include Tee Shot Landing Location and Made putt distance

- **Round Detail**

Provides 173 separate data points for each round for each golfer. This has a similar list of fields to Event Detail**.**

- **Stroke Detail**

This has 38 data points for each shot for each golfer on a given hole. Some examples of data points include X,Y,Z coordinates, distance to pin, lie and elevation.

- **Radar Launch**

Radar Launch data contains 50 data points for shots taken on either one or two holes per round. Data is collected on par 4 or par 5 holes only. Examples of data points include Ball Speed, Launch Spin and Flight Time.

- **Radar Trajactory**

Similar to the Radar Launch dataset with the additon of 6 additional data points containing x,y,z co-ordinates for the ball as it travels. This enables plotting of the flight path of the ball.

- **Data Size**

The size of data collected is available in Table 1.2. There are circa. 940,000 rows of data at stroke level, representing every shot taken on the PGA Tour in 2013. This gives some indication as to the comprehensive nature of the dataset.

- **Data Dictionary**

A data dictionary in PDF format exists for each level of detail available from the Shotlink website. For the most part, any jargon and the meaning of these terms is explained reasonably well. “Smash Factor” is such an example and is actually the ratio of ball speed to club head speed.

2.3 Conclusion

This chapter has introduced the game of golf, outlining the objective of the sport, together with the popular formats of play. It has also detailed the terms used throughout the project and provided a synopses of the organisations that govern the sport at both amateur and professional levels. Finally details of the Shotlink dataset were discussed in order to outline its history, uses and the sheer vast amount of data which is available for analysis.

Chapter Three will discuss the research which has been previously carried out using both Shotlink data, and other sources. It will examine the relevance of this research to the project and how it may be incorporated into the project experimentation.

3. GOLF SCIENCE AND ANALYSIS

This chapter will discuss previously conducted research and how it relates and contributes to the objectives of this project. Subject areas include performance assessment measures, performance analysis and external influences on a professional golfer.

3.1 Performance Assessment

Performance Assessment research is predominantly concerned with creating new statistics, or refining existing ones and then re-ranking golfers based on their scores of the new statistics. They focus on one individual aspect of the game and analyse what is wrong with current statistics for that skill and how their new statistics are much better at describing what is really going on.

Researchers in this category postulate that the current statistical measures do not reveal the whole truth, and can indeed mask certain key performance measures in the game. This is in a similar vein to the case outlined by (Lewis, 2003) in which protagonists such as Bill James wrote extensively on the unsuitability of baseball statistics to describe the true value of baseball players.

Most of the research in this category has utilised the Shotlink dataset which is particularly relevant to the direction of this project as it outlines how the data has been exploited and what value it has added to the body of knowledge.

3.1.1 Fairway Ball Striking

As winner of the “Shotlink Intelligence Prize”, (Riccio Ph.D, 2012) demonstrates which golfer is the best fairway ball striker. The case for the research is outlined firstly by the statement that there are already well developed measures such as longest driver, and most accurate driver. (Riccio Ph.D, 2012) also mentions that Greens in Regulation is a *“very good tee to green”* statistic. In essence, these well-developed measures are at their lowest granular level and are easily understood by anyone who as a basic comprehension of the game of golf and its facets, at a high level.

(Riccio Ph.D, 2012) argues that the PGA Tour "Ball Striking" statistic is of limited (or no) value for telling the story of who is the best fairway ball striker. This is because the Ball Striking statistic is actually the summation of a golfer's "Total Driving" (which is itself a summation on Driving Distance and Driving accuracy ranks) rank and their Greens in Regulation rank.

This statistic doesn't allude to the details of shots in between the tee shot and getting on the green. A measure of Fairway ball striking, also known as long approach shot ball striking is not exposed for easy consumption.

By studying the Shotlink data at a stroke level of detail the creation of a new metric for long approach shot accuracy is proposed. In order to do so (Riccio Ph.D, 2012) focuses on approach shots between 150 and 225 yards from the hole, on par 4 holes only. The reasoning behind this specific and limited set of criteria is, as contended, that it is the golfer's intention for these shots to hit the green. They are also likely to be a full swing iron shot, not a wedge or wood shot.

One of the main aims of the new metric is to normalise distances so that long drivers of the ball such as Bubba Watson can be accurately compared to shorter drivers of the ball such as Luke Donald. What this means in practise is that long drivers of the ball typically have shorter approach shots whereas shorter drivers of the ball have longer approach shots.

Typically, shot accuracy increases as the distance of the shot decreases. This means that longer drivers with shorter approach shots will have a higher number of greens hit in regulation where the opposite is true for golfers who regularly face longer approach shots.

Following a similar approach to (Broadie, 2008) and (Broadie, 2011a), a standard or average is introduced. This is simply the average percentage of greens hit by shots on a par 4 hole where the shot originated between 150 and 225 yards from the hole, for the top 125 golfers in 2012. It is important to note that this is not simply the Greens in Regulation statistic, as shots included in the approach above could be the 5th shot on a par 4 which hits the green.

The collated data produces an almost linear relationship between the percentage of greens hit to the distance of the shot. So, from a distance of 150 yards, 80% of shots hit the green while from distance of 210 yards only 40% hit the green, on average. This makes perfect sense and supports the notion of accuracy versus distance.

Using a linear regression, an equation is given which models the expected outcome for an approach shot from any given yardage in the 150 to 225 yard range. This is the “standard” or average benchmark that his results will be compared against.

From an evaluation of the results of this benchmark (Riccio Ph.D, 2012) states that the new measure normalises distances for long and short hitters, as was the aim. To confirm this, two sets of results are presented; the first is just the percentage of greens hit by long approach shots, the second compares this percentage to the average and ranks accordingly.

For example, Justin Rose hit 77.9% of greens, which was 10.74% better than average. Bubba Watson hit 73.56% of his greens but this was only 3.5% better than average. This results in Bubba Watson being ranked 28th on the list of the best long approach shot hitters.

These findings could provoke some debate. To state that Bubba Watson is only an average fairway ball striker is somewhat misleading and depends on the context. From the standpoint of this new metric alone, it shows that for a shot based on very specific criteria, who is better than average at hitting the green.

The stated aim of the metric to create a level playing field so as to accurately compare long and short hitters may actually be itself biased. For example, Bubba Watson for the time period in question was ranked 2nd for greens hit in regulation, so there is a discrepancy here.

This discrepancy between his GIR rank and new fairway striking rank can be explained by the fact that the GIR rank includes holes of all pars – 3s, 4s, and 5s - the new metric is just for par 4s. Also, a golfer can hit a shot from the rough, or a bunker say and still reach the green in regulation with that shot. These shots are explicitly excluded from the new metric.

Watson drops down the ranking for the new metric because, being a long driver of the ball, he is less likely to have an approach shot that satisfies the criteria for the metric, which is acknowledged. In addition to this however, when Watson does have an approach shot in the 150 – 225 yard range, he has less practise with these shots so his desired outcome, to hit the green, is less than a certainty.

It is also possible that if Watson is hitting an approach shot from this distance, on a par 4, that something has gone wrong with his tee shot and potentially this error could be on his mind when hitting the approach shot, one bad shot could lead to another.

As it stands, the metric is biased towards shorter drivers of the ball, who happen to have more approach shots at the specific distances. Shorter hitters such as Curtis, Moulder and Toms all have significant improvements in their rankings when compared to the average. Longer hitters like Dustin Johnson are last on the list of improvers.

While the fairway ball striking metric's purpose is to determine the best golfer for this particular shot, it does penalise longer driving players such as Bubba Watson. It may be better to have two "standards", one for long drivers of the ball, and one for shorter drivers of the ball. With this, it would be possible to compare a golfer to his closest peers, rather than from the full set of 125 players.

(Riccio Ph.D, 2012) acknowledges that short drivers of the ball need to compensate by being better fairway ball strikes. This could be said about the skills of every golfer though. There will be some aspects of the game in which a golfer excels and some where they are (relatively) poor. This is why context is so important.

By itself, this metric would have merit for a golfer who is seeking areas of his game in which to improve. For the consumption of the general golf fan or lover of statistics it may also be of some use. However, at a high level, this metric shows a narrow focus on only one aspect of the game that doesn't translate into an all-encompassing predictor of success or failure.

3.1.2 Strokes Gained

The Strokes Gained concept was first introduced by (Broadie, 2008) as an outcome from research measuring amateur golfers performance. Following on from this research, (Broadie, 2011a) outlines the concept in further detail. Previously referred to as "Fractional Par", in conjunction with (Fearing et al., 2010) it evolved into "Strokes Gained - Putting" and has been incorporated into the official statistics of the PGA Tour. The Strokes Gained - Putting metric is available in the Shotlink datasets.

Outlining the case for the necessity for a new metric (Broadie, 2011a) argues that "Putts per round" has deficiencies when trying to measure putting skill because it does not take into account putting distance.

The objective of the Strokes Gained metric is to give credit to those golfers who make a 60ft putt versus those golfers who tap in a putt from 6 inches. The putts gained measure is introduced to cater for the distance / length of the putt. The formula for which is:

Putts gained = Average putts to hole out – Actual putts to hole out

"Average putts to hole out" is the average number of putts needed to complete a hole from any given distance. It is calculated from the sum of putts from the given distance, for example, 10 feet for all for all golfers from all tournaments divided by the number of putts that began at 10 feet.

To emphasis the argument that Strokes Gained - Putting is better than other metrics, (Broadie, 2011a) examines the rounds of two golfers, Angel Cabrera and Ian Poulter from the 2010 Deutsche Bank championship.

From initial analysis it can be seen that Cabrera had a total of 26 putts for his round which was 3 less than the PGA average of 29 putts per round. Ian Poulter on the other hand had 32 putts for his round.

It can be seen from Cabrera's scorecard for the round in question that he had no putts on holes 5 and 13. This suggests that he holed out from off the green for these two holes. These could be considered "lucky" shots; however, Poulter wasn't so fortunate.

Analysed from a Strokes Gained perspective, based on the average putting distance of 9.6 feet, Cabrera "lost" 2.8 putts when compared to what the "PGA Average" golfer would have achieved from those same distances.

Inversely, for Poulter, he gained 2.3 putts compared to the average PGA golfer. For the round in question his average putting distance was 32.3 feet. Essentially the metric states that Poulter was better than average from those putting distances.

However, (Broadie, 2011a) neglects to mention the actual scores in the two example rounds used to illustrate the concept. Cabrera had a round score of 69, 2 strokes under par whereas Poulter had a 2 over par score of 73 for his round. Incidentally, Cabrera finished the tournament tied for 18th position while Poulter was tied for 45th place.

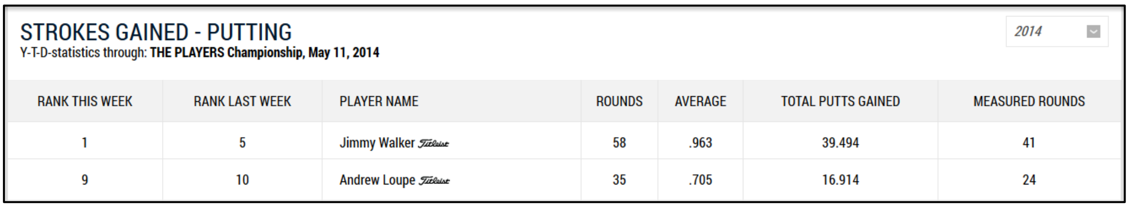
So while Poulter may indeed have been putting better than average for those distances, it was not reflected in his scoring. For this particular round, Poulter's desired requirement would have been to get his approach shots landing closer to the pin than his average of 32 feet.

Ultimately what matters is that a golfer plays a competitive round with as low a score as possible. If a player is fortunate enough to hole out and have no putts on a few holes of a given round they'll be grateful and move on. Over the course of four rounds of competitive golf this good fortune will balance out and a golfer's skill will determine their overall position.

The creation of new statistics such as Strokes Gained - Putting allow for the construction of new ranking lists on the PGA Tour's website. The leader of these specific metrics will vary as more tournaments are played. The current leader of Strokes Gained - Putting is Jimmy Walker ("PGA Tour Strokes Gained Putting Stats," 2014).

Walker has won two tournaments in 2014 and has only missed 2 cuts from 16 tournaments with earnings of $4.5 million. It would seem reasonable that Walker is top of the Strokes Gained - Putting rankings. Inversely Andrew Loupe, who is ranked 9th for Strokes Gained - Putting in 2014 has missed 7 cuts from 12 tournaments with earnings of $427,000. Loupe is ranked 307th in the official world golf rankings. The reader is referred to Appendix A for full details.

Loupe missed the cut of the most recent tournament played (Wells Fargo Championship) but still rose 1 place in the Strokes Gained -Putting ranks from the previous week from 10th to 9th as can be seen in Figure 3.1.



STROKES GAINED - PUTTING

Y-T-D-statistics through: **THE PLAYERS Championship, May 11, 2014**

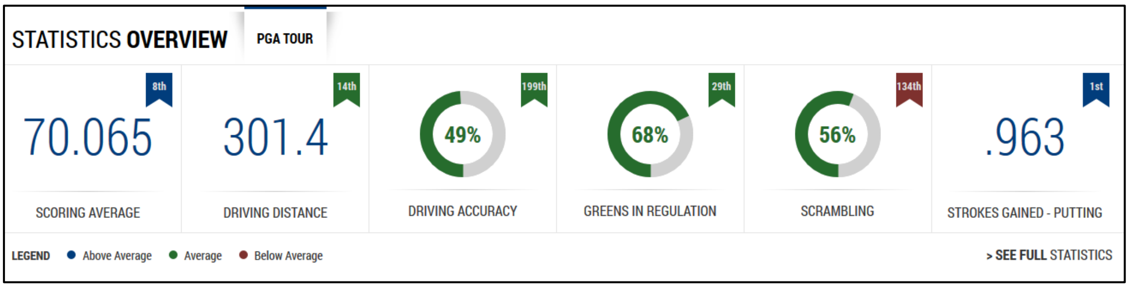
2014

RANK THIS WEEK	RANK LAST WEEK	PLAYER NAME	ROUNDS	AVERAGE	TOTAL PUTTS GAINED	MEASURED ROUNDS
1	5	Jimmy Walker	58	.963	39.494	41
9	10	Andrew Loupe	35	.705	16.914	24

Figure 3.1: Strokes Gained - Putting Statistics for Jimmy Walker & Andrew Loupe (PGA Tour, 2014)

This example illustrates the true value and indeed the problem of any one statistic. That Loupe increased his ranking in Strokes Gained - Putting while missing the cut is of scant consolation as there is no prize money for missing the cut. Looking at this statistic alone it could be assumed that Loupe was playing better than the contrasting tournament results.

Another example of the power of statistics to misdirect can be seen in Figure 3.2. This is a summary of some of the Jimmy Walker's traditional statistics



STATISTICS **OVERVIEW** | PGA TOUR

8th	14th	199th	29th	134th	1st
70.065	301.4	49%	68%	56%	.963
SCORING AVERAGE	DRIVING DISTANCE	DRIVING ACCURACY	GREENS IN REGULATION	SCRAMBLING	STROKES GAINED - PUTTING

LEGEND ● Above Average ● Average ● Below Average

> SEE FULL STATISTICS

Figure 3.2: Statistics Overview for Jimmy Walker (PGA Tour, 2014)

Walker is above average for scoring average (Ranked 8th) and Strokes Gained - Putting (Ranked 1st). For driving distance, driving accuracy and Greens in Regulation Walker is average while being below average for scrambling. What this shows is that there are clearly many different ways to achieve success on the PGA Tour and no one statistic will give an indicator of form, apart from the final score.

Analysis of one statistic alone does not reveal the full picture of a player's skills, or trends. It is when correlation or association analysis with other metrics is conducted that the true value can be revealed. Nevertheless if a golfer is consistently finding themselves losing strokes to the field (the average) then metrics such as Strokes Gained - Putting will allow the professional golfer to focus in on a particular problem area and work on that aspect of their game.

3.1.3 Success in a single metric

(Sen, 2012) seeks to explain the success of a golfer by creating a single metric called the “Key Criterion for Success”, abbreviated to KCS.

(Sen, 2012) contends that the KCS or a similar metric is necessary as many researchers have found anomalies between a professional golfer’s predicted and actual earnings. The objective of the KCS metric is to narrow the gap between predicted and actual earnings.

To illustrate the rationale behind the metric, (Sen, 2012) uses the examples of two professional golfers who played in the 1998 season, Steve Jones & Vijay Singh. In 1998, Jones was ranked higher than Singh in most of the statistical categories but Jones had much lower earnings compared to Singh - $741,544 for Jones vs Singh’s earning of $2,238,998 [5].

The stated goal of the KCS metric is to provide a simple way to describe a golfer’s performance more accurately than the use of individual statistics. (Sen, 2012) realises that the power of each individual golfing statistics is of limited value by itself. For the creation of the KCS metric, two existing “measures of success” are incorporated into measure – adjusted weighted score and earnings per event.

The “adjusted weighted score” statistic appears to have been picked simply because it hadn’t been used in previous research prior to the time of publication. Earnings per event is chosen because it as an expected goal of any professional golfer to maximise this statistic. There is no suggestion that (Sen, 2012) used any other scientific approaches such as correlation analysis for picking these two statistics for incorporation into the metric.

In order to define some parameters for the basis of the metric, (Sen, 2012) introduces a theoretical version of the game of golf with the following strict rules

- A golfer can only score a birdie if they reach the green in regulation
- If a golfer misses a green in regulation then they can only score a bogie
- A golfer can only deviate from par by 1 stroke

[5] http://www.pgatour.com/stats/stat.109.1998.html

In this theoretical model, it is only possible for a golfer to score birdie, bogie or par. Eagles & Double bogies cannot happen in this model.

With these rules in place, the KCS metric derived will give a better score to a golfer with a greater proportion of birdies than par or bogies. The formula for the metric is given as:

$$\frac{\frac{Birdie\ Conversion}{Failure\ to\ Scramble}}{\frac{(1 - Greens\ in\ Regulation)}{Greens\ in\ Regulation}}$$

Unfortunately, (Sen, 2012) neglects to mention the reasons as to why these particular statistics were chosen or the process behind the formulation of the metric.

With the KCS metric, it is possible for golfers with varying skills to achieve the same KCS score. Whether a golfer has a better short game or long game compared to other golfers doesn’t really matter, it’s the final score for the hole which counts.

The idea behind the KCS metric is beneficial. Having a derived single metric to score a player would be highly desirable for the models to be designed in chapter five of this project. However, the assertion by (Sen, 2012) on the merits of the KCS metric are open to debate.

(Sen, 2012) states that the PGA does not provide details on the over / under par rounds of a professional golfer. This claim is incorrect as Shotlink data provides all such data. (Sen, 2012) obtained the data used to test the KCS metric from the PGA tour website and other sources which do not offer the same level of detail as provided for in the Shotlink dataset.

(Sen, 2012) also implies that the KCS metric would perform better if this “over / under” par data had been available at the time of the research. In light of the fact that the data is available it would be interesting to see a continuation of the work in order to further refine the metric’s effectiveness.

One of the disadvantages of the model behind the KCS metric is that it does not take into consideration any scores greater than a bogie or lower than a birdie. It is also very possible to make a birdie even though a green has not been hit in regulation. While

(Sen, 2012) acknowledges these restrictions they would appear, at a high level, to limit the value of the metric.

As contested by (Sen, 2012) it would appear that the aggregation of two statistics enhances their predictive functions. For the comparison of Steve Jones and Vijay Singh, Singh had a higher KCS score reflecting his real money list ranking (2nd place in 1998).

Even without necessarily realising it, (Sen, 2012) appears to accept that there are more limitations to the metric than benefits. One accomplishment is discussed compared to three limitations.

These limitations include the large variance in tournament prize funds where not all PGA Tour golfers are eligible to enter each tournament. These results in cases where golfers do not overlap in tournaments played therefore making it difficult to have a true like for like comparison.

In the case of Jones Vs Singh, (Sen, 2012) notes that they only had 53% overlap in tournaments and Singh also played in all of the four major tournaments. This will have an obvious impact on Singh’s earnings versus Jones’ and so would affect the accuracy of the computed KCS score for ranking purposes.

It appears that the KCS metric may be better refined to compare golfers who have more features, such as tournaments in common. Perhaps when there is tournament overlap by the featured golfers of 70% or more it would give a better result.

As the KCS metric is based on a theoretical model of golf which has some significant limitations and assumptions, it appears to be of limited value. The premise of combining the power of two or more individual statistics to derive a more descriptive measure is logically sound and this approach will be used in the project.

In terms of the promise of the predictive nature of the KCS metric, (Sen, 2012) does not elude as to how it can predict future earnings. As such this would classify the KCS metric as another performance assessment measure which can be used to re-rank golfers based on past tournament performances.

3.2 Performance Analysis

Juxtaposed with performance assessment is performance analysis. The research carried out in this area typically seeks to outline which particular sets of skills are most important in determining a golfer’s scoring averages or winnings. Regression analysis is the recurring technique involved in determining which skills are significant. The reader is referred to Section 4.4.3 for details on regression analysis interpretability.

3.2.1 Regression Analysis

By examining the determinants of golfer performance (Peters, 2008) debates that golfers have to specialise in the aspects of the game they would like to excel at. Spending more time practising one aspect of the game, for example putting, is to the detriment of time available to practice their driving. If a golfer knew for certain which skills are most important then they could make a more intelligent decision regarding which areas of their skillset need more time to be honed.

As (Peters, 2008) points out, the golfers who compete on the PGA Tour are acknowledged as being some of the best in the world. The difference between success and failure can come down to small margins and this similarity in skill level suggests that by analysis of their combined skills data, scoring averages and earnings can be accurately predicted.

(Lewis, 2003) offers a similar idea with the notion that every shot that happens in baseball has already happened thousands of times before. The outcome of these shots is also recorded making it possible to estimate a probability of that same outcome occurring again, based on the trajectory, speed, and finishing point of the shot. It doesn’t matter who pitched the ball, who struck it, or who caught it.

During his discussion (Peters, 2008) refers to the research by (Alexander and Kern, 2005) which concluded that putting was the significant skill factor when attempting to determine golfer earnings. They also noted that driving distance was becoming a larger determinant over time. This appears to be corroborated by the research conducted by (Broadie, 2008).

(Peters, 2008) seeks to differentiate his research from previous related studies by including a player’s experience (the number of years on the PGA Tour), as well as the number of events played per year. For the first time, fatigue is introduced as a possible factor in a golfer's ability to earn.

In the regression model produced, the following explanatory variables where used to describe the dependent variable "Scoring Average"

- Driving Accuracy
- Driving Distance
- Percentage of Greens in Regulation
- Average number of putts per green
- Percentage of Sand Saves
- Experience (The Number of Years a golfer has played on the PGA Tour)

The results showed that Driving Accuracy has no significance to the scoring average (with 99% confidence). All the other variables were significant, with results showing that if a golfer increased their driving distance by 10 yards, they would only gain 0.1 of a stroke per round. Yet if they increased their putting average by 10% a golfer could gain 2.3 strokes per round. The other variables were in between these two ranges for potential strokes gained. Based on these results, putting was deemed to be the most important skill in determining scoring average.

For the regression model that incorporated experience, it was shown to have significance in the explanation of scoring average with a higher significance level than both driving distance and driving accuracy. (Peters, 2008) does admit that there may be some issues with multi-collinearity as the coefficients for the other variables did not change as a result of the inclusion of the experience variable in the model.

Further analysis shows that driving distance is appreciably higher for younger players with less experience of playing on tour. This makes sense and could be attributed to the advantage of their youth. Inversely, and somewhat expectedly driving accuracy increases with age, though whether this is directly attributable to the wisdom of age, or a side effect of shorter driving as mooted by (Riccio Ph.D, 2012) could be a matter of debate by itself.

(Peters, 2008) finds that Average Putts Per Round followed by Greens in Regulation to be the most important factors which determine scoring average. Apparently, the number of tournaments a golfer entered per year does not have any bearing on a golfer's earnings. This is not a measure of fatigue however; as it does not consider how many successive tournaments a golfer has entered.

Similar studies have been conducted by (Heiny, 2008) and (Ridenoure, 2005). The additional variables of height and weight of golfers was introduced into regression analysis by (Ridenoure, 2005) but was found to be an insignificant explanatory factor for both money earned and scoring average.

From all three studies, the main factors which were deemed to contribute to either scoring averages or money earned were:

- Driving Distance
- Driving Accuracy
- Greens In Regulation
- The Number of Putts per Round
- Scrambling
- Sand Saves (to a lesser extent)

These variables can be included in the experiments of this project to discover if they also explain a golfer's ability to make the cut in a given tournament.

3.2.2 Strokes Gained

Similar to (Peters, 2008), (Broadie, 2011b) attempts to assess which skills contribute most to victory on the PGA Tour and unsurprisingly the Strokes Gained concept is the foundation of the presented research.

By calculating a benchmark for the expected number of strokes to complete a hole from any distance (based on the PGA average performance) (Broadie, 2011b) can determine the contribution of a golfer's long game, short game and putting to his round.

The use of the Strokes Gained concept to explain the relative importance of the many facets of skill required by a professional golfer is *"an alternative to regression analysis"* (Broadie, 2011b). When compared to regression analysis, it is easy to see that a Strokes Gained alternative would be more appealing to professional golfers to consume and digest, than attempting to explain what the co-efficient is and what adjusted R squared values mean.

In order to achieve the goal of a regression analysis alternative, (Broadie, 2011b) logically evolves the concept behind Strokes Gained - Putting to the other remaining aspects of the game – long and short play.

Calculated in the same way, a “benchmark” is created to determine how the notional “average” PGA Tour golfer would perform, based on the distance from the hole and the lie they find themselves in (on the fairway, in the rough or in sand, for example).

Shotlink data from 2003 to 2010 is used to compute this average for the many permutations possible. Benchmarks are created for Tee Shots, being within 50 yards of the hole, recovery shots and putting. There are slight variations to the way in which each benchmark is calculated, in order to provide the fairest measure for every golf shot to be compared against.

The identification of a recovery shot (for example, a shot which has a tree in line of sight with the hole) is interesting. As there is no explicit detail of which shots were a “recovery” shot, probably due to its subjective nature, (Broadie, 2011b) has devised a formula to create a best guess measure. For example a shot could be considered a recovery shot if the golfer’s previous shot was a shorter than average tee shot, or their drive had an acute angle on launch. These types of shots would indicate that the previous shot went awry.

Using the Strokes Gained metric, (Broadie, 2011b), by way of an example, shows that Tiger Woods typically gained 3.2 strokes per round between 2003 and 2010. Woods gained 2.08 strokes from his long game; 65% of the total strokes gained.

Applying Strokes Gained across all golf shots in order to assess golfer performance is a justification for its existence. It is an interesting statistics in its own right but as with Strokes Gained - Putting, it fails to tell the full story.

Table 3.1 details the Strokes Gained rank provide by (Broadie, 2011b) married with these players PGA tournament earnings for the seven years between 2003 and 2010. It can be seen that there are some notable exceptions where there is a reasonable difference in the two rankings.

Strokes Gained Rank	Player Name	Money List Rank	Prize Money 2003 -2010
1	Woods, Tiger	1	61,053,451
2	Furyk, Jim	4	33,709,120
3	Singh, Vijay	2	45,039,373
4	Els, Ernie	5	25,060,889
5	Mickelson, Phil	3	37,527,420
6	Donald, Luke	22	17,496,243
7	Goosen, Retief	6	22,090,594
8	Garcia, Sergio	15	19,408,485
9	Scott, Adam	8	21,705,217
10	Harrington, Padraig	18	18,510,580

Table 3.1: Strokes Gained Vs Money Earned. (Broadie, 2011b, Shotlink,2014)

Luke Donald, Sergio Garcia and Padraig Harrington all feature prominently in the Strokes Gained ranks but as can be seen in Table 3.1 above there is quite a difference in rankings from the money list. If Luke Donald can be ranked 6th in Strokes Gained but is only 22nd on the money list it shows that it is not an entirely effective measure of assessing golfers performance. The use of aggregate data for the seven year period from 2003-2010 is somewhat questionable. Seven years is a long time to measure most golfers on. A golfer’s form can come and go so it would be interesting to see the results of the research reported on an annual basis.

3.3 External Influences on Golfer Performance

3.3.1 Caddies

All PGA Tour professional are accompanied by a caddie for every round they play in a tournament. The caddy fulfils a multi-purpose role – they have responsibility for carrying the golfer’s bag, ensuring that the correct number of clubs is present etc. They also provide psychological support and can have a calming influence in high pressure situations.

A number of studies have been carried out to determine if the influence or otherwise a caddy has can be observed in a golfer’s scores.

In their study, (Coate and Toomey, 2012) compare two separate tournaments, the Western Open and The Masters. These two tournaments were identified as suitable for the study as they both at one point in their respective histories only allowed local caddies to caddy for the pro golfer. The Masters only allowed local caddies between 1934 and 1982 and the Western Open had the stipulation between 1974 and 1986. From thereon in, competitors in both tournaments were allowed to use their regular caddies.

For each tournament, the last 3 years of local caddy use is compared to the first 3 years of tour or regular caddy use. They estimate models of daily scoring over the 6 years as a function of variables such as player quality, weather conditions, which round was being played, and dummy variables for caddies assigned a value of 1 or 0 depending on the year the round was played.

The results of the studies show that tour caddies gave an advantage of 1.5 strokes per round at the Western Open and an advantage of 0.8 strokes per round at The Masters.

(Coate and Toomey, 2012) suggest that the reason there is an advantage of almost double for the Western Open vs The Masters is because the local caddies at The Western Open were High School graduates with less experience. The local caddies at The Masters typically had fulfilled the role for many years and so had considerably more experience.

This study shows that there does appear to be a difference in a golfer’s scoring when they are forced to pair with a caddy they do not have a regular relationship with. However, by forcing all golfers to use local caddies, the tournaments insured a level playing field. Any “advantage” is academic: as (Coate and Toomey, 2012) state; almost all the golfers used their own chosen caddies when the rules were relaxed for both tournaments. This would suggest that the golfer’s themselves were also aware that they’re scoring was affected by use of the local caddies.

At the very least, (Coate and Toomey, 2012) demonstrate that there is promise in the field of studying caddy influence over golfer scoring. It is a shame that the study was limited by its nature to just two tournaments over a period of 6 years and that no comparison of tour caddies against each other was conducted to supplement their research.

Exploring the role of the Caddy (Lavalle et al., 2004) interviewed 8 professional golfers and 8 caddies who competed on the Australasian PGA Tour. Two pairs of the golfers and caddies worked together, with the remaining interviewees having no relationship to each other.

(Lavalle et al., 2004) mention some interesting facts such that the average golfer will only be planning a shot for just 25% of his time on the course and only 2% of his time will be spent swinging the club to take a shot. This insightful information leads to the possibility of how a golfer uses this downtime. Does a slow analytical player have a tendency to over think and have a higher probability to then execute a bad shot, or indeed dwell on the bad shot that cost them a double bogie?

(Lavalle et al., 2004) describe the role of the caddy as being both technical and psychological. As time and familiarity increase, psychology becomes more important in the golfer caddy relationship. It is also interesting to note that one of their observations was that none of the interviewed golfers or caddies mentioned that goal setting was an explicit part of their usual routines.

Implicitly though (Lavalle et al., 2004) derived that the one of the main goals of the caddie was to "*optimise a player's mental climate*". Echoing the sentiment expressed in (Rotella, 2008), a key part of optimising a golfer's mental state is to ensure the isolation of each shot. This means that should a golfer have a bad shot and they begin to dwell on it, for whatever reason, then the caddy will step in, in an attempt to get them back on track before the next shot.

In their conclusions there is no explicit statement stating that caddies have a positive influence on a golfer's scores. Nonetheless there is a positive relationship between the golfer and their caddies, especially as the relationship matures. The optimal life of this relationship is questioned, due to evidence of an over reliance for assistance from the caddy, to the point where the golfer takes a passive role with club selection etc.

Golf history is littered with examples of golfers firing caddies in an effort to find some new "spark" in their game[6]. There are also plenty of long standing successful partnerships such as Phil Mickelson and his caddy Jim "Bones" Mackay who have been working together since 1992.

[6] Chapter Five outlines selected examples of caddy changes used for experimentation.

There does seem to be some evidence that caddies do indeed exert some influence on the scoring of the golfers they partner for. Unfortunately, the studies discussed, are limited in nature and lacking the depth that would support an unequivocal view on the matter.

Regrettably, the Shotlink dataset does not contain any details of the caddies for each golfer for each tournament. It would be necessary to analyse a number of different external sources in order to compile a thorough list of golfer and their caddies, by tournament. This would not be possible given the timeframe of this project.

Instead an examination will be undertaken to test the significance of the influence of caddies in instances where there have been high profile caddie changes. These can be fulltime moves such as Steve Williams' switch from Tiger Woods to Adam Scott, or temporary changes such as when Padraig Harrington's caddy Ronan Flood collapsed from dehydration at a tournament in January 2014 (Gray, 2014).

For form or cut line prediction, a number of questions arise from (Lavalle et al., 2004) based on the psychological aspects of the game. While less useful to the golfer, it may be easier to predict who will miss the cut after the first round of a tournament. It may be possible to determine if, after changing caddies, a player is better able to recover from a poor first round than with their previous caddy.

The introduction of "downtime" in this research coupled with the statistics suggesting that a golfer has 75% of his time on the course not physically playing golf provides a fascinating piece of information. It leads to the possibility of introducing time spent on course as a potential avenue for analysis and potential correlation to form or cuts made / missed.

The Shotlink dataset does provide details of tee times, along with hole start and end times. There is no aggregate information but it would be possible to derive a golfer's round time and categories them as slow, normal or fast players.

3.3.2 The impact of age on professional golfer performance

Age is the unstoppable force that is one of the greatest factors on any athlete's ability to perform. Depending on the sport and the physical demands of that sport, the age at which an athlete's performance declines can vary. For example (Schulz and Curnow, 1988) find that tennis players peak at the age of 24 while baseball players tend to peak at the age of 28.

Golf is not as physically demanding as tennis which is probably why it is littered with examples of outliers – older golfers who have won tournaments, or come very close to winning tournaments. Jack Nicklaus won his last Masters jacket at age 46 while Tom Watson lost to Stewart Cink in a playoff for The Open in 2009[7]. More recently Miguel Angel Jimenez became the first golfer over the age of 50 to win on the European Tour[8].

Some research has been carried out with respect to the number of years' experience (and implicitly, age) on the PGA Tour by (Peters, 2008). This showed that with experience, a professional golfer's driving distance off the tee decreased but their accuracy levels increased. Moreover, experience was not found to be a significant factor in a professional golfer's earnings. This may be due to the fact that once a golfer turns 50 they tend to play on the Senior Tour which has its own separate dataset within Shotlink.

More specific research has focused on the age of a golfer with respect to peak performance and the ability to perform under pressure. (Schulz and Curnow, 1988) investigated the age of peak performance in a broad range of sports, including professional golf. It was found that professional golfers tended to "peak" at 31 years of age.

As (Schulz and Curnow, 1988) note that the "*…level of measured performance is in part determined by the competition*", it would be difficult to judge when each golfer has reached their peak. Due in part to this issue, peak performance was determined by the age at which a golfer reached number one in the world rankings.

[7] http://sports.espn.go.com/golf/britishopen09/news/story?id=4339293

[8] http://www.independent.ie/sport/golf/miguel-angel-jimenez-makes-history-as-first-winner-over-50-on-european-tour-30284419.html

While (Schulz and Curnow, 1988) found that the age of golfers was tending towards a younger age when achieving number one status, it is arguable whether this should considered “peak” performance. For golfers in the lower ranks, it could be said that their peak performance could be measured by cuts made. For higher ranked players, the age at which a first major is won could be a more important pointer. It would be very much down to the individual golfer’s own opinion as to when they actually peaked.

(Fried and Tauer, 2011) appear to have taken a more considered approach to age and performance – specifically performance under pressure. (Fried and Tauer, 2011) suggest that a golfer’s skill first increases with age, thanks to practise and experience. It then declines as age increases and interacts with physical ability.

This relationship can be described as an inverted U shape. However, as mentioned above, there are numerous examples of golfers attaining a rejuvenation of their skills at an older age which can skew this general inverted U shape relationship.

(Fried and Tauer, 2011) create an “Age Efficiency Ratio” to determine the relationship between age and performance. Changes in physical skills are ignored in order to ensure a score solely based on age. The ratio is between “unconditional” scores which compare all golfers, regardless of age and “conditional” scores which compare golfers within a certain age group.

Based on this ratio, (Fried and Tauer, 2011) discover that, in general, experience is the main driver of performance, up to the age of 36, with performance declining thereafter. At an individual golfer level the so-called efficiency score shows mixed results. Citing the examples of Luke Donald and Geoff Ogilvy in 2005 (both were 28 years old), (Fried and Tauer, 2011) state the Ogilvy managed his age better than Donald.

In spite of this revelation, Donald earned more prize money per event even though Ogilvy had better golfing inputs. The reason for this is that while Donald had a lower conditional score which meant he underperformed compared to his age group, he still had a higher unconditional performance (where age is irrelevant) which means he is better than the collective set of sample golfers. This example highlights the issue with comparing golfers to two groups. Discrepancies may occur in interpretation where one group is a notional comparison – age, versus the group that exists in reality, the money list.

3.3.3 Psychology

Psychology is a very important aspect of the game of golf. If a golfer, either amateur or professional lacks the confidence in their golf skills then their performance will suffer. How confidence (or lack thereof), manifests itself, is individual to each golfer.

Sports psychologists work with athletes to ensure that they have the mental fortitude to perform to the best of their abilities and beyond. This is especially true in cases where an athlete experiences pressure to win, or pressure to bounce back after suffering a loss, (or succession of poor performances).

In golf, one of the most feared psychological phenomena to affect a golfer is called the "Yips". (Klämpfl et al., 2013) describe the Yips as *"involuntary movement during the execution of a skill"* such as spasms, tremors, twists and jerks, predominately in the wrists and lower arm. The Yips can affect any aspect of a golfer's game, but it is typically most commonly observed in the putting stroke.

A golfer afflicted with the Yips can suffer from detrimental effects on performance. Bernhard Langer is one of the more well-known golfers to have suffered from the Yips at various points in his career. In 1991 Bernhard Langer missed a 5 foot putt in the Ryder Cup which cost Europe the trophy[9].

(Klämpfl et al., 2013) discuss the concept of "Reinvestment" as a cause of the Yips. The theory behind reinvestment is that once a golfer is aware of, or believes they have the Yips, they will consciously control their movements in an attempt to override their subconscious, or learned movements which they believe are the Yips. This reinvestment becomes the Yips, thereby creating a vicious cycle.

In order to determine if Reinvestment is indeed a factor in a Yips affected golfer's performance, (Klämpfl et al., 2013) carried out a series of experiments. A total of 22 golfers where either distracted from the act of putting or forced to concentrate on their putting stroke.

[9] http://www.golf.com/tour-and-news/bernhard-langer-golf-magazine-interview-ryder-cup-anchored-putting

The hypothesis of the study was that Yips affected golfers would putt better when their attention was drawn away from the task of putting. However, (Klämpfl et al., 2013) found that contrary to previously studied research there was no evidence to conclude that this was the case.

Unfortunately, the conclusions of (Klämpfl et al., 2013) do not help a Yips affected golfer to overcome their problems. Some sports psychologist such as (Rotella, 2008) suggest that the key to overcoming the Yips is to give each shot the same level of (un)importance as each other. This way, a golfer will not succumb to pressure and treat, for example, a putt to win the Masters the same way as if they were simply playing a practise round with friends.

Psychological Momentum

To any sporting competitor or avid sports fan, the idea of Psychological Momentum is one that they both understand and relate to. When an individual or a team begin to achieve competitive success (whether for the first time, or after some failures) and this success continues it could be perceived that they are experiencing Psychological Momentum (Savage, 2012).

Interestingly, Psychological Momentum appears to only consider positive results. It could equally be reasoned that if a competitor has successively less successful performances that they are experiencing negative Psychological Momentum.

Thus, concentrating on the positive side, momentum in a sporting context is when a competitor experiences a run of good results. In golf, this could translate to 6 birdies in a row, successive wins, or even making successive cuts after a series of missed cuts. This positive momentum give rises to a heightened confidence that changes the competitors opinion of themselves and of others (Iso-Ahola and Mobily, 1980).

Experiments to determine if Psychological Momentum is a measureable phenomenon detailed by (Savage, 2012) showed that, in the case of basketball at least, 88% of analysed players were more likely to make a shot following a missed shot than following a made shot. This notion of success following failure, rather than success breeding success is aligned to the discovery by (Broadie, 2008) that golfers have a higher likelihood of making a par than a birdie.

(Savage, 2012) conducted a number of experiments which are of relevance to this project. The main hypotheses of the experiments are:

1. Cuts occur in sequences
2. A golfer's first round score determines their chances of making the cut
3. A golfer's performance in a previous tournament can predict the same golfer's performance in the next tournament.
4. The performance of each previous round determines the performance of the next round

Based on two years of data for 204 golfers, (Savage, 2012) concluded that cuts made or missed do indeed occur in sequences that could not be explained by chance alone. However, when these golfers were split into quintiles, the opposite conclusion was made, i.e. the sequence of cuts missed or made was down to chance alone.

As the quintile cuts occur by chance, (Savage, 2012) states that the golfers in each group had similar performances, but *"between the groups, there were non-random patterns"*. This would imply that say for the Top 25, it is poor form, bad luck etc. to miss the cut.

However, if a golfer consistently ranks less than 70th, they are probably more likely to miss the cut. Within that golfer's quintile of those with similar performance, this would be chance too. The only differentiator between the quintiles is a different sequence of runs of cuts made or missed.

This discrepancy between the performance of each quintile and the full group of 204 players, causes (Savage, 2012) to conclude that factors of Psychological Momentum must be at work between the highest and lowest ranked players.

Whether a golfer's first round score determines their chances of making the cut is also analysed in (Savage, 2012). Using logistic regression with a binary dependent variable of "making the cut" and a single regressor – 1st round score, it was determined that making the cut was dependent on the 1st round score.

In this instance, a golfer was deemed to have positive Psychological Momentum, only if they shot a score under par. Course difficulty does not appear to have been a determining factor which could easily affect scoring on a given day. It is perfectly plausible to have a difficult course or conditions where only 10 players shoot under par for the first round. It would be unlikely that only these 10 players would have experienced Psychological Momentum.

For the remaining experiments, (Savage, 2012) found no compelling evidence to suggest that a golfer's performance in one tournament would determine their performance in the next. On the other hand, a strong correlation was found to suggest that the score in each round was correlated with the next round.

3.3.4 Quantifying Course Difficulty

Measuring golf course difficulty is initially presented by (Connolly and Rendleman Jr., 2008). By attempting to measure Tiger Woods' dominance on the PGA Tour a generalised additive model was used to "estimate time-dependent mean skill functions and the first-order autocorrelation of residual scores about their means". (Connolly and Rendleman Jr., 2008) seek to remove the course difficulty factor from the models by the removal of the estimated random effects.

While (Connolly and Rendleman Jr., 2008) removed course difficulty from the models created, (Broadie, 2011b) incorporates it into the Strokes Gained process. This initially seems to be fair and valid as it is well known that professional golfers typically have a preference for certain types of golf courses; for example[10], links or parkland.

The desire by (Broadie, 2011b) to create a collective benchmark from numerous courses brings balance to this benchmark allowing the Strokes Gained score to be reflective of the course difficulty. Some tournaments could have a winning score of 21 under par (suggesting an easy course or setup). Others such as the 4 major tournaments typically have a winning score much closer to par, and sometimes over it. To account for course difficulty, (Broadie, 2011b) only uses course length.

As (Cosgrove, 2014) discussed there are many factors which can affect the way a golf course plays. The length of a course is indeed one factor in its difficulty, though (Heiny, 2008) shows that over time, golf equipment has evolved to the point where the

[10] http://www.pga.com/news/champions-tour/rory-mcilroys-criticism-links-golf-british-open-surprises-several-senior-stars

average driving distance on the PGA Tour has increased by approximately 26 yards from 1992 to 2003.

To the avid golf fan, it is well known that advances in club technology have led to longer distances and in turn have caused golf course designers to increase the length of courses in response.

Other factors which affect course difficulty are weather conditions and course setup. Weather conditions, as an "act of God" are beyond the control of both the professional golfer and the tournament organisers. (Cosgrove, 2014) mentions that course setup can change from round to round and especially from year to year if the competition organisers decide that the winning score will not be lower than 10 under par (for example).

Therefore, in the case of predictive modelling, rigidly classifying a golf course as "Easy" or "Difficult" could be flawed; an easy course could become difficult the following year. Knowledge of course setup would be needed in advance in order to introduce it to any models built and expect it to have a positive effect on the model.

Of course, if the analysis is solely to establish what happened in the past as opposed to predicting what will happen in the future (as appears to be the object of the Strokes Gained metric) then advance knowledge of course setup is not required.

An alternative to the gathering of collective data for a measure of course difficulty would be to create a benchmark for each individual golf course, based only on previous scores from past tournaments at that specific course. In this way measures such as "Strokes Gained" could be tailored to that particular course and adjustments made based on how the course played, or is expected to play.

3.4 Conclusion

This chapter has presented research from the areas of performance analysis, performance assessment and the external factors which affect professional golfers.

The performance assessment research is disappointing from the viewpoint of use for predictive analytics. Strokes Gained and Fairway Ball Striking measures offer limited value as they are essential a means of re-ranking the past. They're value lies solely for the purpose of providing a key performance indicator to a professional golfer who is looking for the possible problem area of their game.

The performance analysis research would appear to have more value to be exploited for use in the predictive experiments of this project. Key variables which explain a golfer's performances have been described which can be put to use in the experiments.

The use of Strokes Gained for all parts of the game of golf (not just putting) may well prove to be a useful variable when determining a golfer's form or when a cut will be missed. Unfortunately, the Shotlink dataset only contains data on Strokes Gained for putts.

The external factors discussed also present some interesting questions. The sway of caddies, psychological momentum and the impact of age on performance should be quantifiable to some degree using available Shotlink data. Further analysis is required to establish an effective metric for course difficulty which is neither too complex nor too simple.

Chapter Four will now discuss the various classification methods such as Decision Trees and Neural Networks which will incorporate the ideas from this chapter. It will also introduce hypothesis testing and its uses in this project.

4. CLASSIFICATION PREDICTION AND STATISTICS

From a Data Analytics perspective, classification is the of building a model to predict categorical labels (Han et al., 2012), such as "Buy" or "Sell" in the example of stocks and shares, or "Yes" or "No" in the prediction of a golfer missing the cut or not.

There are numerous methods available to the Data Analyst which allow for the prediction of a particular classification. As they are used in this project, Machine Learning algorithms such as Decision Trees, K Nearest Neighbour & Neural Networks will be discussed. Statistical methods such as Analysis of Variance (ANOVA), T-Tests and Regression will also be introduced so as to allow the reader to have an informed comprehension of the methods used for experimentation.

4.1 Supervised Vs Unsupervised Learning

Machine Learning classification techniques fall under two main categories which predict classification problems, Supervised and Unsupervised Learning.

Supervised learning is where a model learns from the class label during the training of the model (Han et al., 2012). Based on the data provided alongside this known class label, the model can then determine a set of rules on which to classify new unseen data.

Unsupervised learning is where a model has no knowledge of the class label, either it is not provided, or unknown in the first instance. Instead of predicting an exact class label, unsupervised learning models will create groups of similar classes (Han et al., 2012). Analysis of these new classes can help to discover previously unseen groups, or clusters of similar data.

A summary of the machine learning classification techniques used in this project follow:

4.1.1 Decision Trees

Decision Trees can be used to predict discrete classification type values or continuous regression type values (Mac Namee and Kelleher, 2012). Decision Trees are considered to be an "eager learner" classification method as the model is built and ready (eager) to classify new data before this new data is presented to the model (Han et al., 2012). For this project binary classification decision trees will be utilised and hence outlined in some detail. Regression trees have not been employed.

Figure 4.1 illustrates a simple (colour coded) decision tree to predict whether or not a customer will buy a computer. Decision Trees consist of one root node (coded red), multiple internal nodes (coded blue), leaf or terminal nodes (coded green) and branches (coded orange) to connect the nodes.

The Decision Tree predicts a binary classification value (either positive or negative) for a target variable, based on the attributes which best describe the target variable. When training, the decision tree performs a number of tests which are represented by the root and internal nodes with the branches representing the possible outcomes of each test (Han et al., 2012). The leaf nodes contain the final class label to be assigned to the target variable, if that leaf is reached for new, unseen data.

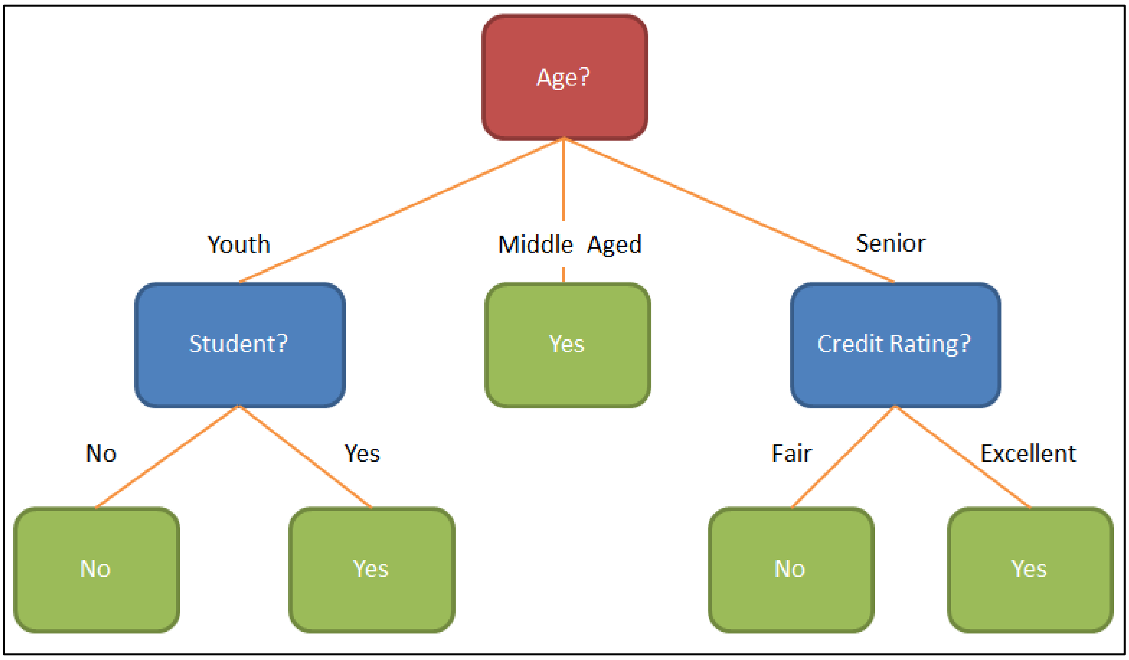


Figure 4.1: Classification Decision Tree - Based on example by (Han et al., 2012)

The root node attribute is chosen by the algorithm as being the attribute which will split the data in the best way, that is, a split with the fewest possible outcomes and lowest number of positive and negative examples for each outcome (Mac Namee and Kelleher, 2012). It achieves this by performing recursive tests on each attribute available. Once the root node attribute has been chosen the possible outcomes will form the branches to the next nodes. Each possible outcome will then be considered to be a new learning problem and by itself can be considered a sub-tree (Mac Namee and Kelleher, 2012).

If all the values in an internal node are either positive or negative (i.e. there is no mix of values) then the node is considered a leaf node and no further tests are performed beyond that node. Therefore, from Figure 4.1, all customers whom are middle aged had a positive “Yes” value; hence there is no need to split this sub-tree further.

The decision tree stops growing once all nodes only contain either positive or negative classification values. If the tree is quite large it can be pruned to become smaller, if based on analysis it is determined to potentially over-fit the data. Analysis of measures such as the average squared error or the misclassification rate on the validation data for any given number of leaves aid in such decisions.

4.1.2 k Nearest Neighbour

In contrast to decision trees, k nearest neighbour classifiers are considered “lazy learners”. It waits until it is presented with test data before deciding upon the classification, based on the similarity of the data to its training data (Han et al., 2012).

Lazy learner classification methods perform most of their processing when predicting the classification. Very little work is performed to train lazy learners. As such, in a real-time environment, lazy learners can be slower than eager learners to provide a classification. However, they have the advantage of supporting incremental learning, each test row can be re-classified based on knowledge gleaned subsequent to the initial prediction[11].

[11] Recommendation systems ubiquitous on most retail website are an example of incremental learning in a real-time environment.

According to (Han et al., 2012), nearest neighbour classifiers learn by comparing test data with training data which is similar. Each row of data corresponds to a point in an n-dimensional space, where n is the number of attributes in the row. Figure 4.2 illustrates an example of a 2-dimensional space based on average driving distance and Greens Hit In Regulation. k nearest neighbour classifiers assume that cases close to each other have a tendency to belong to the same class.

When new data is presented to the model, it searches for *k* training rows closest to the new data – the nearest neighbours. Euclidean distance is used to define which training rows are closest to the new data. The reader is referred to the bibliography for details of further reading on distance measures.

Accuracy of k-nearest neighbour models can depend on how separated the training classes are. When new data is presented which places the data very close to instances of both classes, it may be necessary to ensure the model picks two, three or more of its closest neighbours in order to best decide the new classification. The new data is then assigned the most common class amongst these neighbours (Han et al., 2012).

As discussed by (Mac Namee and Kelleher, 2012) the value of *k* is suggestive of the complexity of the model. Overfitting may occur if *k* is of a low value and inversely underfitting may happen if the value of *k* is too large.

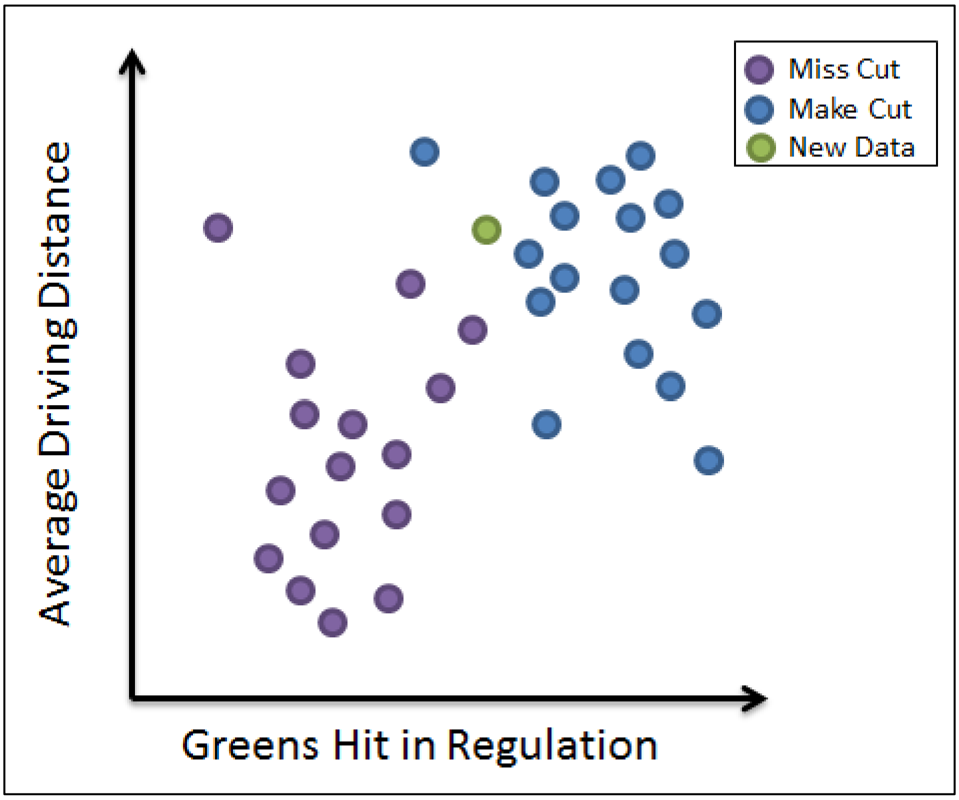


Figure 4.2: Example of k nearest neighbour 2 dimensional space

4.1.3 Clustering

Clustering is a method of unsupervised learning which divides datasets into subsets, or clusters. Each cluster will contain for example, groups of golfers which are similar to each other, but dissimilar to golfers in other clusters (Han et al., 2012). Business uses for clustering include the identification of unique customer segments which can then have tailored treatment strategies.

By creating clusters and visualising them in 3D scatter plots it may be possible to discover previously unknown groupings. Furthermore, analysis of these groups could detect differences in behaviour which leads to separate models being created for each discrete grouping.

Figure 4.3 illustrates five distinct groups of professional golfers mapped on the three dimensions of driving accuracy, average money earned and average greens in regulation. The colours of the data points indicate the cluster which the golfer has been assigned. Figure 4.4 reveals the clusters of golfers with greater separation based only on average money earned and average greens in regulation.

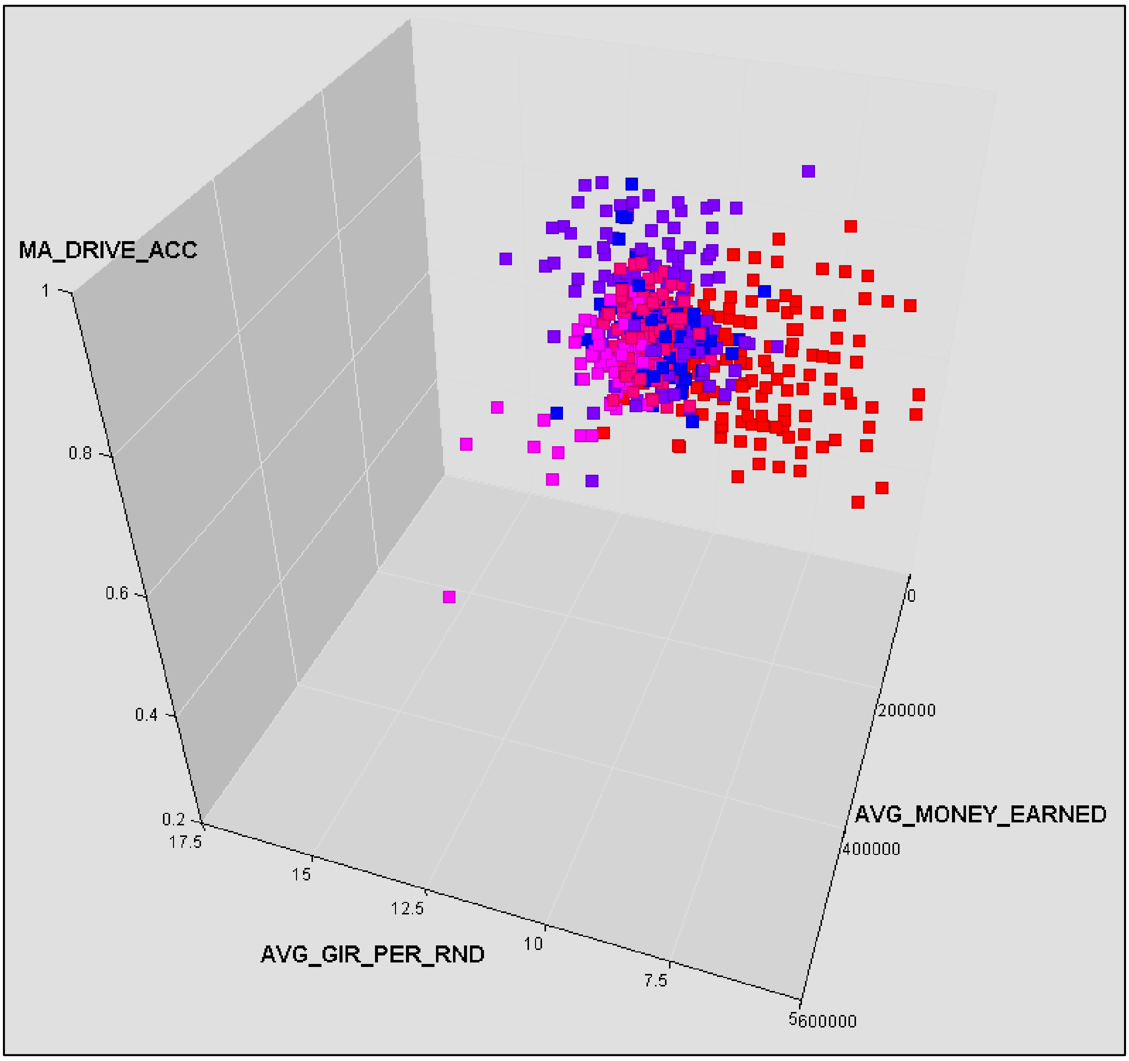


Figure 4.3: 3D scatter plot of golfer clusters

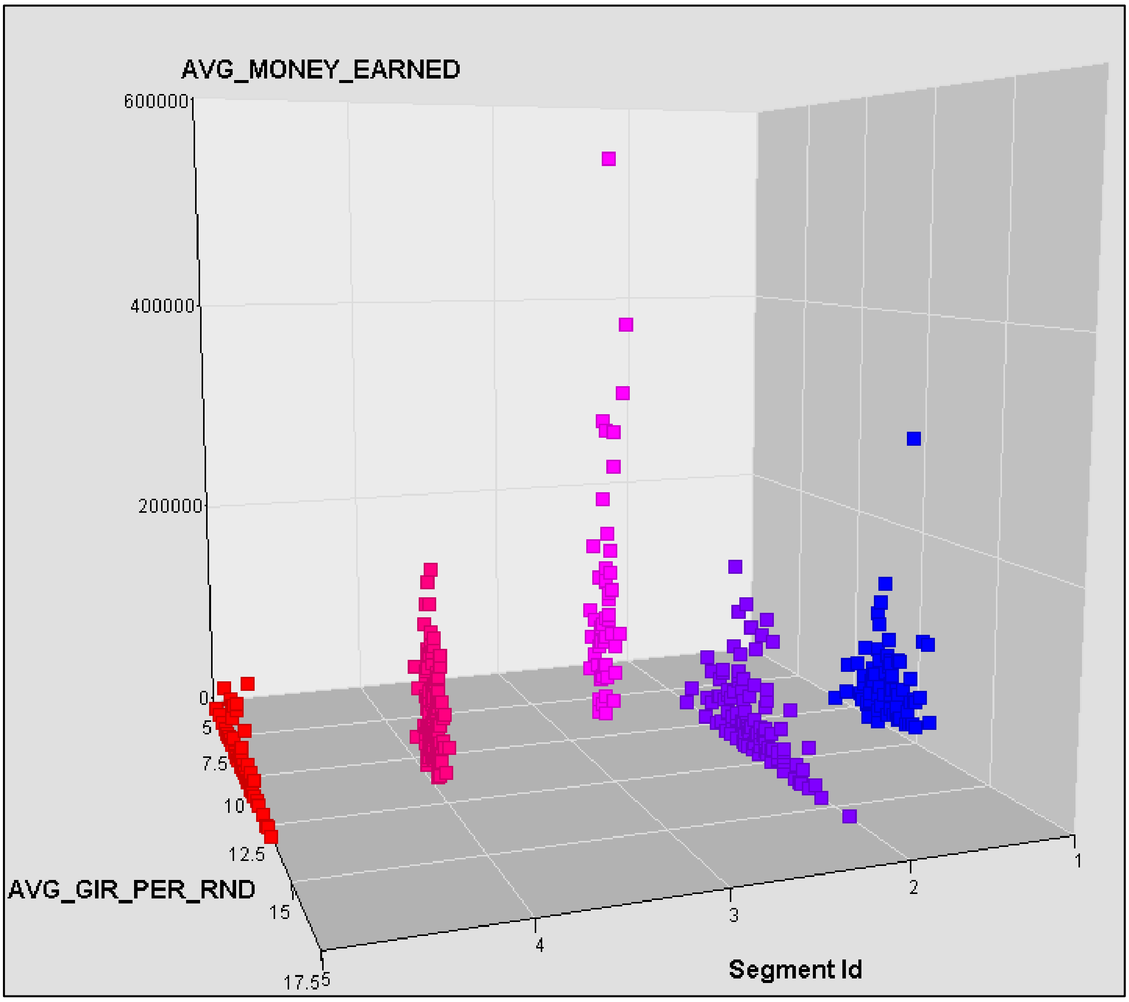


Figure 4.4: Visual separation of golfer clusters

4.1.4 Neural Networks

Neural Networks were originally designed to mimic the workings of neurons in the human brain. A neural network is made up of input and output units, connected by links, an example of which can be seen in Figure 4.5. Each link has an associated weight *W* which indicates the strength of the connection between the units (Mac Namee and Kelleher, 2012).

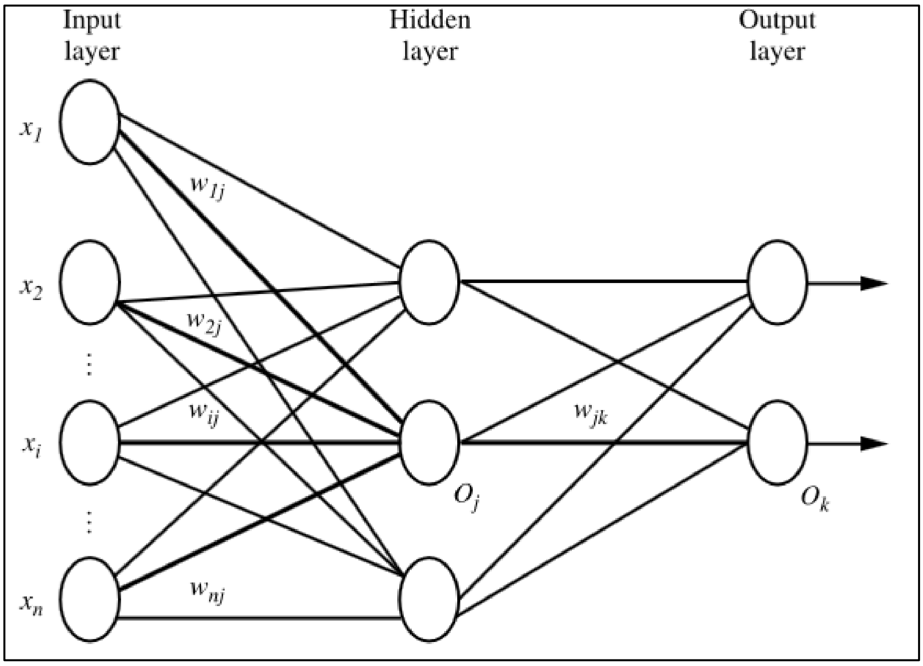


Figure 4.5: Illustration of multilayer "feed-forward" neural network (Han et al., 2012)

Neural Networks can either be feed-forward networks (as in Figure 4.5) or recurrent networks (Mac Namee and Kelleher, 2012). Feed-forward networks are either single layer or multi-layer networks or "perceptrons" and will be the focus of the remainder of this sub-section.

Single Layer perceptrons consist of an input layer and an output layer. The addition of a "hidden" layer creates the multi-layer perceptron which enlarges the hypotheses space (Mac Namee and Kelleher, 2012). A neural network can have any number of hidden layers but additional layers will add to the complexity of the network. The number of hidden layers chosen is typically by a manual process and a result of building many iterations of a network.

The input data is fed into the input layer where the weighted values are calculated and then sent concurrently to the hidden layer. The calculations from the hidden layer form the input for the output layer which provides the predicted classification for each row of data (Han et al., 2012).

Neural Networks are slow to train and their results can be difficult to interpret due to the use of hidden layers and the complexity of the algorithms behind them. Neural

Networks do not perform any automatic data selection; all inputs from a dataset will be used in the training of the network (Han et al., 2012).

Where there is a large number of attributes in a dataset, the use of dimension reduction methods such as those outlined in Chapter 4.2 would be beneficial. A lower number of units in the input layer should result in a quicker training time.

4.2 Dimensionality Reduction

As broached in Chapter One, the "curse of dimensionality" is often an issue which needs to be addressed, especially when choosing which dimensions out of potentially hundreds, or even thousands are most important for the data analysis problem to be addressed. Table 1.2 has outlined the number of attributes (also known as features) in each of the available datasets.

For this project it will be necessary to reduce the number of features used by machine learning algorithms in order to understand which features are key to describing a golfer's performance. While automatic techniques (as outlined in the following paragraphs) are the most efficient means to dimensionality reduction, there are occasions when manual selection is desirable or necessary. The inclusion of derived features could be included in order to discover their effect of models produced, if at all.

Principal Components Analysis is a dimension reduction technique which determines which features are redundant as the variation in the data they explain is contained in other features. Thus, by reducing the feature space it is possible to have more refined datasets which have automatically been chosen for their importance to the dataset.

Synonymous with regression, the use of Stepwise selection is another technique which can be utilised in order to ascertain which features are statistically significant in explaining any variation in a model (Han et al., 2012).

Stepwise forward selection starts with an empty set and determines the "best" attributes in an iterative process until the features remaining are no longer significant to the model. Conversely, stepwise backward selection begins with all attributes in the dataset and then removes the “worst” attributes. It is also possible to have a combination of forward and backward selection (Han et al., 2012).

Decision Trees by their nature also provide dimension reduction. By considering which features split the data in the best way, dimension reduction becomes a bi-product. Any features which do not appear in the decision tree can be considered redundant to the dataset (Han et al., 2012).

4.3 Statistical Analysis

The descriptions of the following statistical topics are intended to provide a brief introduction to the methods used in some of the previous studies reviewed, as well as in Chapter 5 of this project. The reader is referred to the bibliography for details of sources of further reading on these topics.

4.3.1 Hypothesis Testing

Hypothesis Testing is used when there are two competing claims regarding the value of a parameter such as the mean or standard deviation of a sample distribution (Montgomery and Runger, 2011). When testing a claim, there are two hypotheses, the null hypothesis (H_0) and the alternative hypothesis (H_1). The null hypothesis is the claim that is assumed to be true (D’Arcy, 2012).

Z-Tests & T-Tests can be used to test hypotheses. For sample sizes of 30 or larger, (whether they are normally distributed or not[12]), the Z statistic can be used to determine within a specified confidence level (typically 95%) that a claim is valid or not. T-Tests are used for sample sizes less than 30 but the sample means must be from a normal distribution in order for conclusions to be valid (D’Arcy, 2012).

Hypothesis testing can be performed on a single sample to test for example, whether or not a PGA Tour golfer’s average earnings are equal to $500,000. It can also be used to compare two independent samples known as a two sample t-test or two dependent samples, known as a paired t-test.

[12] The Central Limit Theorem allows non-normally distributed sample means to be tested as if normal.

When testing a hypothesis with a certain level of confidence, what is known as one-tailed or two-tailed tests can be performed. For a one-tailed test, the hypothesis is that the sample statistic is either greater than, or less than the mean. So, as per the right hand diagram in Figure 4.6, the shaded area under the curve is the region where the t-statistic lies in order to reject the null hypothesis.

For a two-tailed test, the hypothesis states that the value must be different to the mean (for example). In this instance, the confidence level is halved so that a t-statistic in either of the shaded areas, as per the left hand diagram in Figure 4.6 rejects the null hypothesis. This means that it is more difficult to reject a null hypothesis for a two-tailed test, as there is less area under the curve in which the t-statistic can lie.

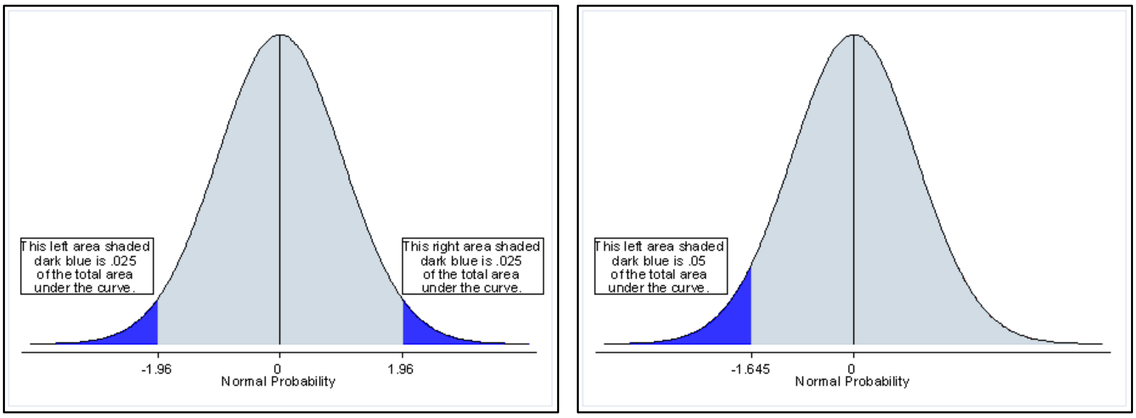


Figure 4.6: Two Tailed T-Test (left) and One Tailed T-Test (right)

Source (http://www.ats.ucla.edu/stat/mult_pkg/faq/general/tail_tests.htm)

4.4.1.1 Analysis of Variance

When testing a chosen parameter (mean, standard deviation etc.) on more than two samples[13], Analysis of Variance (ANOVA) is used. ANOVA is a model which describes whether, for example, at least two of the sample means are significantly different, and how much of the variability (if any) is explained by the "treatment" (the different populations) (Rumsey, 2007). The null hypothesis for ANOVA is $\mu_1 = \mu_2 =$

[13] ANOVA can also be used to compare just two samples as an alternative to a t-test

$\mu_3 = \mu_4$ and the alternative hypothesis is that at least two of the sample means are different.

For example, a golfer wishes to discover if there is any difference in distances hit with four different makes of 5-iron. Once enough data had been recorded from each club, ANOVA could be used to determine if there was a difference between at least two of the clubs.

While ANOVA is useful to determine if there is a difference between many samples, further tests such as Fisher's paired differences and Tukey's simultaneous confidence intervals are required in order to pinpoint which samples are different (Rumsey, 2007).

4.3.2 Contingency Tables

Contingency Tables are used to test whether or not two categorical methods of classification are independent of one another (Montgomery and Runger, 2011). By cross classifying the data into a table as in Table 4.1, a Chi Square test of independence can be carried out. The null hypothesis is that both variables are independent of each other, the alternative hypothesis being that there is some association between the variables.

	Belief in Afterlife	
Gender	Yes	No
Female	509	116
Male	398	104

Table 4.1: Contingency Table - Belief in Afterlife (Agresti, 2007)

According to (Agresti, 2007), the two variables are statistically independent if the population distributions (from the example in Table 4.1) of Gender are identical at each level of "Belief in Afterlife".

4.3.3 Regression Analysis

Regression analysis comes in many forms but the two most pertinent to this project, Linear & Logistic regression will be outlined briefly here. This will give context and comprehension to the results of previous research and some of the experiments carried out for the project.

4.4.3.1 Linear Regression

In simple terms, when there is a linear relationship between two continuous variables, linear regression analysis allows the prediction of the value of a dependent or response variable, Y based on the values of single or multiple predictors (also known as regressors) $x_1..x_n$ (D'Arcy, 2012). When a single predictor is used it is known as Simple Linear Regression and it is called Multiple Linear Regression when a model contains more than one predictor.

A linear relationship between two variables can be determined using the Pearson Correlation Coefficient. As illustrated in Figure 4.7, the coefficient is a number between -1 and 1, where a value of -1 is a strong negative correlation and 1 is a strong positive correlation. According to (D'Arcy, 2012) a coefficient value of between -0.3 and 0.3 should be considered a weak correlation.

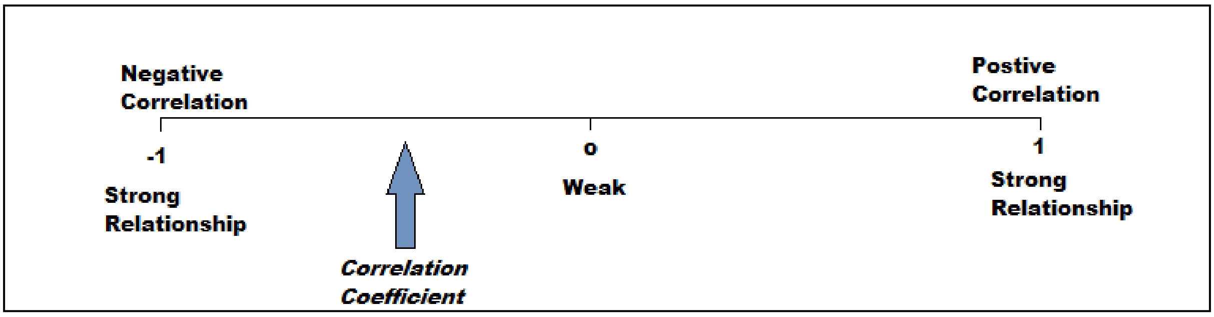


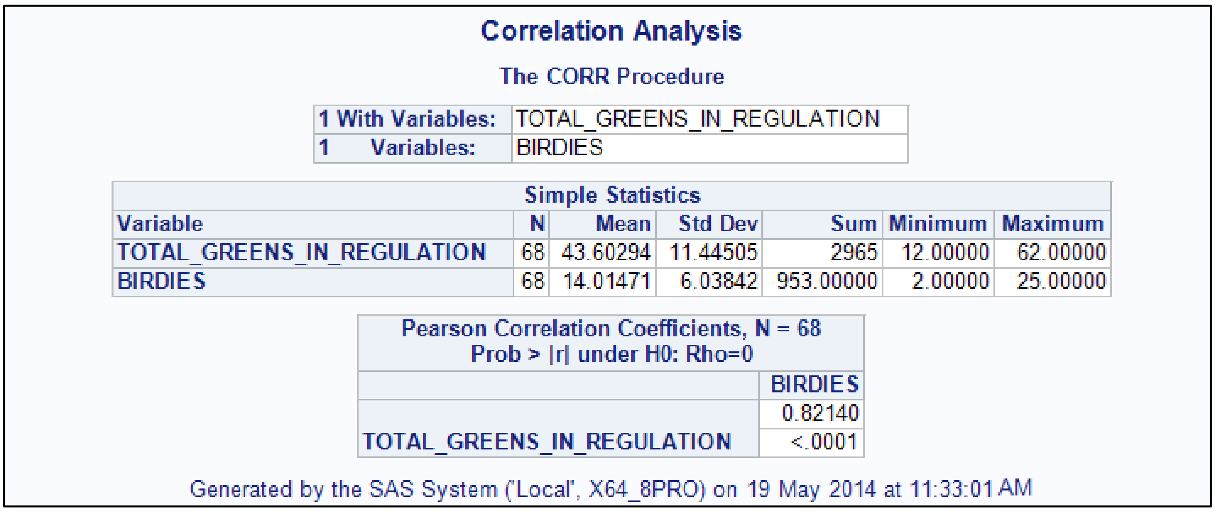
Figure 4.7: Pearson Correlation Coefficient (D'Arcy, 2012)

If a strong correlation between the dependent and (potential) predictor variables has been established then a regression model can be built to create an equation which can be used to predict the dependent variable Y. The equation used to describe the simple linear regression model is: $Y = \beta_0 + \beta_1 x + \varepsilon$. β_0 represents the intercept of the line and β_1 represents the slope of the line (D'Arcy, 2012). These are known as either the parameter estimates or the regression coefficients. These values will remain constant, only the value of x will change and therefore provide a varying prediction for Y .

- **Golfing Example**

Using Shotlink Event Level Data for Rory McIlroy from 2009 onwards, the correlation between greens in regulation and the number of birdies scored will be analysed.

Firstly the correlation between the two variables is analysed. As can be seen in Figure 4.8 which is the output of the Correlations node from SAS Enterprise Guide, the Pearson Correlation Coefficient score is 0.82140. This indicates a strong positive correlation between the number of birdies scored and the total number of greens hit in regulation.



Correlation Analysis

The CORR Procedure

1 With Variables:	TOTAL_GREENS_IN_REGULATION
1 Variables:	BIRDIES

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
TOTAL_GREENS_IN_REGULATION	68	43.60294	11.44505	2965	12.00000	62.00000
BIRDIES	68	14.01471	6.03842	953.00000	2.00000	25.00000

Pearson Correlation Coefficients, N = 68
Prob > |r| under H0: Rho=0

	BIRDIES
TOTAL_GREENS_IN_REGULATION	0.82140 <.0001

Generated by the SAS System ('Local', X64_8PRO) on 19 May 2014 at 11:33:01 AM

Figure 4.8: Output of Correlation Analysis between Greens in Regulation and Birdies

This positive correlation can also be visually assessed using a scatter plot as shown in Figure 4.9.

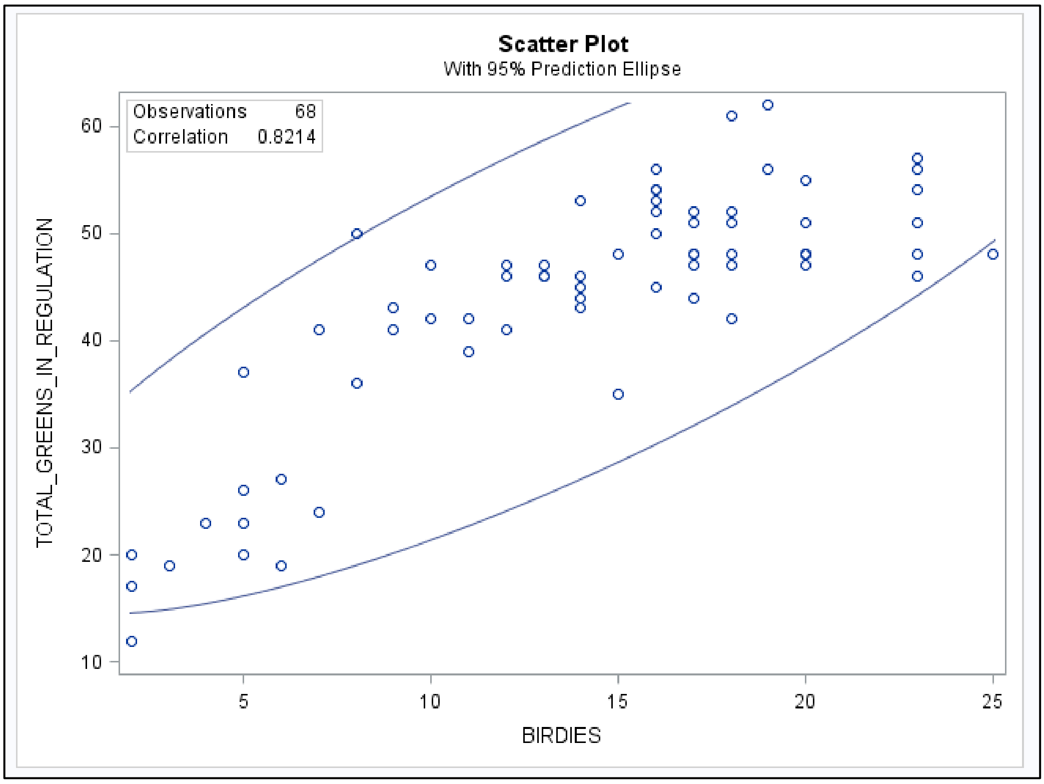
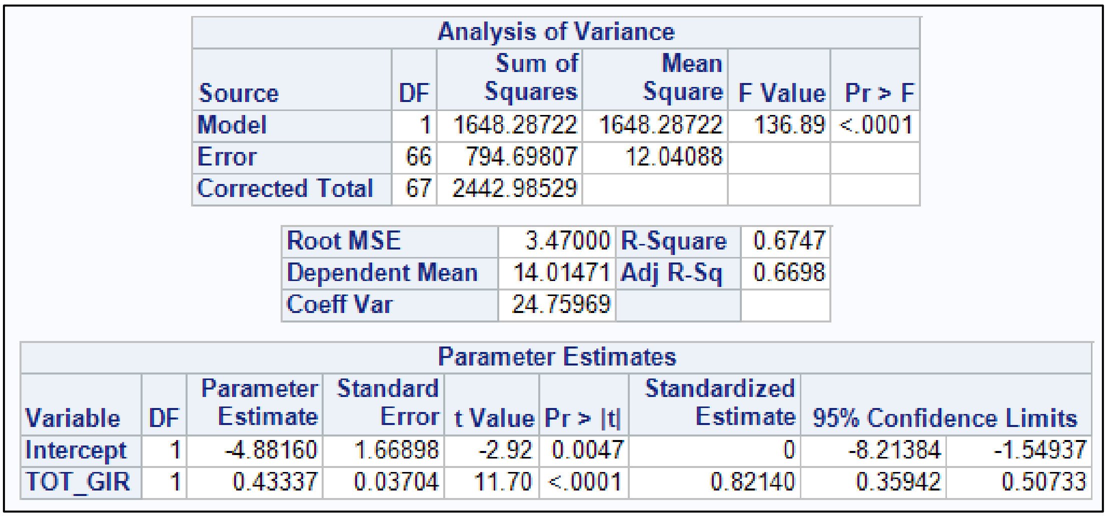


Figure 4.9: Scatterplot of Greens in Regulation and Birdies

This strong correlation implies that the total greens hit in regulation should be a useful regressor in an attempt to predict the number of birdies that McIlroy will score per tournament.

The simple linear regression model is built where "birdies" is the response variable, Y and total greens in regulation is the single predictor x. The results of the model are available in Figure 4.10.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1648.28722	1648.28722	136.89	<.0001
Error	66	794.69807	12.04088		
Corrected Total	67	2442.98529			

Root MSE	3.47000	R-Square	0.6747
Dependent Mean	14.01471	Adj R-Sq	0.6698
Coeff Var	24.75969		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	95% Confidence Limits	
Intercept	1	-4.88160	1.66898	-2.92	0.0047	0	-8.21384	-1.54937
TOT_GIR	1	0.43337	0.03704	11.70	<.0001	0.82140	0.35942	0.50733

Figure 4.10: Simple Linear Regression results

Along with the parameter estimates, the results of the ANOVA test is provided to show whether or not the model is better than just using the mean number of birdies scored. The P-value of < 0.0001 suggests that this model is indeed better than using the mean.

The R-Square value of 0.6747 means that greens hit in regulation explains 67% of the variation in the number of birdies Rory McIlroy scores.

Finally, in order to predict the number of birdies scored given the total greens hit in regulation, the parameter estimates can be substituted in the simple linear regression equation referenced above.

The prediction formula is:

$$\textit{Number of birdies} = -4.88160 + 0.43337 * (\textit{Total Greens In Regulation})$$

Multiple Linear Regression is very similar to simple linear regression; there are just multiple variables which can be used as predictors. Using the R-Square values for each predictor, it is also useful for discovering which variables are most important to a model and so can be used as a dimensional reduction technique which can be beneficial when creating inputs for other models such as neural networks. It is also worth noting that for a multiple regression model, the adjusted R square value should be used when judging how much variability is explained by the model, even though R square is also typically provided by software.

4.4.3.2 Logistic Regression

Logistic Regression is used to predict the probability of a categorical (generally binary) response such as whether or not a golfer will miss the cut – Yes or No (or 1 or 0). Unlike Linear Regression which predicts a value for the response variable, Logistic Regression predicts the probability that the response variable will be Yes or No (Rumsey, 2007), (Agresti, 2007). Similar to linear regression, logistics regression can involve single or multiple predictors.

Logistic regression is also known as a Logit model because it models the Logarithm of the Odds, that is the odds of the response variable (Miss Cut) being 1. The odds are the ratio of the probability that the response variable will be 1 to the probability that the response variable will be 0 (Agresti, 2007).

Figure 4.11 illustrates a hypothetical logistic function of the probability of a golfer missing the cut based on their combined score for rounds 1 and 2 (for example). The inverted S curve implies that a change in x will have less of an impact when the probability is already closer to 1 or 0 than if x was in the range of say 4 to – 4 (Der and Everitt, 2007). In practical terms this means that whether a golfer is 6 over par after 2 rounds or 10 over par makes very little difference to their chances of missing the cut – they are both highly likely to miss it.

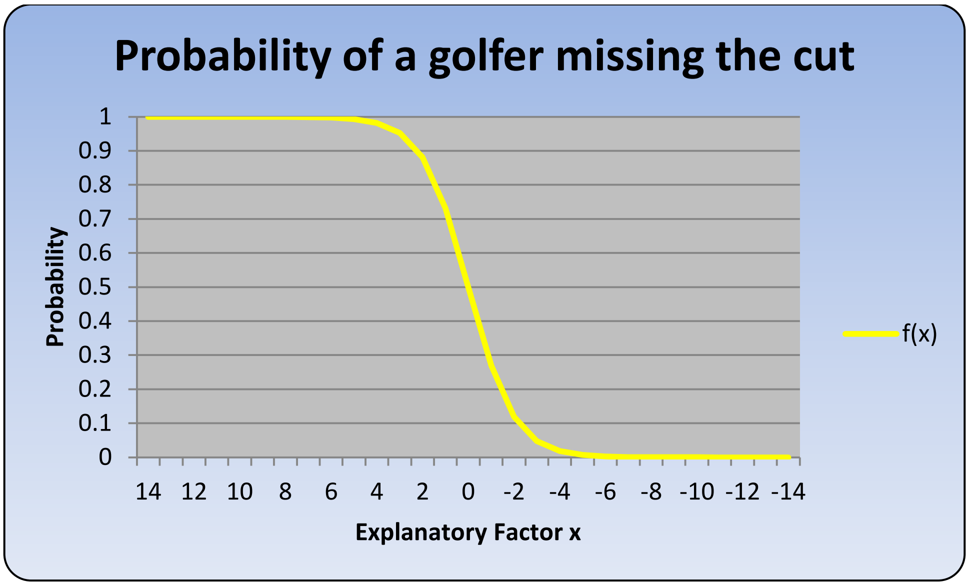


Figure 4.11: Illustration of Logistic Regression inverted "S" curve.

The equation for a single predictor logistic regression model is: $\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta_1 x$ and can be re-written in terms of probability as: $= \frac{e^{\beta_0+\beta_1 x}}{1+ e^{\beta_0+\beta_1 x}}$. The sign on β_1 indicates whether the S curve is inverted or not, so in the case of Figure 4.11, β_1 is negative giving the inverted S shape. The absolute value of β_1 specifies the steepness of the curve. High values will have a steep curve and low values will have a more elongated S shape.

Interpretation of Logistic Regression Models

Based on the output results from Logistic Regression in SAS Enterprise Guide, the tests in Table 4.2 indicate whether or not the chosen model is a good fit.

Test	Null Hypthothesis	Description
Hosmer-Lemeshow Goodness-of-fit	Model is a good fit	If P value is greater than 0.05 then accept the null hypothesis that the model is a good fit
Testing global null hypothesis : BETA=0	All of the predictors' regression coefficient are equal to zero in the model.	A P Value less than 0.05 is good. Conclude that at least one of the regression coefficients in the model is not equal to zero.[14]
Analysis of Maximum Likelihood Estimates (Pr > ChiSq)	The individual regression coefficient is zero given the other variables are in the model	If P values is less than 0.05 then conclude that the variable is significant to the model (statistically different from zero)

Table 4.2 Logistic Regression tests and interpretation (Der and Everitt, 2007)

4.4 Conclusion

The chapter has presented details of the various machine learning algorithms which will be used in the implementation of the predictive experiments of the project. The basic workings of each algorithm have been discussed, coupled with relevant golfing examples to aid comprehension. Similarly it has provided an overview of the statistical methods to be implemented in the project.

The next chapter will introduce the experiment methodologies and designs in detail, together with information regarding the preparation of the Shotlink dataset and the common features throughout all the experiments.

[14] According to (Agresti, 2007) , if β = 0 the curve becomes a horizontal line which would indicate that Y is independent of X. So values where β is close to zero would indicate a weak dependence between Y and X.

5. EXPERIMENT DESIGN

This chapter will discuss the details of the design of all experiments performed. It will outline preliminary statistical analysis on the data, the predictive models created subsequent to this analysis and testing of ideas introduced in Chapter Two.

The primary objective of this project is the prediction of the cut line for any PGA Tour tournament. Numerous approaches, experiments and ideas will be used in order to achieve the greatest accuracy possible within the constraints of the project.

In addition to the main research topic, further experiments are outlined to determine the answers to the following questions:

- How similar is Jordan Spieth's performance to Rory McIlory's at the same age?[15]

- How much influence does a caddie have in a professional golfer's performance?

- Does a golfer's age have an effect on their ability to make the cut?

5.1 Shotlink Data Understanding

Of the available levels of data, Event Level detail was chosen as the main dataset to work with, since Stroke or Hole level detail was deemed to be too granular to be of use given the time scale of the project. The individual specifics of each shot are unlikely to yield any insight for the questions posed in this project that could not be attained at a higher aggregate level.

Event level detail is an aggregation of Round level and contains on row per tournament, per competitor. The 190 columns it contains can be found in Appendix B.

[15] Since the 2013 season, Jordan Spieth has been consistently compared to Rory McIlory. http://www.golfwrx.com/139451/spieth-to-supplant-mcilroy-as-the-next-great-young-golfer/

The event details from 2009 to 2014 where loaded into a MYSQL database and appended into one table. This table "SL_EVENT_DETAIL" is the base for all queries used to create datasets for training and testing models.

Some hole level detail was loaded into the database in order to carry out analysis of how far an approach shot finished from the pin. The inspiration for this analysis was based on the research by (Broadie, 2011a) which outlined the case for Strokes Gained - putting. The basis for the analysis was to determine if there was a relationship between a golfer's average distance to the pin after an approach shot and whether they made the cut or not.

The box plot[16] in Figure 5.1 compares the average distances to the pin after approach shots which landed on the green for Robert Allenby and Angel Cabrera in the 2013 season. It can be seen that Angel Cabrera's distances where typically further from the pin than Robert Allenby's.

When the tournament results of these two golfers are compared however, there is a marked contrast. Robert Allenby missed the cut 17 times while Angel Cabrera missed only 3 cuts. This contrast between the average distance to the pin and the golfer's results was observed in many cases. Further investigation will be required in order to determine the value of hole level detail but this will not be possible given the time constraints of this project.

From this initial analysis it was decided not to incorporate the average distance from the pin metric into the experiment datasets, although it could be useful for future research on different questions.

[16] For details on how to interpret a box plot, the reader is referred to Appendix D.

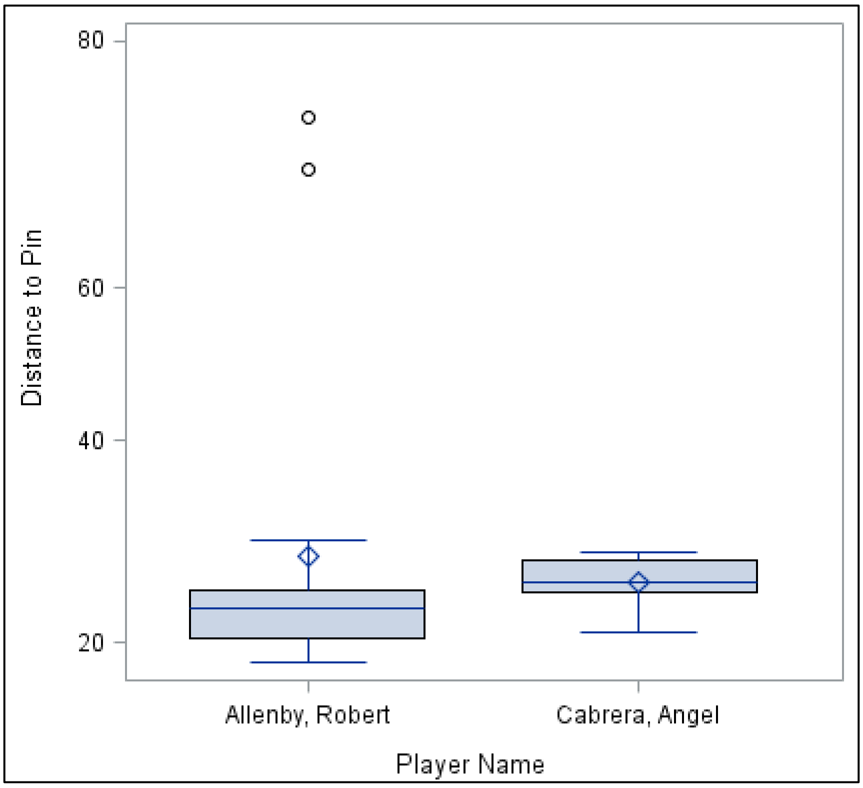


Figure 5.1: Average distance (feet) to the pin from approach shots for 2013

5.1.1 Data Quality

Data Quality within the Shotlink data is of a good standard but some minor issues were encountered during the loading and use of the data which are briefly outlined.

- **Null and Blank Values**

Null values appear quite frequently throughout the dataset for numeric data columns. In most cases a column is assigned a null value as data does not exist for that player. An example of this is the “Money” column which contains numeric data for all the golfers who made the cut and therefore earned prize money in that specific tournament. Any golfers who missed the cut received no prize money but rather than assign a value of zero to this column, it is assigned a null value.

Occasionally blank values, (which are just whitespace) appear erratically throughout the dataset. Though limited to numeric data fields, there is no consistency to which fields or the scenarios in which it appears.

Null and blank values need to be transformed in order for SAS to treat the relevant columns as numeric fields and to ensure the rows are not rejected due to missing data. In almost all cases the assignment of a zero value to replace nulls or blanks was the correct course of action and did not cause any skewing of data.

- **Miscellaneous Observations**

A value of 999 is assigned to columns were the player missed the cut. Examples include the "FINISH_POSITIONNUMERIC" column which give the final finish position of each player who made the cut and then a value of 999 for the players who missed the cut.

This value is also assigned to fields such as the number of birdies, eagles and greens in regulation when the player missed the cut. This is an obstacle to the use of such fields due to the fact that in spite of playing two rounds of golf, these metrics do not exist at event level for these golfers.

In these instances, substitution of the 999 value with a zero or another value would be ill-advised. Techniques such as imputation to calculate a mean value of the field in question (excluding the 999 values) could produce an inaccurate replacement value. In order to obtain these metrics and ensure accuracy, it would be necessary to observe the round level data for each golfer that made the cut.

Both column names and non-numeric fields had occurrences of leading or trailing spaces. For the column names, manual removal of the spaces was sufficient as it was a once off process to load the data. The TRIM() function was used where necessary to remove both trailing and leading spaces on columns included in training or evaluation data.

5.1.2 Derived Data

The Shotlink Event Level detail data can be divided into the categories outlined in Table 5.1. Each category is then further broken down to lower levels such as distance ranges in putting from 3 feet, 4 feet etc.

Main Subjects in Event Level Detail	
Round Scores	Approach Shots
Round Positions	Shot Locations
Scoring Averages	Scrambling Metrics
Score Types	Sand Saves
Driving Metrics	Putting Metrics

Table 5.1: The main subject areas in Event Detail

From preliminary analysis of the data, additional columns were derived with the goal of producing variables that may explain some extra variability in a golfer's performance regarding making or missing the cut.

- **Form Score**

Inspired by the KCS metric described in Section 3.1.3, the "Form Score" measure tracks a player's proximity to the cut line over time. By calculating how many strokes above or below the cut line a player was, after 2 rounds, a moving average over the previous X number of tournaments is calculated. This provides a simple means to tracking a player's performance relative to making the cut or not.

The Form Score attribute could also be consider as means of tracking psychological momentum discussed in Section 3.3.3. By analysing the trend of the form score over time, this could indicate whether a golfer is gaining or losing momentum.

- **Final Round Meltdown**

The final round meltdown indicates when a golfer has suffered from a "poor performance" in the final round of a tournament. "Poor performance" is defined as when a golfer shoots their final round with 5 or more strokes than the previous round. In addition to this, their final round finish position must be lower than their position in the penultimate round.

- **Moving Averages**

With the exception of Experiment 1, all averages are moving averages. That is, they are calculated on a "sliding window" basis that only the last X[17] number of tournaments count towards the calculation of the average.

- **Tournament ID**

This is an amalgamation of the two fields, TOURNAMENT YEAR and TOURNAMENT_NUMBER. It is used as an index throughout all SQL queries created but it is not intended for use in the training of models and will be distinctly assigned the role of "ID" in SAS Enterprise Miner to ensure it unutilised by any models.

- **Just Missed / Just Made Cut Count**

The number of times a golfer just made the cut by one stroke or just missed the cut by one stroke is counted in separate variables.

- **Cut Missed Ratio**

This is the proportion of cuts missed to cuts made over the last X number of tournaments. It indicates the propensity of the golfer to miss or make the cut and could be considered another gauge of form.

- **Finish Position Groupings**

The number of times a golfer has finished in the Top 10, Top 20, Top 30 etc.

5.2 Software

SAS Enterprise Miner and SAS Enterprise Guide will be used to design, build and implement all the models of the forthcoming experiments. SAS Enterprise Miner is a specific Data Mining tools with many features and machine learning algorithms. Experiments in this project will utilise the appropriate classification machine learning algorithms provided by this software.

[17] The value of X can be between 2 and 5.

SAS Enterprise Guide is a statistical analysis tool which will be used to discover associations, correlations, distribution analysis for data exploration. It will also be used to create most of the Logistic Regression models for this project in tandem with the design and implementation of individual statistical experiments.

Data Manipulation and transformation will be performed using SQL queries run against the aforementioned MySQL database created to store the data for this project.

5.3 Choice of Models

The following models have been chosen for use in the predictive experimentation part of this project:

- Decision Trees
- Neural Networks
- K-Nearest Neighbour
- Logistic Regression

Decision Trees were chosen for a number of reasons. It is a very mature, tried, tested and trusted method for solving classification problems. It is quick to train and to classify data and is known to be quite accurate (Han et al., 2012). Decision Trees can also be easily interpreted and the tree's rules can be easily extracted and implemented in programs using CASE statements, for example.

K-Nearest Neighbour will be used in this project primarily due to the fact that it trains and classifies based on case similarity (Han et al., 2012). It is naturally suited to the predictive experiments as, by their very nature these experiments are founded on the idea that similar past performance from any golfer will aid in a classification for new data.

Neural Network models will also be built because they are well suited to classification problems. As Neural Networks iteratively learn and adjust their weights during training, it is expected that will create models which are highly accurate. While they take some time to train, with the resources available this is not expected to be a disadvantage.

Logistic Regression has been chosen for this project for a number of reasons. Firstly it is more synonymous with the world of Statistics than Data Mining so it introduces another discipline which provides for classification solutions. Secondly it is perfectly suited for a binary classification problem which is exactly what the predictive experiments for this project are.

5.4 Predictive Models

The main aim of the predictive experiments is to forecast whether or not a golfer will miss the cut in any given tournament. The following experiments outline the different approaches taken in order to discover the most accurate approach.

5.4.1 Collective Performance Methodology

The first set of experiments draws inspiration from (Lewis, 2003) and general classification problems / exercises. In the sport of baseball, one of the main ideas behind SABERMETRICS is that every baseball shot as been played thousands of times before and the outcomes, or finishing positions on the field for these shots has been recorded and measured (Lewis, 2003). Based on this finish position it is possible to see what happened next and how often each specific event happened. If an event occurs a high majority of the time, then the predicted result is that event.

For example, if a batter has two strikes and hits their 3rd pitch which lands in left of outfield, the statistics will show that say 70% of the time the batter will make 2 bases.

The key to this approach is that it is the combined collective data which is used to measure each player, rather than using only that player's data. This is because it is possible that an individual never hit a shot with a particular outcome before but other players of similar skill have.

Likewise in a business setting, if an analyst seeks to determine which customers are likely to churn (leave the company) they analyse the behaviour of previous customers. The behaviour of these customers who churned, mixed with those who did not can be used to build models to determine which current customers are likely to churn, even if they never left the company before.

This is the premise of using past behaviour to predict future behaviour. The exhibition of key behaviours or attributes from independent observations can be used to predict the future behaviour of other independent observations.

For this project, models built using collective data will be based on event level data from 2009 onwards. Experiments based on this will determine if there is merit to the collective approach. As each golf shot has been played thousands of times before, it is hypothesised that if a golfer is experiencing similar performance to that of previously recorded performances from other golfers, then their tournament outcome should be predictable.

Collective Advantage

One of the key advantages to using collective data is that there is sufficient data to train and validate any model created. The more data available to train a model the more accurate (in theory) it should be.

In any data mining exercise the balance of the dataset has to be considered. A dataset has unbalanced classification if one of the target labels is rare in the training data. Take for example a dataset with 10,000 rows, of which 9,000 had a classification label of "Y" indicating that the golfer made the cut. As the remaining 1,000 rows are classified as "N" (indicating that the golfer missed the cut) this would be considered an imbalanced dataset for training purposes.

Fortunately, neither of the target labels of "Y" (Missed Cut) or "N" (Made Cut) are rare in the Shotlink data. Of the 12,097 rows of data available for training in the collective experiments, 7,327 (61%) rows are labelled as "N" and 4,770 (39%) are labelled as "Y".

5.4.2 Individual Performance Methodology

The second experiment methodology predicts a golfer's performance based solely on that individual golfer's previous performances, so no other golfer's performances are used to train the models. Models are built for each individual golfer and therefore are more specific to the nuances of the golfer in question.

While gaining the more personal touch, the drawback of building models based only on that golfer's past performances is the lack of data. In any given year, there are at most, 45 official PGA Tour tournaments which a professional golfer can enter.

So, from 2009 onwards the most tournaments a model can be trained with is 180. In practise, most golfers would play between 15 and 25 tournaments a year so even less data is potentially available. Generally, in order to build accurate models hundreds of rows of data are required in order to have sufficient data to train, validate and test the generated model.

Another potential issue is that a golfer's previous performances from 2 or 3 years ago will not necessarily be indicative of their current performances. As all their data will be used to build the model, the "old" data could have a negative impact on the model's ability to predict effectively.

Even if a golfer was more likely to miss the cut 2 or 3 years ago than in their more recent history, the factors which causes them to miss the cut could have changed so it is important to ensure newer data is given more relevance. In order to counterbalance this "old" data, it may be necessary to introduce a weighting to the data so that more recent tournaments are given more importance. Contrasting experiments will be implemented in order to determine whether the weighting of new data has a positive effect on model accuracy.

5.4.3 Shared Services

- Training data

For the collective experiments, training data will contain results from all previous tournaments from the beginning of the 2009 season up to the 2014 AT&T Pebble Beach National Pro-Am tournament. This amounts to some 12,000 rows and will be split into training and validation dataset partitions of varying proportions between 50/50 and 70/30 for training and validation respectively.

The training data used for the experiments using only individual training data are outlined in Table 5.7.

- Model Scoring

For the collective experiments, all models were scored using the entry list for the 2014 Northern Trust Open which took place between the 13th and 16th of February 2014. The tournament had 146 entrants, 77 of whom made the cut, competing for a prize fund of $6.7 million.

Assignment of the classification label will be performed on a manual basis based on probability scores predicted by each model. The reasons why this approach is necessary is discussed in detail in Sub-chapter 7.1

5.4.4 Experiment 1 – Granular Data

The idea behind this experiment is to use the previous tournament's performance metrics to predict the next tournament.

The approach for this experiment will be to transform the individual rows in the SL_EVENT DETAIL table into a training row. The training row contains all the relevant metric columns (170 numeric attributes) from the previous tournament coupled with an indicator column - "MISSED_CUT_IND" as the class label.

MISSED_CUT_IND is a binary classifier containing either "Y" (the golfer missed the cut in this tournament) or "N" (the golfer made the cut in this tournament).

Table 5.2 provides an example of the source data containing two rows of data for separate tournament results.

TID	START DATE	PLAYER NUMBER	R1	R2	R3	R4	MISSED CUT IND
2014010	25/10/2013	7066	68	68	73	70	N
2014020	05/11/2013	7066	73	73	999	999	Y

Table 5.2: Example of Source Data available in Experiment 1

Table 5.3 outlines the newly created training row. The training data consists of the scores from rounds 1 to 4 from tournament number 2014010 coupled with the actual outcome of tournament number 2014020 for this golfer (in this case that the cut was missed)**.** As such, the model will be able to utilise the actual outcome of the tournament to be trained with the data from the previous tournament.

TID	START DATE	PLAYER NUMBER	R1	R2	R3	R4	MISSED CUT IND
2014020	41583	7066	68	68	73	70	Y

Table 5.3: Example of training row for Experiment 1

Finally, in order to predict the next tournament which is unseen by the model, data which contain the round details for tournament number 2014020, as in Table 5.4 would be presented to the model for a classification prediction for tournament number 2014030.

TID	START DATE	PLAYER NUMBER	R1	R2	R3	R4	MISSED CUT IND
2014030	12/11/2013	28237	73	73	999	999	?

Table 5.4: Example of new unseen data awaiting classification

Dimension Reduction

In order to determine which of the 170 columns are most significant in explaining variability, the dimension reduction technique of Principal Components Analysis will be employed before models are trained. This will be achieved by first deriving the principal components which will in turn become the input for subsequent models. In SAS Enterprise Miner, the default number of principal components to be created is 20 and this will not be modified.

Once PCA has been deployed, the models to be built in SAS Enterprise Miner will be Decision Trees, K-Nearest Neighbour, Neural Networks and Logistic Regression.

5.4.5 Experiment 2 – Aggregation

This experiment will transform the source data to an aggregate level. The general concept is to take the numeric columns from X number of previous tournaments for each player and then create derived columns containing the sum and moving average of these metrics. Models will be built using the same techniques / algorithms as used in Experiment 1.

The experiment is designed in such a way that the data from two or more tournaments can be aggregated quite easily. For the purposes of this project, the maximum number of previous tournaments used is 5. Based on discussion with (Cosgrove, 2014), it was deemed that any more tournaments than this are unlikely to provide an indication of a professional golfer's current form.

Figure 5.2 illustrates the "sliding window" nature of the aggregate data for the previous two tournaments (x = 2) coupled with the outcome from the third event which is the one the model will train on.

Whatever the number of tournaments to aggregate is chosen, it has the limitation that is unaware of the time between the most recent tournament played and say the fifth most recently played. It is possible that a golfer played a tournament a week for five weeks, in which case the data would most likely indicate current form.

It is also possible that a golfer who is a dual tour member could have gap of two months or more (for example) between their five tournaments. In this example the data may not have the desired integrity and hence accuracy could decrease as it is not indicative of current form.

It is theorised however, that the collective approach of using all golfers' data to predict the performance of an individual golfer would smooth out any of these gaps in tournament play.

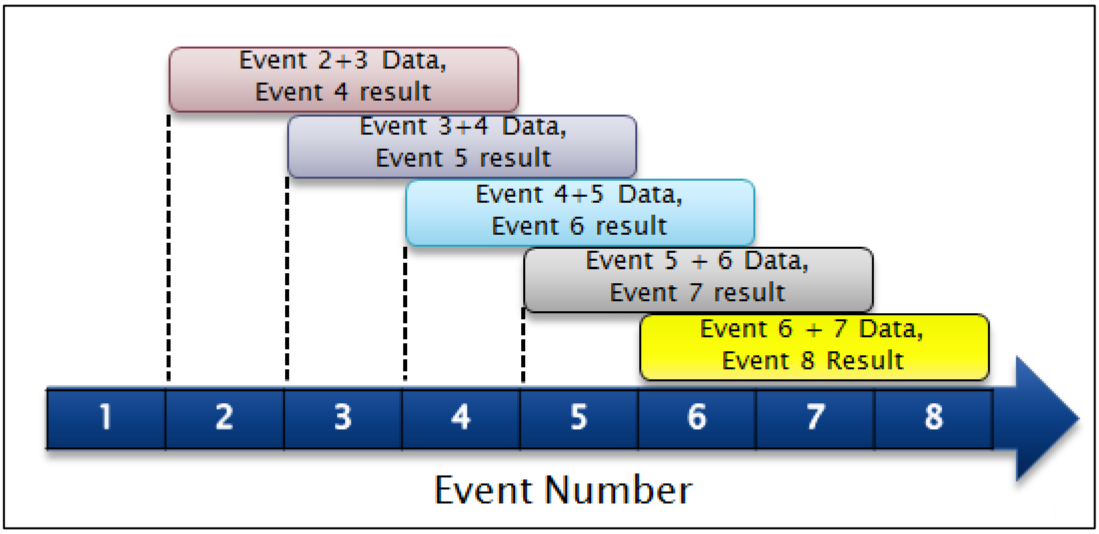


Figure 5.2: Illustration of the use of "Sliding Windows" to build aggregate data

The attributes used build the training dataset are outlined in Table 5.5. These attributes have been chosen based on research reviewed in Sub-chapter 3.2, coupled with the master columns from the main subject areas as noted in Table 5.1.

Attribute Name	
PREDICT_TID	AVG_PARS
MISSED_CUT_IND	AVG_BOGEYS
PLAYER_NAME	AVG_DOUBLES
AVG_GIR_PER_RND	AVG_OTHERS
AVG_DIST_TO_HOLE_AFTER_APP	AVG_TOTAL_GREENS_IN_REGULATION
MA_DRIVE_ACC	AVG_DRV_ACC
PREDICT_FINISHPOS	AVG_STROKE_AVERAGE_RANK
MOV_AVG_FINISH_POS	AVG_EAGLES_RANK
NO_OF_TOURNAMENTS_PLAYED	AVG_BIRDIES_RANK
NO_OF_CUTS_MADE	AVG_BOGEYS_RANK
NO_OF_CUTS_MISSED	AVG_BOGEY_AVOIDANCE_RANK
FORM_SCORE	AVG_GIR_RANK
LAST_TRN_4TH_MELT	AVG_SCRAMBLING_RANK
JUST_MADE_CUT_CNT	AVG_PUTTS_GAINED_RANK
JUST_MISSED_CUT_CNT	AVG_SAND_SAVE_RANK
CUTS_MISSED_RATIO	NO_OF_PUTTS
TOT_GREENS_IN_REGULATION	AVG_NO_OF_PUTTS
AVG_EAGLES	NO_OF_ONE_PUTTS
AVG_BIRDIES	NO_OF_THREE_PUTTS

Table 5.5 Attributes used to build training dataset for Experiment 2

5.4.6 Experiment 3 – Clustering

Moving closer towards using an individual's data to only predict said individual's performance, clustering will be performed on the 558 golfers who played in at least one PGA Tour event in 2013.

This is a similar approach to the one taken by (Savage, 2012) in that golfers were split into quintiles based on current ranking. In this experiment though, the segmentation will be based on the similarity of the golfers to each other, as decided by the clustering algorithms.

By splitting the golfers into groups of similar performance, it is theorised that the less generic collective approach should yield more accurate models. The code from Experiment 2 can easily be leveraged and adapted to account for smaller numbers of golfers. For the clusters identified, models will be trained on tailored aggregate data from Experiment 2 for each separate cluster.

Table 5.6 outlines the attributes which will be used to generate the golfer clusters. These have been chosen based on the performance analysis research reviewed in Chapter 3.2, together with some of the derived metrics referenced in Section 5.1.2.

Cluster Attributes	
NO_OF_PUTTS	AVG_MONEY_EARNED
NO_OF_CUTS_MADE	JUST_MADE_CUT_CNT
NO_OF_ONE_PUTTS	AVG_GIR_PER_RND
MOV_AVG_FINISH_POS	AVG_NO_OF_PUTTS
NO_OF_CUTS_MISSED	MA_DRIVE_ACC
NO_OF_THREE_PUTTS	JUST_MISSED_CUT_CNT
CUTS_MISSED_RATIO	

Table 5.6 Attributes used to build golfer clusters

The number of clusters chosen to be created is 5. As separate models will have to be trained for each cluster, this is a compromise between the time needed to create the training data for each cluster, the building of the models, and the scoring of them.

A higher number of clusters could create more accurate models for each grouping but it is theorised that the accuracy of each cluster should be higher than for the more general collective approach which would be sufficient to prove the concept is worthwhile, at the very least.

Models for each cluster will only be trained using the data from golfers in that cluster. The predicted probabilities of each model will be collated into one combined score and sorted in order of lowest probability of missing the cut to the highest.

5.4.7 Experiment 4 – Individual Performance

By utilising a golfer’s own unique past performance data to only train a specific model for that golfer it is theorised that models produced will be more accurate. Taking into consideration the comparatively small amount of data available for each individual golfer, it is still deemed worthy of experiment and analysis.

From the 2013 season, 10 golfers were chosen as subjects for this experiment. These golfers, together with additional relevant details and the number of training rows available for the models are detailed in Table 5.7.

PLAYER NAME	TOURNAMENTS PLAYED	CUTS MADE	CUTS MISSED	CUTS MISSED RATIO	TOTAL MONEY EARNED	TRAINING ROWS
Bradley, Keegan	25	21	4	0.16	3,636,812.59	68
Spieth, Jordan	23	18	5	0.2174	3,879,819.57	33
Appleby, Stuart	25	19	6	0.24	538,332.64	84
McDowell, Graeme	16	11	5	0.3125	2,174,595.40	38
Harrington, Padraig	17	10	7	0.4118	711,244.22	54
Immelman, Trevor	24	12	12	0.5	360,548.67	85
Bowditch, Steven	22	11	11	0.5	697,774.65	80
Beljan, Charlie	23	7	16	0.6957	916,228.66	57
Verplank, Scott	14	4	10	0.7143	62,905.00	44
Allenby, Robert	23	5	18	0.7826	204,272.00	84

Table 5.7 Golfers for which individual models are to be built (2013 Season Data)

Four golfers are from the top echelon of those who make the cut most often. Three golfers missed or made the cut roughly on a 50/50 spilt. The final three were chosen due to their ability to miss the cut in the majority of the tournaments played. Jordan Spieth has the least number of rows available for training models, while Stuart Appleby and Robert Allenby both have 84 rows available for training.

Iterative models will be created for each of these golfers. Training data will be created using the following logic:

- From 2011 to end of 2013 season, create aggregate data for X number of tournaments together with the outcome (Missed Cut, Y/N) of the “focus” tournament.
- Train models and predict the first tournament of 2014.
- Add metrics of the first tournament in 2014 to the training data and rebuild the model to create a new model to predict the next tournament, and so on.

The prediction for each tournament played by each golfer in 2014 will be scored and recorded. Once all iterations of the model have been complete for each golfer, the predictions will be compared to the actual results of these tournaments and accuracy will be calculated.

The attributes used for this experiment are the same as detailed in Table 5.5, with the addition of a “Weight” variable. The intention of this variable is to assign more importance to newer data than to older data if, after initial trials of the experiment it is deemed necessary to introduce it.

SAS Enterprise Guide will be used to develop the Logistic Regression models for this experiment. This is mainly due to the slightly easier implementation of building and implementing an increasing number of models. This, coupled with the fine tuning of filters and the extra detail provided by the models provides much more control and information than is possible with SAS Enterprise Miner for this task.

Stepwise selection will be used to pick the explanatory variables most significant to each model. It is hypothesised that the explanatory variables will change from golfer to golfer, and also between iterations of models for an individual golfer.

5.5 Statistical Experiments

5.5.1 Experiment 5 - Measuring the influence of caddies

As discussed in Chapter Three, caddies are perceived to have an influence on a professional golfer’s performance. This experiment will test this hypothesis by measuring scoring averages and average money earned using paired t-tests to compare performance before and after golfers changed their caddies.

The golfer’s whom are the subject of this experiment are Lee Westwood and Adam Scott.

Lee Westwood was chosen due to the fact that his caddie from 2009, Billy Foster was injured in May 2012[18]. After an 18 month gap which saw Foster replaced as caddie, he returned to partner Westwood in December 2013[19].

Adam Scott was chosen due to his high profile hiring of Tiger Woods former caddie, Steve Williams in 2011[20]. Beginning with the US Open in the same year, Williams has caddied for Scott since then.

Between 2009 and up until Billy Foster was injured in 2012, Lee Westwood played in 30 stroke play tournaments on the PGA Tour. He only played in 27 PGA Tour tournaments in the 18 month period when Billy Foster was not his caddy. In order to achieve balanced sample sizes and for the sake of simplicity 3 tournaments from the 2009-2012 seasons where randomly excluded from the data.

For Adam Scott the data selection process is more straightforward. The 30 most recent stroke play tournaments before the 2011 US Open were chosen as the “before” sample. For the “after” sample, the 30 stroke play tournament from the 2011 US Open onwards, were selected.

As both Lee Westwood and Adam Scott have competed in tournaments organised by other tours such as the European or Asian Tours, the PGA data is considered to be a sample from the overall “population” of tournaments played by each player in their respective timelines.

Of the 41 tournaments Adam Scott played in, 9 were non PGA Tour events which counted towards world ranking points between 2012 and 2013. Lee Westwood competed in 18 non PGA Tour events out of 52 in total for the same time period.

[18] http://www.pga.com/news/pga-tour/lee-westwood-keep-mike-kerr-permanent-caddie-billy-foster-rehabs-injured-knee

[19] http://www.dailymail.co.uk/sport/golf/article-2513452/Inevitable-reunion-Lee-Westwood-Billy-Foster--Derek-Lawrensons-World-Golf.html

[20] http://sports.espn.go.com/golf/usopen11/news/story?id=6651746

5.5.2 Experiment 6 - Jordan Spieth Vs Rory McIlroy

Since the beginning of the 2013 season, 20 year old Jordan Spieth has been compared to 24 year old Rory McIlroy in terms of talent and performance. There are numerous articles on the internet comparing Spieth to McIlroy, as well as other successful golfers[21]. This experiment will seek to determine if these comparisons are justified.

In this experiment, the scoring averages of both players and separately, their average money earned will be compared in order to determine if there is a statistically significant difference or similarity in their performances.

As there is an age difference of 4 years between the two players, a direct comparison based on commonly played tournaments is deemed to be unfair. In order to have a balanced experiment, data will be collected based on the first 30 PGA Tour stroke play tournaments of each player.

Both players were 19 years of age competing in their first full season so this should give unbiased samples based on increasing levels of experience as they are exposed to more pressure following on from impressive tournament results.

In 2013, his first full season, Jordan Spieth played the vast majority of his tournaments on the PGA Tour, 23 out of 27. This is in contrast to Rory McIlroy who played 11 PGA Tour events out of a total 28 tournaments on all tours in his first PGA Tour season, 2009.

5.5.3 Experiment 7 - Age Profile of the Cut Line

Chapter Three introduced the perception that the age of a professional golfer can have a bearing on their performance. Analysis will be carried out in order to determine if there is any significance between the age of the golfer and their likelihood to miss or make the cut. 9 age groups were created for the analysis (available in Figure 5.3).

The treemap of Figure 5.3 shows the proportions of 8 of the 9 groups compared to the total number of golfers who played on the PGA Tour in 2013. The 0-18 age group is

[21] http://www.golfdigest.com/blogs/the-loop/2014/01/at-this-point-in-his-career-jordan-spieth-is-outperforming-a.html

missing from the treemap due to the very small proportion of the golfer population it makes up (0.13%).

Figure 5.3 can be interpreted as follows:

The outer rectangle represents 100% of the population of golfers. The sub-rectangles are proportionally sized based on the population in each age group. The age group category (e.g. 30-35) is displayed first, followed by the percentage of the population which the age group consists of.

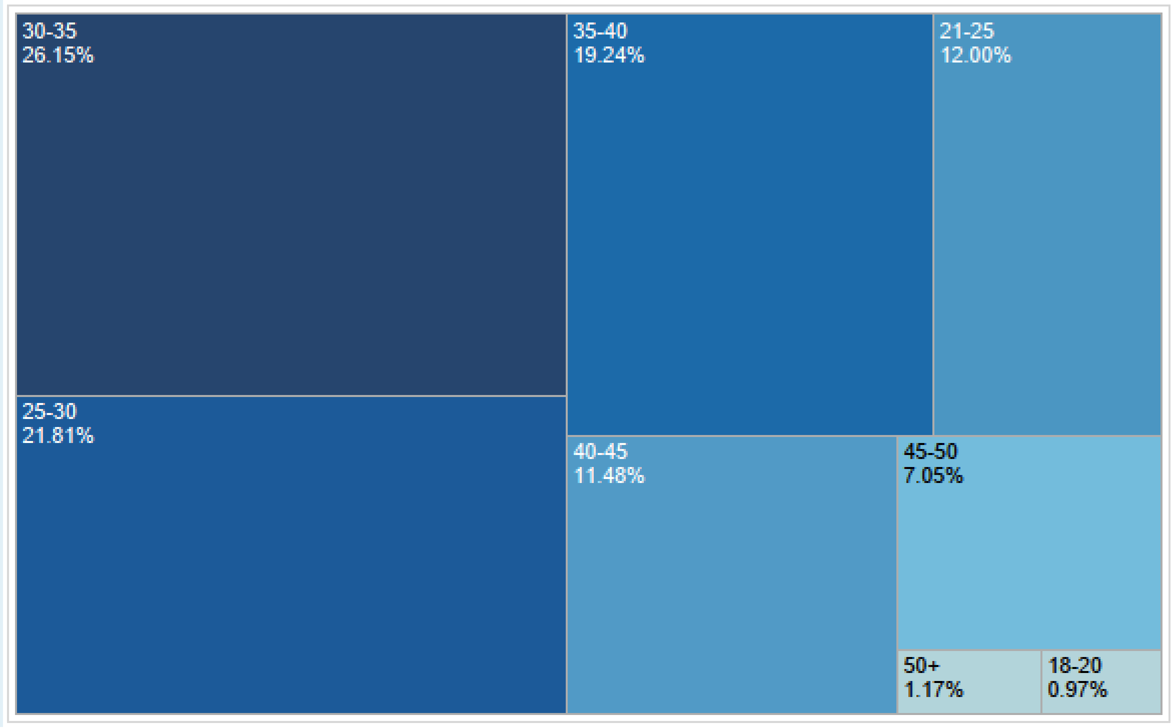


Figure 5.3: Treemap of the age profile of golfers who played on the PGA Tour in 2013. With Age Group and the group percentage of the total population. (Shotlink, 2014)

Figure 5.3 shows that the largest age group is the 30-35s who made up 26.15% of the total golfer population in 2013. Additional analysis shows that this group missed the cut 23.96% of the time in 2013 which is also the highest missed cut proportion. This would be expected considering 30-35s made up the highest numbers.

Breaking down the numbers further, Figure 5.4 shows the proportion of the number of cuts made, to cuts missed, for each age group, for the total number of tournaments played by each group.

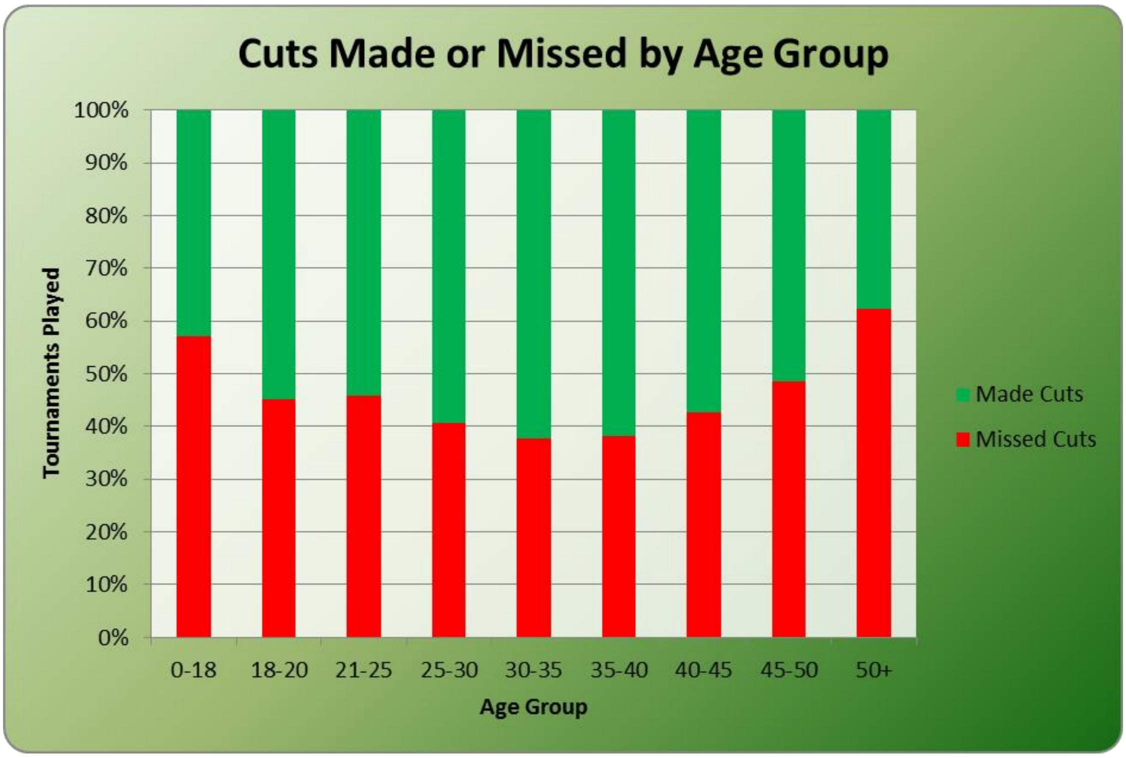


Figure 5.4: Cuts Made or Missed by Golfer Age Group

Figure 5.4 helps to show that despite having the highest number of missed cuts, the 30-35 age group has the highest proportion of cuts made of all the groups. As this cohort have played the most number of tournaments this also seems fair.

Based on this analysis, a two way Chi Squared test will be carried out in order to determine if there is statistical evidence to prove an association between the age of a professional golfer and making or missing the cut.

The input data for this experiment will be from the 2013 season only. It will include the derived age group for each golfer in a given tournament, their outcome, (whether they missed the cut or not) and the count of the number of occurrences for each scenario. In total there are 5,235 data points split out by 2,153 missed cuts and 3,082 made cuts.

5.6 Conclusion

This chapter has outlined the concepts, designs and methodologies to be used for the seven experiments to be implemented for this project. It has introduced the dual methodologies of using collective training data and individual golfer training data, outlining the nuances of each approach. The four experiments to be performed for predictive modelling have been detailed, with the differences between each of them discussed, as well as their expected outcomes.

The design of the statistical experiments has also been outlined with details on why the specific golfers were suitable subjects for the experiments. Exploratory data was also presented to support the rational for Experiment 7 - “Age profile of the Cut Line”.

Chapter Six will outline the results of these experiments in detail and discuss observations and highlights within these results.

6. EXPERIMENT RESULTS

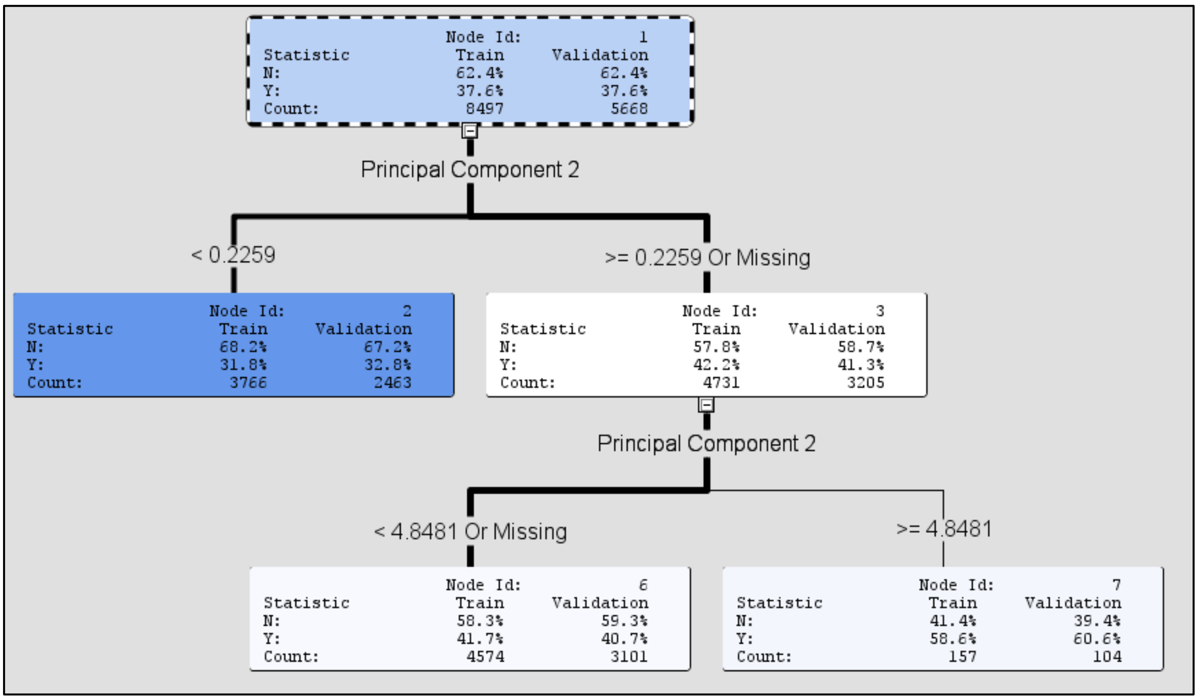
This chapter will detail the results of all seven experiments outlined in Chapter Five. It will present the results for each predictive model, for each experiment, under the headings of overall accuracy, make cut accuracy and miss cut accuracy. This will allow for simple comparison of models both within and between experiments. Specific details of the golfer segmentation process and its outcomes will also be presented.

The results of the statistical experiments will also be outlined with additional commentary on the interpretation of their validity as statistical models.

6.1 Predictive Models

6.1.1 Experiment 1 – Granular Data

The first experiment conducted used 170 numeric attributes from the 190 available in the Event Level dataset. Principal Components Analysis (PCA) was first used to reduce the number of features down to 20 (the default setting) Principal Components. The output from the PCA node was then used as input firstly for a Decision Tree model to train and validate the data using a 60/40 split.



Statistic	Node Id: 1 Train	Validation
N:	62.4%	62.4%
Y:	37.6%	37.6%
Count:	8497	5668

Statistic	Node Id: 2 Train	Validation
N:	68.2%	67.2%
Y:	31.8%	32.8%
Count:	3766	2463

Statistic	Node Id: 3 Train	Validation
N:	57.8%	58.7%
Y:	42.2%	41.3%
Count:	4731	3205

Statistic	Node Id: 6 Train	Validation
N:	58.3%	59.3%
Y:	41.7%	40.7%
Count:	4574	3101

Statistic	Node Id: 7 Train	Validation
N:	41.4%	39.4%
Y:	58.6%	60.6%
Count:	157	104

Figure 6.1: Decision Tree from PCA Inputs

As can be seen in Figure 6.1 the only variable used in the tree was "Principal Component 2".

Table 6.1 holds the results of experiment 1, measured under the categories of overall accuracy, make cut accuracy and miss cut accuracy. Overall accuracy tallies the actual outcome (missed or made the cut) against the predicted outcome for all 146 golfers who entered the Northern Trust Open. "Make Cut" tallies the golfers who's predicted outcome was "N" (Make the Cut) with the actual outcome. Similarly, "Missed Cut" accuracy takes the predicted outcome "Y" (Miss the Cut) and tests it against the actual outcome.

Model	Overall Accuracy %	Make Cut Accuracy %	Miss Cut Accuracy %
Decision Tree	▨	▨	▨
K-NN	53.84	59.74	46.96
Neural Network	58.04	63.63	51.51
Logistic Regression	51.74	55.69	44.77

Table 6.1: Experiment 1: Model Accuracy incorporating Principal Components

It can be seen that no results were obtained for the decision tree model and this can be explained as follows:

The output from this model produced only two distinct probabilities for a golfer, either 0.3175 or 0.4167 for missing the cut. 91 golfers were assigned a probability of 0.3175. As only 77 golfers made the cut for the Northern Trust Open, this poses the problem of selecting a cut-off point. As no other means of separating the golfers for scoring purposes was logical the results are deemed invalid for the Decision Tree model for this experiment.

Of the remaining 3 models, the Neural Network had the greatest overall accuracy at 58%. Its ability to predict who would make the cut at 63.6% was significantly better than its ability to determine who would miss the cut (51.5%). It should be noted that as the cut line of the Northern Trust Open was not an even split, 77 golfers made the cut and 69 missed it. This imbalance could have the potential to cause differences in the accuracy between the three aforementioned results categories.

6.1.2 Experiment 2 – Aggregation

Experiment 2 was implemented using two different training datasets. The first was based on aggregate data from the previous five tournaments prior to the tournament being predicted. The second was based on aggregate data from just the previous two tournaments. Table 6.2 outlines the experiment results under the same headings as outlined for Experiment 1.

	Overall Accuracy %		Make Cut Accuracy %		Miss Cut Accuracy %	
Model	5 Events	2 Events	5 Events	2 Events	5 Events	2 Events
Decision Tree	58.57	63.82	64.86	68.83	52.30	57.81
K-NN (12 Ns)	66.18	63.82	71.62	68.83	60.00	57.81
Neural Network	58.27	52.48	64.38	58.44	51.51	45.31
Logistic Regression	59.71	58.15	65.75	63.63	53.03	51.56

Table 6.2: Experiment 2: Model Accuracy for 5 and 2 tournament aggregates[22]

Over all of the categories, with the exception of "5 Events, Decision Tree", there was greater predictive accuracy when using the aggregate data from the five most recent tournaments prior to the Northern Trust Open, rather than using just the two previous tournaments. This is contrary to discussion with (Cosgrove, 2014) where it was thought unlikely that five tournaments of data would give a reliable indication of current form.

It should also be noted that in contrast to Experiment 1, the Decision Tree model produced sufficiently different probabilities so that it was possible to split golfers appropriately.

[22] Decision Trees & k-NN for 2 tournament aggregation data both had same numbers predicted however; there were difference in the predictions for golfers between models. The reader is referred to Appendix C for further details.

6.1.3 Experiment 3 – Clustering

The 558 golfers who played in PGA sanctioned tournaments during the 2013 season where segmented into one of 5 different clusters. The importance of each of the input variables into the segmentation process is shown in Table 6.3.

Column	Importance
NO_OF_PUTTS	1
NO_OF_CUTS_MADE	0.988417049
NO_OF_ONE_PUTTS	0.93388861
MOV_AVG_FINISH_POS	0.86814397
NO_OF_CUTS_MISSED	0.785207178
NO_OF_THREE_PUTTS	0.776004579
CUTS_MISSED_RATIO	0.637799701
AVG_MONEY_EARNED	0.631056496
JUST_MADE_CUT_CNT	0.603591537
AVG_GIR_PER_RND	0.342015728
AVG_NO_OF_PUTTS	0.257764829
MA_DRIVE_ACC	0.210555236
JUST_MISSED_CUT_CNT	0

Table 6.3: Importance of each variable for creation of clusters

Figure 6.2 illustrates a 3D representation of the defined clusters based on the three most important variables, the number of putts, cuts made and one putts. With the use of colour coded data points representing each of the five segments, the separation of the individual clusters is clearly observed.

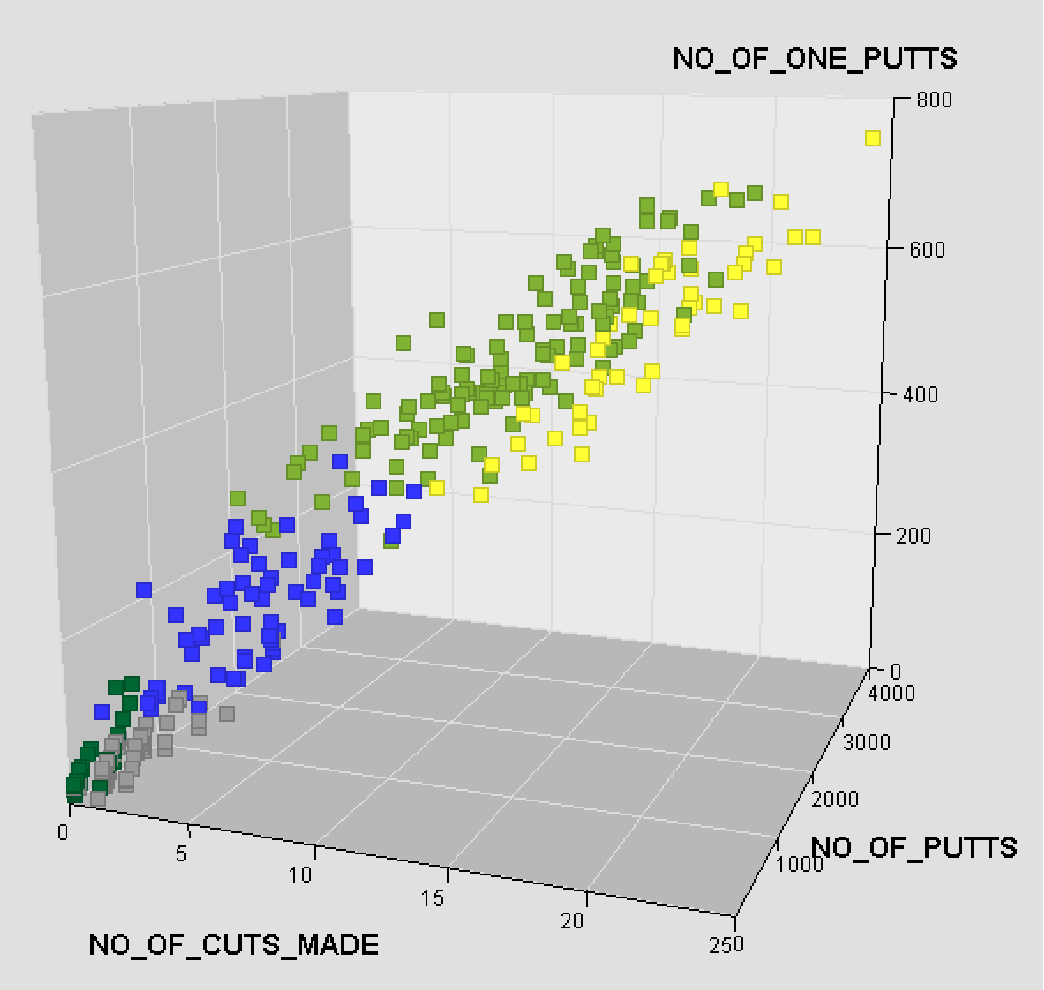


Figure 6.2: 3D Scatter Plot of golfer segments (Segments divided by colour)

Table 6.4 contains the number of golfers in each cluster along with the average values of the primary attributes. The differences between each cluster are also quite discernable in this tabular format.

Segment ID	Number of Golfers in Cluster	Avg Money Earned	Avg No of One Putts	Avg No of Putts	Avg No of cuts made
5	183	2,888.49	18.89	89.02	0.10
2	136	25,102.33	35.53	165.71	1.15
4	123	64,127.09	484.46	2024.77	12.59
1	67	44,855.67	205.75	875.03	5.06
3	49	159,966.60	528.18	2198.10	17.78

Table 6.4: Population of clusters, with key metrics

Following the creation of the clusters, training dataset were built for each of them, details of which can be found in Table 6.5.

Segment ID	Training Data Size	No of golfers in Northern Trust Open
1	2484	24
2	1158	8
3	2700	30
4	7127	72
5	788	14

Table 6.5: Training data size per cluster

Training data was built using aggregated data based on the previous 2 tournaments only. The scores from each model, for each cluster where combined and then sorted by descending probability score for assignment of the classification. Table 6.6 outlines the results from each combined scoring model.

Model	Overall Accuracy %	Make Cut Accuracy %	Miss Cut Accuracy %
Decision Tree	50.67	57.14	43.62
K-NN	45.94	40.25	52.11
Neural Network	63.51	68.83	57.74
Logistic Regression	65.54	71.42	59.15

Table 6.6: Experiment 3 - Overall Combined Results

Analysis of the results show that K-NN suffers badly with this approach compared to the collective aggregation utilised in Experiment 2. Logistic Regression would appear to favour this segmented approach.

6.1.4 Experiment 4 – Individual Performance

Due to the time consuming and iterative nature of this experiment, logistic regression was the sole model used for this experiment. Within the group of chosen golfers, the minimum number of tournaments played was 3 and the maximum was 14 tournaments played in the 2014 season, up to the Shell Houston Open.

With a base point of training data up to the end of the 2013 season to predict the first of tournament of 2014, the model was re-trained with each new tournament's data to predict the next. This required 102 re-runs of the logistic regression model coupled with experiment setup.

During the experiment where Trevor Immelman was the focus golfer, it was noted that the output prediction for all the 2014 tournaments played by him had a prediction of "N". So according to the model Trevor Immelman would not miss any cuts for the tournaments played. When the model was run excluding 2011 season data, there was more variation in the predictions. As it is counter intuitive to use less data to train a model, a weighted variable was introduced into the training dataset. This allows older data to have less importance than newer data which makes sense when considering a golfer's current form. By assigning newer data a higher weight it ensures that current form has more influence on the models predictions.

Table 6.7 shows the high level results from the logistic regression experiments. The logistic regression model for Jordan Spieth found that no variables were significant therefore only an intercept model was produced. The result of this is a model that would be no better than the mean and therefore is considered an unsuccessful result for Jordan Spieth.

Player Name	Tournaments Played in 2014	Cuts Made	Cuts Missed	Tournaments Correctly Predicted	Prediction Accuracy %
Stuart Appleby	14	11	3	10	71.42
Graeme McDowell	3	3	0	2	66.66
Keegan Bradley	7	6	1	4	57.14
Jordan Spieth	7	5	2	0	0
Trevor Immelman	13	7	6	6	46.15
Stephen Bowditch	14	9	5	8	57.14
Padraig Harrington	8	3	5	6	75
Robert Allenby	14	5	9	10	74.42
Scott Verplank	8	2	6	5	62.5
Charlie Beljan	14	6	8	8	57.14

Table 6.7: Final results of individual prediction models

3 tournaments versus 2 tournaments

Initial trials of the experiment were conducted to examine the most effective number of tournaments upon which to aggregate data for the training datasets. Models were built for Robert Allenby based on aggregate data from the previous 3 and the previous 2 tournaments.

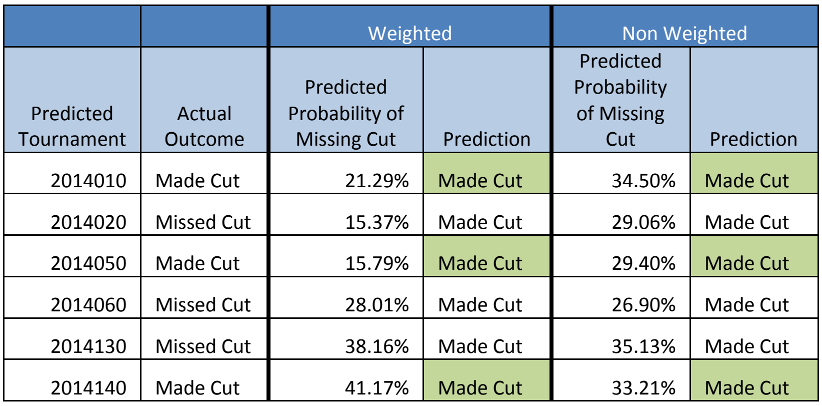
The models based on 2 tournaments of training data predicted 10 out 14 outcomes correctly. The 3 tournament models predicted 7 out 14 correctly. Based on this result all model for Experiment 4 were based on aggregate data containing details from the previous two tournaments prior to the tournament being predicted.

This finding is in line with the thoughts of (Cosgrove, 2014) but contrary to the results of experiment 2 which showed that data from the 5 previous tournaments produced better accuracy, at the collective level at least.

With and without weights

The introduction of a weight variable into the Logistic Regression models has produced varying results. For Trevor Immelman's models, use of weights changed the predicted outcome from 13 "No" values down to 10, from a total of 13 events played. While this may seem a small reduction it shows that the newer data is affecting the model equation more significantly than without any weighting given to the newer data.

Table 6.8 provides a comparison of the results of models produced for Trevor Immelman. For weighted models, the prediction accuracy for Trevor Immelman is 6 out of 13 (46.15%). For non-weighted models, the accuracy is 7 out of 13 (53.8%).



		Weighted		Non Weighted	
Predicted Tournament	Actual Outcome	Predicted Probability of Missing Cut	Prediction	Predicted Probability of Missing Cut	Prediction
2014010	Made Cut	21.29%	Made Cut	34.50%	Made Cut
2014020	Missed Cut	15.37%	Made Cut	29.06%	Made Cut
2014050	Made Cut	15.79%	Made Cut	29.40%	Made Cut
2014060	Missed Cut	28.01%	Made Cut	26.90%	Made Cut
2014130	Missed Cut	38.16%	Made Cut	35.13%	Made Cut
2014140	Made Cut	41.17%	Made Cut	33.21%	Made Cut

2014160	Missed Cut	32.83%	Made Cut	31.75%	Made Cut
2014170	Missed Cut	87.55%	Missed Cut	36.59%	Made Cut
2014190	Made Cut	97.49%	Missed Cut	40.05%	Made Cut
2014200	Made Cut	52.20%	Missed Cut	32.63%	Made Cut
2014230	Made Cut	40.68%	Made Cut	33.78%	Made Cut
2014240	Made Cut	27.74%	Made Cut	33.49%	Made Cut
2014250	Missed Cut	36.51%	Made Cut	34.86%	Made Cut

Table 6.8: Comparison of Weighted and Non Weighted predictions for Trevor Immelman. Cells shaded in green indicate correct prediction.

Table 6.9 compares results for Padraig Harrington, showing that the introduction of weighting had the opposite result when compared to Trevor Immelman's models. The weighted models correctly predicted the outcome of Padraig Harrington's tournaments 6 out of 8 times (75%). The non-weighted models predicted the outcome correctly 4 out of 8 times (50%).

		Weighted		Non Weighted	
Predicted Tournament	Actual Outcome	Predicted Probability of Missing Cut	Prediction	Predicted Probability of Missing Cut	Prediction
2014150	Missed Cut	52.09%	Missed Cut	12.43%	Made Cut
2014160	Made Cut	37.63%	Made Cut	51.45%	Missed Cut
2014170	Missed Cut	59.81%	Missed Cut	80.73%	Missed Cut
2014190	Missed Cut	97.99%	Missed Cut	76.21%	Missed Cut
2014220	Made Cut	94.18%	Missed Cut	35.37%	Made Cut
2014230	Made Cut	1.66%	Made Cut	17.24%	Made Cut
2014240	Missed Cut	1.16%	Made Cut	12.30%	Made Cut
2014250	Missed Cut	78.65%	Missed Cut	36.08%	Made Cut

Table 6.9: Comparison of Weighted and Non Weighted predictions for Padraig Harrington. Cells shaded in green indicate correct prediction.

Analysis of the models produced for Trevor Immelman would suggest that non-weighted models are more accurate; however, this is probably down to chance. As mentioned, the non-weighted model predicted a "No" value for all of Trevor Immelman's events in 2014. As such, it has a 50/50 chance of being correct. Though the non-weighted model is less accurate in this case, it has more credence, showing its ability to adapt and give impetus to change according to the newer data.

It appears to be logical to weight newer data to ensure its influence on any models built. A golfer’s form can be fairly steady over a number of years, or it can be erratic. It is also possible that a golfer could experience a sudden loss of form, or develop the Yips and in this way, a weighting will bring flexibility to the model so that it can quickly change its predictive behaviour.

Evidence of change as newer data introduced to models.

The flexibility of re-building models with the additional data from the previous tournament is evident in a number of cases. For the example of Robert Allenby, he made the cut in the Valspar Championship after missing 5 cuts out of 6, including 2 consecutive missed cuts just prior to the Valspar.

The models had predicted that Allenby would miss all 6 cuts prior to the Valspar Championship but successfully predicted that he would make the cut for the Valspar. The explanatory variables used in the Robert Allenby models were “Average Greens Hit in Regulation” and “Average Driving Accuracy”.

Analysis of these for the tournaments prior to the Valspar Championship show slight increases in their numbers which when added to the existing training data was enough to cause the switch in the predicted classification. Allenby’s Average Greens Hit in Regulation had increased from 10 to 11.16 over previous four tournaments. His average driving accuracy recovered from a decline of 15% from his forth last tournament to only an 8% decline in his last tournament prior to the Valspar Championship.

Similar results are evident for Stephen Bowditch for whom the explanatory variables in his model where “Average Others”[23], “Missed Cuts Ratio” and “Number of 3 Putts”.

If automated, individual models could prove to be very accurate predictors of a golfer’s ability to make the cut or not. It may also be possible to predict other factors of a golfer’s game, such as whether a golfer will finish in the top 5, 10, or 20 places.

[23] Scores such as double bogies or an albatross

There may also be ways to quantify and factor in external influences that affect the physiological aspect of a golfer's game such as relationship breakups or fatigue.

Logistic Regression Interpretation

The reader is referred to Section 4.4.3.2 for details of the tests used to interpret the results of a logistic regression model.

Tables 6.10, 6.11 and 6.12 contain the details of the Goodness of Fit tests, Maximum Likelihood Estimates analysis and "Global Null Hypothesis: BETA=0" testing. For the sake of brevity only two golfers and two tournament models are shown from the 102 models created. During the implementation of each model, the results of each of the three aforementioned tests were observed to ensure the models were a good fit. Only the models of Jordan Spieth were not a good fit.

Hosmer and Lemeshow Goodness-of-Fit Test				
Player	Tournament ID	Chi-Square	DF	Pr > ChiSq
Stephen Bowditch	2014010	8.647	6	0.1944
Stephen Bowditch	2014170	10.6038	7	0.1569
Scott Verplank	2014010	10.5601	7	0.159
Scott Verplank	2014130	12.9244	8	0.1145

Table 6.10: Goodness of Fit Test for Logistic Regression Models

Table 6.10 shows that all P values are greater than 0.05 which accepts the null hypothesis that the model is a good fit

Analysis of Maximum Likelihood Estimates							
Player	Tournament ID	Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Stephen Bowditch	2014010	Intercept	1	-0.1289	0.4669	0.0762	0.7825
		NO_OF_CUTS_MADE	1	1.5404	0.405	14.4643	0.0001
		AVG_OTHERS	1	7.4755	2.0577	13.1989	0.0003
		NO_OF_THREE_PUTTS	1	-0.4811	0.1392	11.946	0.0005
	2014170	Intercept	1	2.5686	0.6593	15.1798	<.0001
		CUTS_MISSED_RATIO	1	-2.7129	0.6919	15.3715	<.0001
		AVG_OTHERS	1	7.0952	1.821	15.1819	<.0001
		NO_OF_THREE_PUTTS	1	-0.5139	0.1186	18.7645	<.0001

Scott Verplank	2014010	Intercept	1	12.5872	4.4629	7.9547	0.0048
		MA_DRIVE_ACC	1	-22.1086	6.5902	11.2546	0.0008
		MOV_AVG_FINISH_POS	1	0.1501	0.0696	4.6526	0.031
		NO_OF_CUTS_MISSED	1	-3.1474	0.9635	10.67	0.0011
		AVG_EAGLES	1	9.3255	2.9371	10.0812	0.0015
	2014130	Intercept	1	14.3784	3.6016	15.9382	<.0001
		MA_DRIVE_ACC	1	-21.1498	5.3298	15.7468	<.0001
		NO_OF_CUTS_MISSED	1	-0.9069	0.4387	4.2727	0.0387
		AVG_EAGLES	1	6.0754	2.1774	7.7854	0.0053

Table 6.11: Maximum Likelihood Estimates for Logistic Regression Models

Table 6.11 shows that with the exception of the intercept for Stephen Bowditch in Tournament ID 2014010, all the P-values are less than 0.05 meaning that they are significant to the model.

Testing Global Null Hypothesis: BETA=0					
Player	Tournament ID	Test	Chi-Square	DF	Pr > ChiSq
Stephen Bowditch	2014010	Likelihood Ratio	31.7283	3	<.0001
		Score	27.6202	3	<.0001
		Wald	21.235	3	<.0001
	2014170	Likelihood Ratio	37.7222	3	<.0001
		Score	33.4518	3	<.0001
		Wald	26.5204	3	<.0001
Scott Verplank	2014010	Likelihood Ratio	34.7802	4	<.0001
		Score	27.0782	4	<.0001
		Wald	15.8305	4	0.0033
	2014130	Likelihood Ratio	30.4687	3	<.0001
		Score	26.0152	3	<.0001
		Wald	17.6804	3	0.0005

Table 6.12: Global Null Hypothesis results for Logistic Regression

The P-values in Table 6.12 are all less than 0.05 indicating that at least one of the regression coefficients is not equal to zero. This signifies that there is a dependency between missing the cut and at least one of the explanatory variables.

Explanatory factors changing as new data added

Table 6.13 contains examples of the statistically significant explanatory variables in three different models for both Trevor Immelman and Padraig Harrington. It can be seen that not only are there differences in which variables are important in explaining each golfers performance, but also between tournaments for the same golfer.

Focusing on Trevor Immelman, for one tournament, the average number of double bogies was an additionally significant variable but became insignificant subsequently. Birdie and Eagle scores are clearly the main indicators of Immelman’s form.

Padraig Harrington on the other hand had a minimum of five different explanatory factors which were significant in explaining his performances. His number of three putts, one putts and greens in regulation all make appearances in some of the models, but not all of them. Harrington’s performance indicators appears to be more complex than Trevor Immelman’s.

The fluid nature of the individually constructed models which are iteratively updated on a continuous basis as new data becomes available exemplifies the complex nature of predicting a golfer’s performance. While the individual scores are promising there would still be potential issues when combining 120 individual model results into a prediction for one specific tournament.

This change indicates the evolution of the models, given new training data and also shows how there are differences in the factors that explain individual golfer’s performances.

	Example 1	Example 2	Example 3
Trevor Immelman	AVG_BIRDIES	AVG_BIRDIES	AVG_BIRDIES
	AVG_EAGLES	AVG_EAGLES	AVG_EAGLES
		AVG_DOUBLES	
Padraig Harrington	AVG_DRV_ACC	NO_OF_CUTS_MADE	NO_OF_ONE_PUTTS
	NO_OF_CUTS_MADE	AVG_DRV_ACC	AVG_DRV_ACC
	AVG_BOGEYS	AVG_BOGEYS	NO_OF_CUTS_MADE
	NO_OF_THREE_PUTTS	JUST_MISSED_CUT_CNT	AVG_GIR_PER_RND
	JUST_MISSED_CUT_CNT	FORM_SCORE	JUST_MISSED_CUT_CNT
	FORM_SCORE		FORM_SCORE
			AVG_BOGEYS

Table 6.13: Explanatory variables used in different models (in descending order of significance)

6.2 Statistical Experiments

The reader is referred to Section 4.7.1 for the reasons why the subjects of the following experiments were chosen.

6.2.1 Experiment 5 - Measuring the influence of caddies

Adam Scott

A two sample t-test with a confidence level of 95%[24] was performed on tournaments entered by Adam Scott. With a total sample size of 60 tournaments these were split into 30 before Adam Scott teamed up with Steve Williams & 30 after the team up. The sample data consists of Money Earned and Scoring Average.

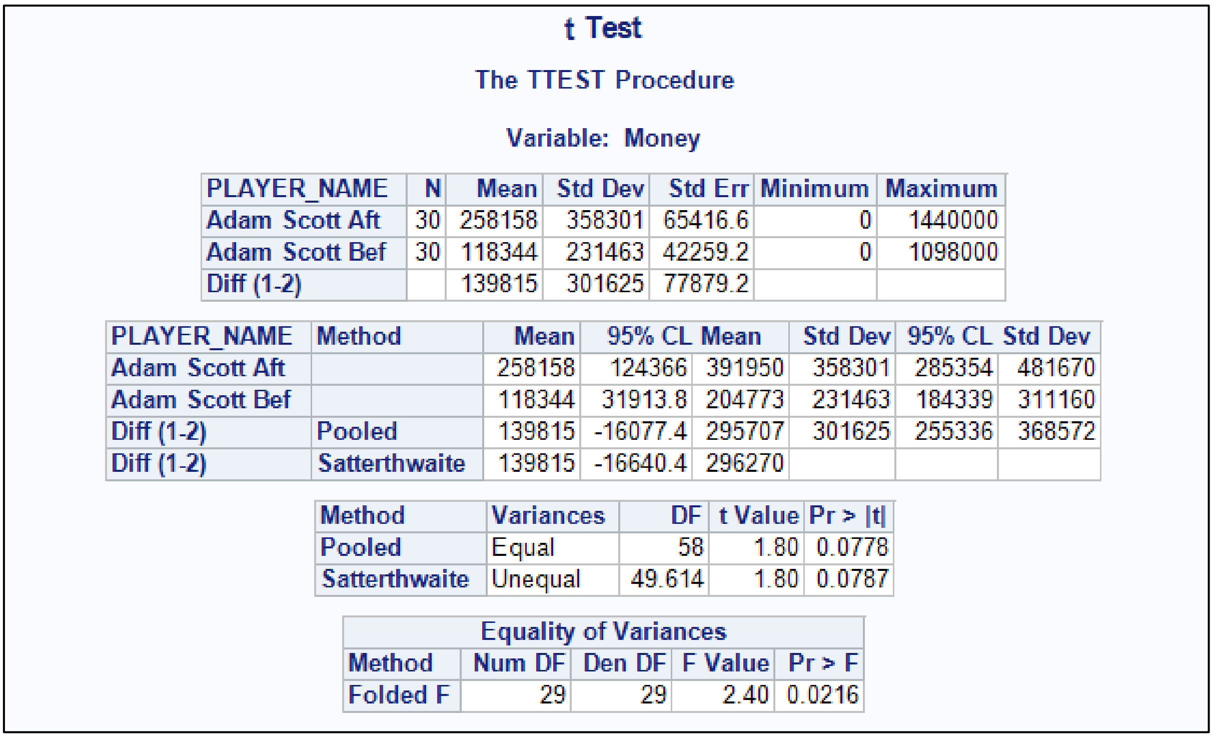
In the 30 tournaments before the team up, Adam Scott won a total of $3.5 million but won $7.7 million after. Figure 6.3 shows the P-Value for a two tail test (signified by |t|) of 0.0787. This means that the null hypothesis that there is no difference between the means cannot be rejected.

In order to obtain a one tail test result the P-Value is halved to 0.03935. As the t-statistic is positive with a value of 1.80 this suggests that the mean of sample 1 (Adam Scott After) is greater than the mean of sample 2 (Adam Scott Before). With a P-Value of less than 0.05 this result is statistically significant.

The result of the two tail test that there is no difference between the means may be surprising considering the fact that Adam Scott practically doubled his earnings since teaming up with Steve Williams (based on PGA Tour prize money). It has to be noted that a one tail test has more power to detect an affect as the rejection region is larger in a one tailed test than a two tailed test. The reader is referred to Chapter Four for details.

Figure 6.4 illustrates the distribution of Adam Scott's earnings between the two samples. The larger outliers are due mainly to his high finish positions in the more recent major tournaments.

[24] At 95% confidence, the P-value must be <0.05 in order to reject the null hypothesis



t Test

The TTEST Procedure

Variable: Money

PLAYER_NAME	N	Mean	Std Dev	Std Err	Minimum	Maximum
Adam Scott Aft	30	258158	358301	65416.6	0	1440000
Adam Scott Bef	30	118344	231463	42259.2	0	1098000
Diff (1-2)		139815	301625	77879.2		

PLAYER_NAME	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Adam Scott Aft		258158	124366	391950	358301	285354	481670
Adam Scott Bef		118344	31913.8	204773	231463	184339	311160
Diff (1-2)	Pooled	139815	-16077.4	295707	301625	255336	368572
Diff (1-2)	Satterthwaite	139815	-16640.4	296270			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	58	1.80	0.0778
Satterthwaite	Unequal	49.614	1.80	0.0787

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	29	2.40	0.0216

Figure 6.3: t-test results for Adam Scott's “Before” and “After” earnings

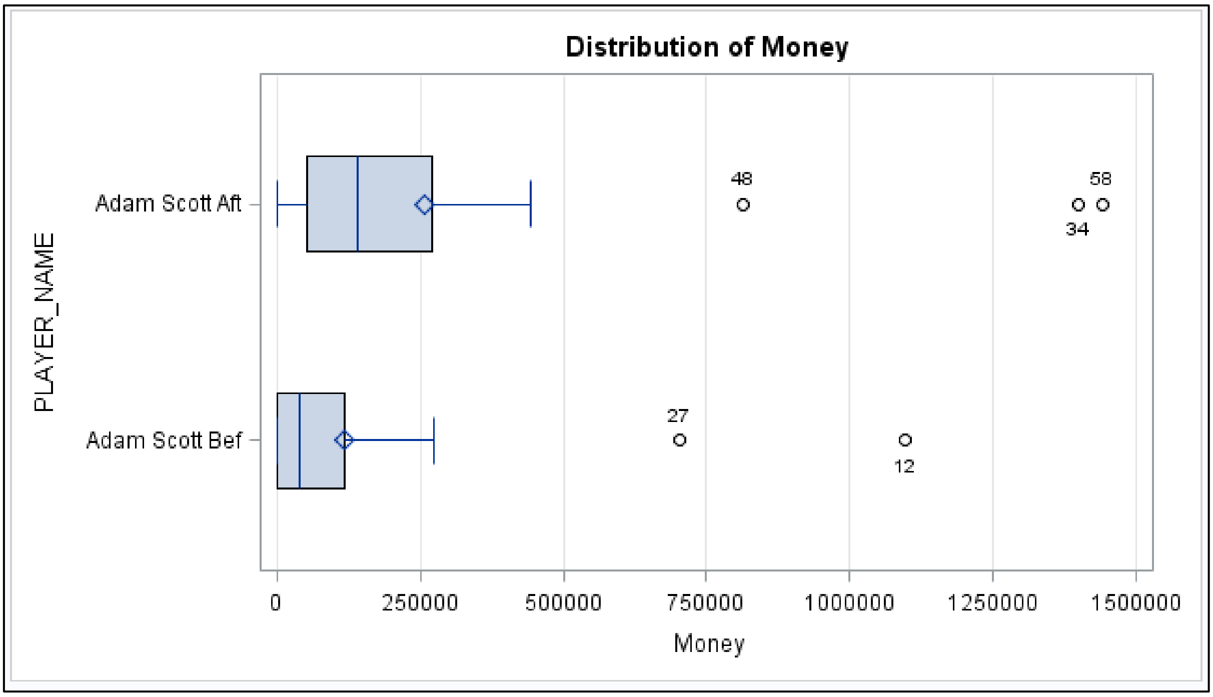
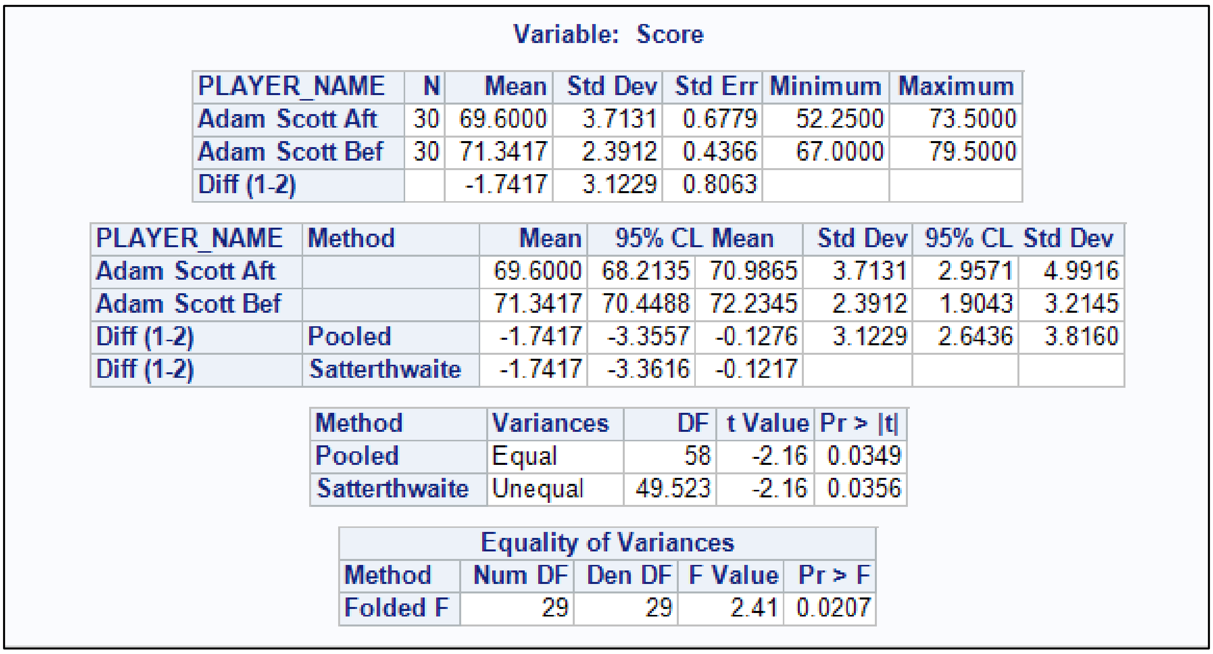


Figure 6.4: Distribution of Adam Scott's "Before" and "After" earnings

In contrast to Money earned, Adam Scott's scoring average with a P-Value of 0.0356 for a two tail test is statistically significant; there is evidence of a difference in the mean of the two distributions. The results of Figure 6.5 display the details.



Variable: Score

PLAYER_NAME	N	Mean	Std Dev	Std Err	Minimum	Maximum
Adam Scott Aft	30	69.6000	3.7131	0.6779	52.2500	73.5000
Adam Scott Bef	30	71.3417	2.3912	0.4366	67.0000	79.5000
Diff (1-2)		-1.7417	3.1229	0.8063		

PLAYER_NAME	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Adam Scott Aft		69.6000	68.2135	70.9865	3.7131	2.9571	4.9916
Adam Scott Bef		71.3417	70.4488	72.2345	2.3912	1.9043	3.2145
Diff (1-2)	Pooled	-1.7417	-3.3557	-0.1276	3.1229	2.6436	3.8160
Diff (1-2)	Satterthwaite	-1.7417	-3.3616	-0.1217			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	58	-2.16	0.0349
Satterthwaite	Unequal	49.523	-2.16	0.0356

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	29	2.41	0.0207

Figure 6.5: t-test results for Adam Scott's "Before" and "After" scoring average

The t-statistic value at -2.16 indicates that the actual mean of the population of the first sample (Adam Scott After) is less than the actual mean of the second sample (Adam Scott After). As golf is all about lower numbers being better, with the obvious exception of prize money, this is a logical result.

The results of both t-tests on Adam Scott's earnings and his scoring average suggest that hiring Steve Williams was a good decision. Of course, it could be just coincidence and Adam Scott's earnings may well have increased regardless of his caddy. However, considering Steve Williams extremely successful partnership with his former employer Tiger Woods, his influence cannot be discounted.

Lee Westwood

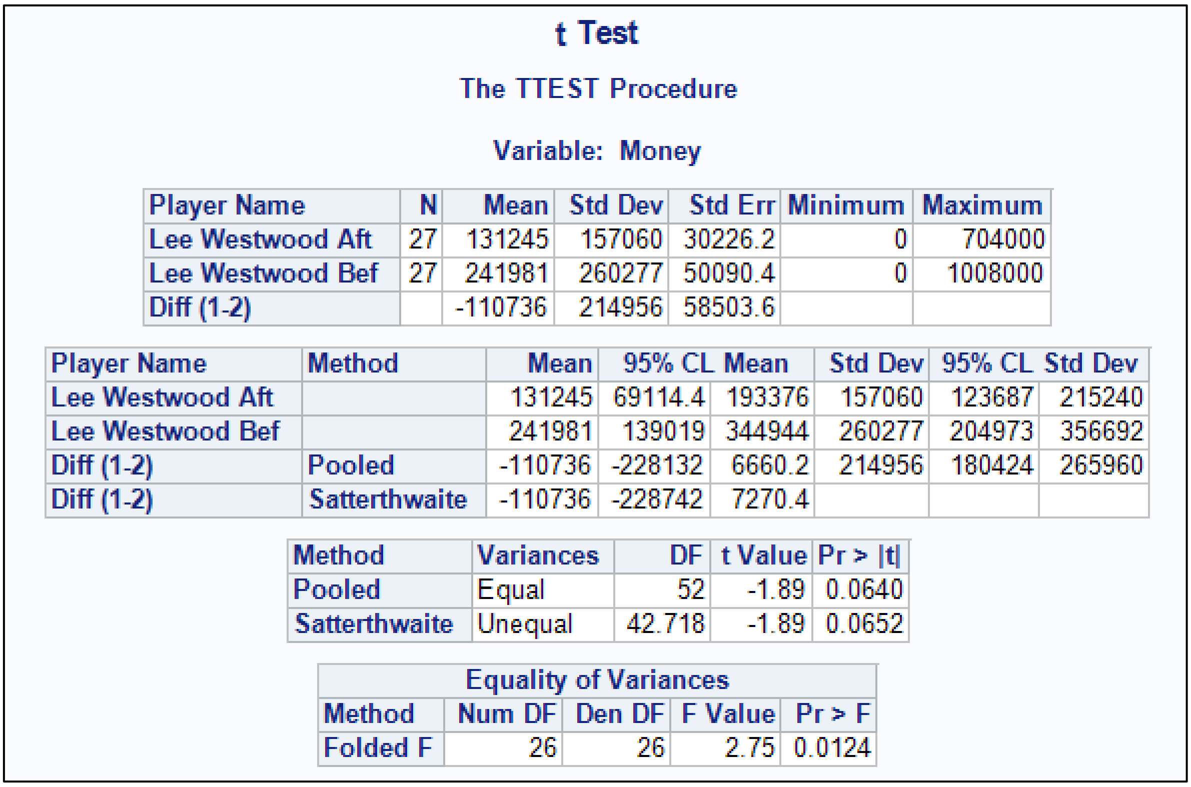
The exact same experiment was performed on Lee Westwood's data with the exception that the total sample size was 54 tournaments split evenly into two samples of tournaments. 27 when Billy Foster was his caddy (Lee Westwood Before) and 27 after Billy Foster was injured (Lee Westwood After). The sample data consists of Money Earned and Scoring Average.

When Billy Foster caddied for Lee Westwood in the period between 2009 and 2012, Lee Westwood earned approximately $7.5 million on the PGA Tour alone. After Billy Foster's injury, Lee Westwood earned $3.6 million.

Figure 6.6 shows the P-Value for a two tail test of 0.0652. This means that the null hypothesis that there is no difference between the means cannot be rejected.

In contrast to Adam Scott's results, the one tailed test provides a P-Value of 0.0326 and a negative t-statistic of -1.89. The inference of the test is that the population mean of the "Lee Westwood After" sample is less than the population mean of the "Lee Westwood Before" sample, with 95% confidence.

Figure 6.7 outlines the distribution of prize money earned on the PGA Tour by Lee Westwood for the two time periods in question.



t Test

The TTEST Procedure

Variable: Money

Player Name	N	Mean	Std Dev	Std Err	Minimum	Maximum
Lee Westwood Aft	27	131245	157060	30226.2	0	704000
Lee Westwood Bef	27	241981	260277	50090.4	0	1008000
Diff (1-2)		-110736	214956	58503.6		

Player Name	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Lee Westwood Aft		131245	69114.4	193376	157060	123687	215240
Lee Westwood Bef		241981	139019	344944	260277	204973	356692
Diff (1-2)	Pooled	-110736	-228132	6660.2	214956	180424	265960
Diff (1-2)	Satterthwaite	-110736	-228742	7270.4			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	52	-1.89	0.0640
Satterthwaite	Unequal	42.718	-1.89	0.0652

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	26	26	2.75	0.0124

Figure 6.6: t-test results for Lee Westwood's "Before" and "After" earnings

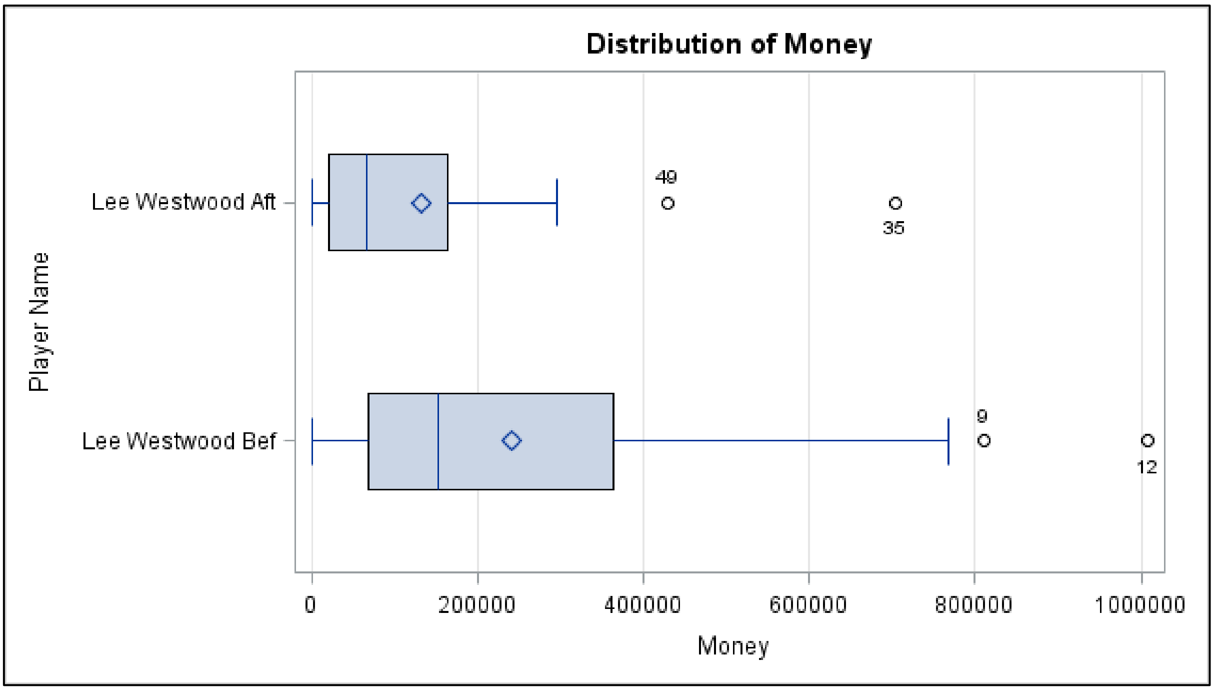
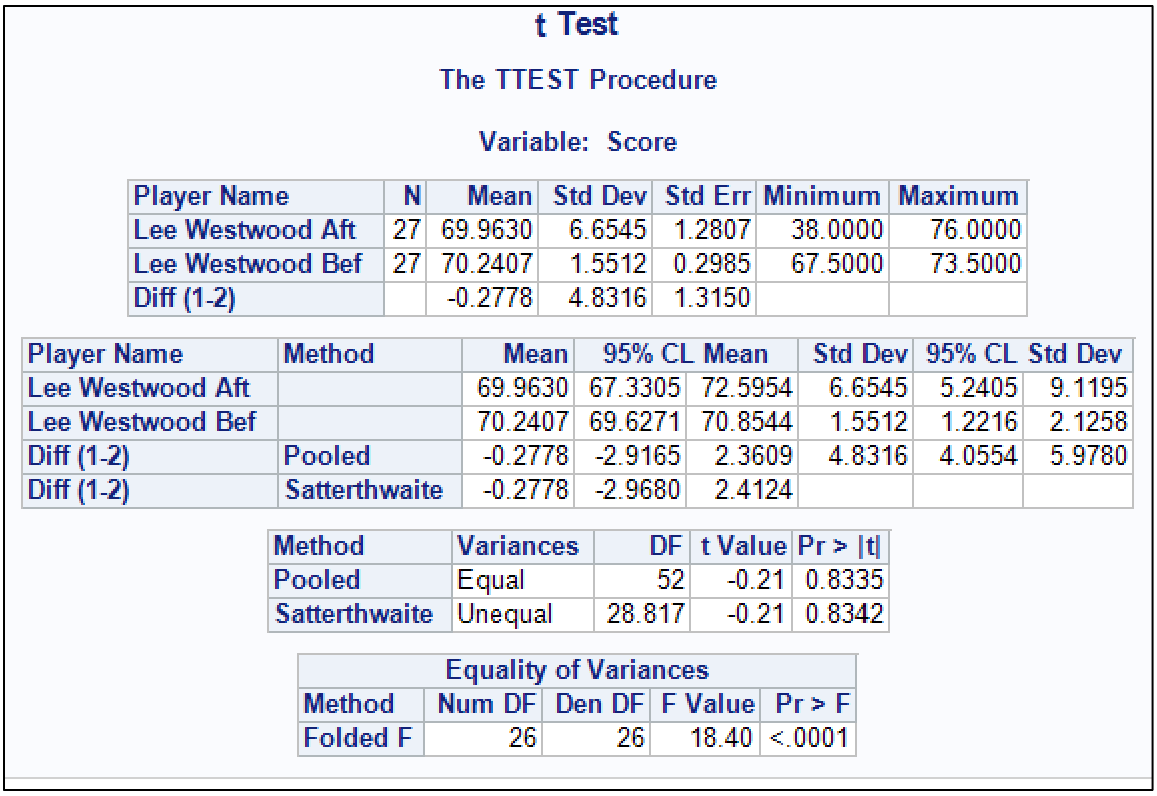


Figure 6.7: Distribution of Lee Westwood's "Before" and "After" earnings

Lee Westwood's Scoring Average results displayed in Figure 6.8 that for either one or two tailed tests, there is no evidence to suggest any difference between the population means of the scoring averages for each time period in question.



t Test

The TTEST Procedure

Variable: Score

Player Name	N	Mean	Std Dev	Std Err	Minimum	Maximum
Lee Westwood Aft	27	69.9630	6.6545	1.2807	38.0000	76.0000
Lee Westwood Bef	27	70.2407	1.5512	0.2985	67.5000	73.5000
Diff (1-2)		-0.2778	4.8316	1.3150		

Player Name	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Lee Westwood Aft		69.9630	67.3305	72.5954	6.6545	5.2405	9.1195
Lee Westwood Bef		70.2407	69.6271	70.8544	1.5512	1.2216	2.1258
Diff (1-2)	Pooled	-0.2778	-2.9165	2.3609	4.8316	4.0554	5.9780
Diff (1-2)	Satterthwaite	-0.2778	-2.9680	2.4124			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	52	-0.21	0.8335
Satterthwaite	Unequal	28.817	-0.21	0.8342

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	26	26	18.40	<.0001

Figure 6.8: t-test results for Lee Westwood's "Before" and "After" scoring average

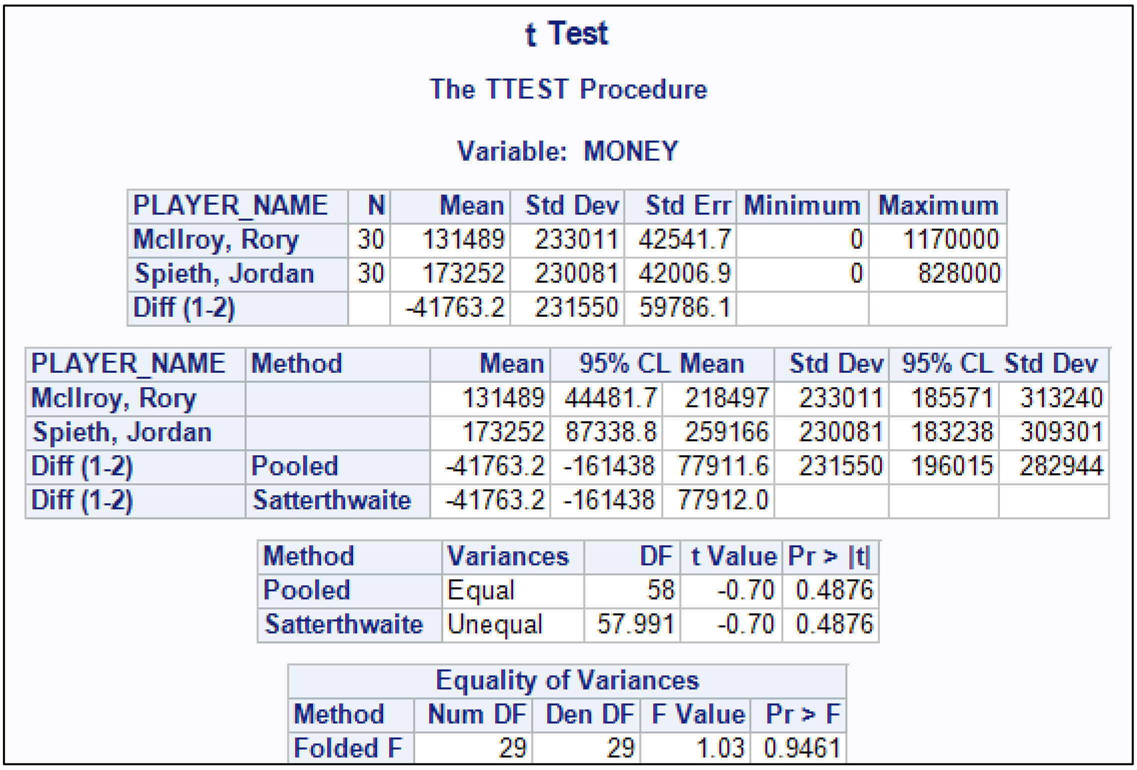
The loss of Billy Foster as Lee Westwood's caddie appears to have made a significant difference to Lee Westwood's earnings at least. By maintaining his scoring average yet experiencing decreased earnings it could be suggested that there are other factors in Lee Westwood's golf game that require further analysis. Perhaps other golfers have lowered their scoring averages to the point where metaphorically standing still results in moving backwards.

It is no surprise that Lee Westwood has re-hired Billy Foster for the 2014 season but more data will be required in order to have certainty as to the effect on Westwood's performance.

6.2.2 Experiment 6 - Jordan Spieth Vs Rory McIlroy

A two sample t-test with a confidence level of 95% was performed on tournaments entered by Rory McIlroy and Jordan Spieth. The total sample size is 60 tournaments with 30 tournament for Rory McIlroy and 30 for Jordan Spieth. These are the first 30 tournaments on the PGA Tour for each player.

As can be seen in Figure 6.9, there is no evidence to suggest that the inferred distribution mean is different between the two sample populations. The two-tailed P-Value of 0.4876 means that one tailed tests would also be insignificant with a P-Value of 0.2438.



t Test

The TTEST Procedure

Variable: MONEY

PLAYER_NAME	N	Mean	Std Dev	Std Err	Minimum	Maximum
McIlroy, Rory	30	131489	233011	42541.7	0	1170000
Spieth, Jordan	30	173252	230081	42006.9	0	828000
Diff (1-2)		-41763.2	231550	59786.1		

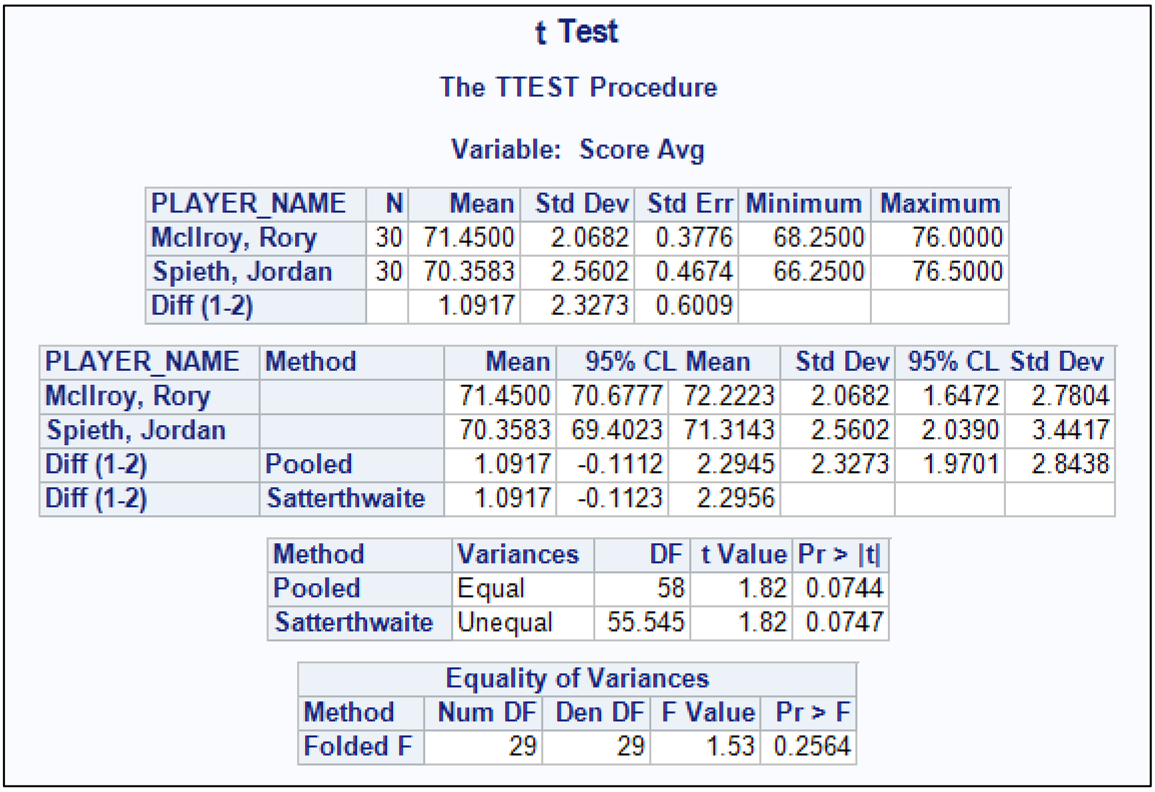
PLAYER_NAME	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
McIlroy, Rory		131489	44481.7	218497	233011	185571	313240
Spieth, Jordan		173252	87338.8	259166	230081	183238	309301
Diff (1-2)	Pooled	-41763.2	-161438	77911.6	231550	196015	282944
Diff (1-2)	Satterthwaite	-41763.2	-161438	77912.0			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	58	-0.70	0.4876
Satterthwaite	Unequal	57.991	-0.70	0.4876

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	29	1.03	0.9461

Figure 6.9: t-test results for the comparison of earnings between Rory McIlroy and Jordan Spieth

The two-tail scoring average t-test seen in Figure 6.10 is also insignificant with a P-Value of 0.0744. Nonetheless, the one tailed test with a P-value of 0.0372 shows that there is sufficient evidence to suggest that Rory McIlroy's scoring average was higher for his entire schedule of tournaments in that period.



t Test

The TTEST Procedure

Variable: Score Avg

PLAYER_NAME	N	Mean	Std Dev	Std Err	Minimum	Maximum
McIlroy, Rory	30	71.4500	2.0682	0.3776	68.2500	76.0000
Spieth, Jordan	30	70.3583	2.5602	0.4674	66.2500	76.5000
Diff (1-2)		1.0917	2.3273	0.6009		

PLAYER_NAME	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
McIlroy, Rory		71.4500	70.6777	72.2223	2.0682	1.6472	2.7804
Spieth, Jordan		70.3583	69.4023	71.3143	2.5602	2.0390	3.4417
Diff (1-2)	Pooled	1.0917	-0.1112	2.2945	2.3273	1.9701	2.8438
Diff (1-2)	Satterthwaite	1.0917	-0.1123	2.2956			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	58	1.82	0.0744
Satterthwaite	Unequal	55.545	1.82	0.0747

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	29	1.53	0.2564

Figure 6.10: t-test results for the comparison of scoring averages between Rory McIlroy and Jordan Spieth

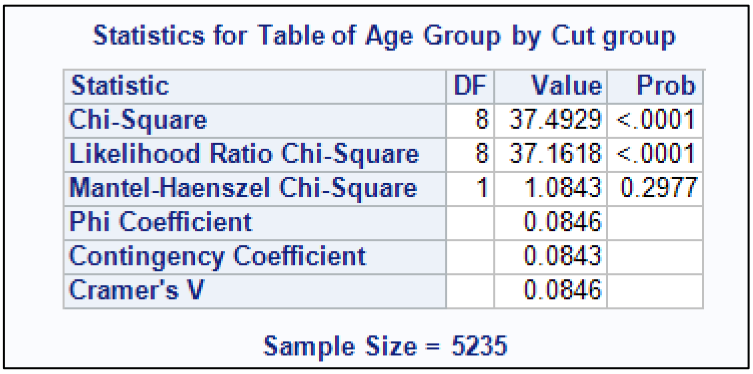
The media comparison of these two players appears to be fair. In terms of money earned at the start of their professional careers, as far a t-test is concerned, the data might as well come from the same golfer. In other words, they are rightly compared and as such, it could be fairly expected to see Jordan Spieth win major championships as Rory McIlroy did early in his career.

6.2.3 Experiment 7 - Age Profile of the Cut Line

In order to determine if there is an association between the age of a golfer and whether they will miss the cut or not, a two way Chi Squared test was implemented. According to (Meyers et al., 2009) Chi Squared tests can be used to test associations between two categorical variables. The null hypothesis is that both variables are independent of each other.

The output from the chi square test for the categorical variables of “Age Group” and “Cut Group” is available in Figure 6.11. The P value for the Chi Square statistic is less than 0.05 (<0.0001) so the null hypothesis is rejected and it can be concluded that there is an association between the age of a golfer and whether they will miss the cut or not.

The reader is referred to Appendix A for further details of the output from the Chi Square test.



Statistics for Table of Age Group by Cut group

Statistic	DF	Value	Prob
Chi-Square	8	37.4929	<.0001
Likelihood Ratio Chi-Square	8	37.1618	<.0001
Mantel-Haenszel Chi-Square	1	1.0843	0.2977
Phi Coefficient		0.0846	
Contingency Coefficient		0.0843	
Cramer's V		0.0846	

Sample Size = 5235

Figure 6.11: Chi Square statistic results for Age Group by Performance

As proven using the chi square test it can be concluded that age does have an impact on a professional golfer’s performance. For a given tournament, Shotlink provides the exact age of each golfer. It can easily be included in the predictive experiment design and a determination can be made as to its overall contribution to the models built.

6.3 Conclusion

This chapter has presented the results from the predictive and statistical experiments in detail. When relevant, initial commentary has been provided in order to explain differences between results, as well as other factors which influenced the experiments during implementation.

Chapter Seven will discuss the evaluation methodology developed for the project, followed by an evaluation the results from this chapter. It will provide an elaboration on initial comments as well as discussing the effectiveness of each experiment. A discussion of model interpretability will also be provided.

7. EVALUATION

In this chapter the results of all the experiments in Chapter Six will be evaluated. Firstly it will discuss the unique challenges of evaluating the predictions for a golfing tournament. It will then separately evaluate the two predictive model methodologies used. The results from each methodology will be discussed and then a comparison of the two methodologies will be presented.

Other areas to be evaluated are predictive model interpretability and the use of the derived attributes created previously. Finally the results from the statistical experiments will be appraised and interpreted.

7.1 The Evaluation Challenge

Typical examples of “real world” or “business world” classification projects are customer churn prediction, targeted marketing campaigns or predicting customer propensity to purchase given products.

Any models built for these types of projects can be easily evaluated using a combination of training, validation and test data. Once the model has been built using the training data it is then validated using a hold out sample of the data which was not used to build the model. This helps to measure if the model over-fits the training data or not.

Tools such as the confusion matrix which tabulates True Positives, False Positives, True Negatives and False Negatives are utilised in order to help choose which model is the best amongst competing models, for example.

As discussed in (Han et al., 2012), analysts can also measure the sensitivity[25] (the true positive recognition rate), specificity (the true negative rate) and precision (the percentage of rows predicted positively and actually are). Visual aids such as ROC Curves also assist an analyst with their model selection decisions.

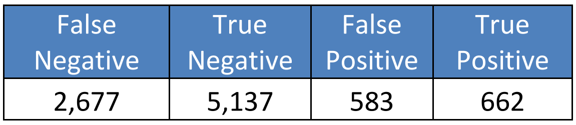
When the best model has been chosen, it can then be used to score data for which a classification prediction is desired. In SAS Enterprise Miner, the analyst is provided

[25] Sensitivity is also known as recall.

with a probability score for each outcome, for example 0.5894 for a No classification and 0.4106 for a Yes classification. Therefore with a cut-off point of 50%, in this example the row will be assigned the NO classification as it is the higher probability.

The training datasets used to build models for Cut Line prediction in this project have been assigned a training label called “MISSED_CUT_IND” (Missed Cut Indicator) which contains a binary classification of either Y (Yes) or N (No).

When this data is used to build a model based on any of the Machine Learning algorithms it will validate the data using a misclassification rate which will compare the actual target value with the predicted target value provided by the newly trained model. Table 7.1 is the Event Classification Table from one of the aggregate models built which shows that the model validation has a high number of False Negatives and False Positives.



False Negative	True Negative	False Positive	True Positive
2,677	5,137	583	662

Table 7.1: Sample Event Classification Table

These traditional methods of model evaluation are unfortunately redundant as an effective means of appraisal for this project. In a business classification problem such as whether a customer is likely to take out a loan or not, all customers are treated equally. That is, customers are not competing against each other for a loan. When the models are evaluated there is no need to consider ranks.

With this in mind, given the nature of the domain being predicted it is erroneous to trust the validation results. This is because the training data is based on the collective data from all golfers since 2009. For the aggregation based models, this is 12,097 rows of data with a split of 7,327 (61%) making the cut to 4,770 (39%) missing the cut.

Obviously this is over multiple tournaments, but the algorithm is not aware of this. When the validation is performed on the hold out set, again it is unaware of the subtlety of the golfing tournament domain and classifies based purely on the rules of the trained model. It is oblivious to the fact that for each tournament with a cut, at least 70 players tend to make the cut.

In essence it is the probability of making the cut which is a much more important indicator than the spurious classification assigned in validation or the scoring of new data.

By taking the probability of missing the cut for each golfer and sorting in descending order such that the lowest probability of missing the cut is first, it is possible to manually assign a classification. This assignment will be N (golfer will not miss the cut) to the top 70 or so golfers who entered the tournament being predicted and Y (golfer will miss the cut) to the remaining golfers.

Evaluation is performed by simply comparing the actual outcome of the tournament to the predicted outcome when manually assigned based on the descending order of probability.

This approach works quite well for Logistic Regression. It is rare for golfers in a given tournament to have the exact same probability of missing the cut. However, by its very nature, decision trees will provide repeating probabilities. This is due to the fact that cases in each node / leaf of a decision tree will be assigned the same predicted probability.

When manually assigning a classification based on the probability score from decision tree models, the issue arose of multiple golfers with the same probability of missing the cut. If these golfers were around the cut line of the lowest 70 golfers or more it means that manual assignment of a Yes or No classification becomes arbitrary. Some of these golfers would be assigned a No classification and others a Yes classification even though they had the same probability of missing the cut. There is no distinguishing feature to fairly separate these golfers.

Table 7.2 illustrates this issue and provides a sample of the predicted probabilities for the Augusta Masters tournament in 2014. Out of 84 players for which data was available, 60 players received a predicted probability of 0.272 of missing the cut. The remaining 24 had a predicted probability of 0.427 or missing the cut.

	MISSED_CUT_IND	
Player	Probability of Y	Probability of N
Mickelson, Phil	0.272299708	0.727700292
Els, Ernie	0.272299708	0.727700292
Stricker, Steve	0.272299708	0.727700292

Furyk, Jim	0.272299708	0.727700292
Jimenez, Miguel A.	0.272299708	0.727700292
Gallacher, Stephen	0.272299708	0.727700292
Senden, John	0.272299708	0.727700292
Couples, Fred	0.427977641	0.572022359
Watson, Tom	0.427977641	0.572022359
Singh, Vijay	0.427977641	0.572022359
Bjorn, Thomas	0.427977641	0.572022359

Table 7.2: Sample Prediction Probabilities for the 2014 Masters

In principle this hampers decision trees from being an effective method of predicting the cut line, when using the collective data from all golfers. More precision is required in order to distinguish between each golfer's chances of making the cut. This could be achieved if the decision trees built have enough distinct terminal nodes providing good separation in probability values.

7.1.1 Cost of Misclassification

As with any classification problem, the benefit and/or cost of predicting a class label correctly or incorrectly needs to be considered. In a medical context, if a model incorrectly predicted that a patient did not have a disease (a false negative) the cost of this misclassification could have serious consequences.

In a golfing context, misclassification may not have the life or death consequences of the previous example. Nonetheless, it could have financial consequences if the models where used to aid a professional golfers decision to play in a tournament or not.

If the model predicted that golfer will not miss the cut but the golfer actually misses the cut, the golfer still has to pay for travel and accommodation to the tournament, as well as paying support staff such as a sports psychologist and their caddy.

Conversely if a golfer chooses not to play a tournament based on the models predictions, they will never know if it was the right decision or not. This introduces the problem that once a model states that a golfer will miss the cut, more data will be required in order for the model to re-assess the golfer's chances of making the cut.

One potential solution to this problem would be to add practice round data to the model when no new tournament data is available. In this way, any changes in a golfer's form could then be used to re-evaluate the prediction.

There will also be times when the professional golfer would have to go against the advice of the model and play regardless, otherwise they may never compete again. If the model's output could be interpreted via an additional descriptive report so that its decision could be explained, then the golfer may be more likely to believe the results.

For example, along with the Play / Don't Play classifier, a report shows that a golfer's trend in Greens In Regulation is falling and the last time this happened for this golfer, they missed 3 cuts in succession. A report such as this would allow the golfer to make a more informed decision, even if that decision is contrary to the models prediction.

7.2 Evaluation of Predictive Models

7.2.1 Collective Methodology Experiment Evaluation

The results from the three collective experiments implemented where relatively disappointing. At first glance, the overall accuracy would appear to be reasonably high given the subject domain and the general difficulty in predicting sporting outcomes. The k-Nearest Neighbour model from Experiment 2 (Table 6.2) produced the highest overall accuracy from the collective experiments at 66.18%. This was followed closely by the segmented Logistic Regression model in Experiment 3 (Table 6.6) with 65.54% overall accuracy.

However the primary question of these experiments was if the golfer will miss the cut for a specific tournament. Making the cut is not the real concern. With this in mind, the missed cut accuracy levels are probably a truer reflection of the results of the experiment. Any models which produced an accuracy rating of 50% or less is considered a failure. This is because it would be expected that based on random guessing that predictions would be correct 50% of the time.

The model with the highest missed cut accuracy was again the k-Nearest Neighbour model from Experiment 2 (Table 6.2**).** This correctly predicted 60% of the golfers who would miss the cut in the Northern Trust Open. Similarly, it was the Logistic Regression model in Experiment 3 (Table 6.6) which produced the second highest result with 59.15% missed cut accuracy.

As alluded to in Chapter Six, it is unlikely to be a coincidence that the Made Cut accuracy results were (nearly) always higher than the Missed Cut accuracy results. The split for the cut line is slightly disproportionate, with 77 golfers making the cut and 69 missing it (a 52/48% split). This room for the extra 8 golfers above the cut line means that there is a greater chance of predicting more of these golfers correctly.

Experiment 1, which used the granular data, based on just the previous tournaments performance, produced the lowest missed cut accuracy rates between 44.77% and 51.51%. From these results, data aggregation appears to be a more effective method of extracting higher accuracy from the models than the use of granular data.

7.2.2 Cluster Evaluation

Analysis of the clustering process suggests that it may require further refinement. The 5 clusters created from the 558 golfer produced unbalanced numbers in each cluster. The smallest segment contained only 49 golfers while the largest contained 183 golfers. Analysis of the entry list for the Northern Trust Open shows that there were 70 golfers in one segment alone.

The results show that the k-Nearest Neighbour model had the worst overall accuracy at 45.94%. Considering that the k-Nearest Neighbour model produced the best results in Experiment 2, this was unexpected. This is because clustering and k-Nearest Neighbour are closely related in that they both consider the similarity of the data they are presented with. The drop off in accuracy for the k-Nearest Neighbour model is the main indicator that refinement of the segmentation process is necessary. This could be achieved by either increasing the number of clusters to be formed, or refining the variables on which the clusters are founded.

7.2.3 Individual Methodology Experiment Evaluation

The results of the individual experiments performed in Experiment 4 sway from the impressive to the disappointing. The prediction accuracy rates for Stuart Appleby (71.42%), Robert Allenby (74.42%) and Padraig Harrington (75%) are very encouraging. The lower predictive accuracy rates for Charlie Beljan (57.14%) and Trevor Immelman (46.15%) suggest that further analysis of their individual skills may be required in order to achieve greater accuracy. This is especially true for Jordan Spieth for whom Experiment 4 was an unqualified failure.

The results also show that sometimes it takes more than one new tournament's training data to "flip" the predicted outcome. In Table 6.9 the predicted probability of Padraig Harrington missing the cut is 97.99% for tournament number 2014190. The outcome of this tournament was that Harrington did indeed miss the cut. For the following tournament (number 2014220) the predicted probability of missing the cut remained high at 94.18%, but Harrington made the cut.

It is then observed that for the next two tournaments the predicted probabilities of missing the cut were 1.66% and 1.16% yet the actual outcome was a made cut followed by a missed cut. In this case it would appear the there is some lag between the introduction of the new data and its influence on the model. One possible trial would be to provide an even higher weighting to the newest data.

7.2.4 Comparison of Methodologies

The collective methodology offers optimism that it could become more accurate with further refinement. This would be the preferred option as it has the advantage of being less time consuming to build training data for. It also requires less manual intervention as changing filters to alter the golfer in focus, or the training data in focus is not required, as is the case with the individual methodology.

However, the individual methodology may offer the best potential for predicting a professional golfer's chances of making the cut in a given tournament. They do not generalise for all golfers and are extremely adaptable as evident by the changes in explanatory variables used from tournament to tournament (Table 6.13).

The results from Experiment 4 suggest that new models would have to be built and customised for every single golfer, as the results range from 46% to 75% accuracy. It may be necessary to further tailor the model to perhaps give varying weightings to each golfers data or to change the cut-off point at which it is determined to output either yes or no.

Experiment 4 relied on a cut-off probability of 50%, so if the golfer was determined to have a probability of 50% or more of missing the cut, the prediction was "Y". It could well be the case that for individual golfers the cut-off-point should be adjusted to 60%, for example, in order to attain better results.

This approach would require a more automated means to extracting data, transforming it and then then re-building the models but is achievable with the appropriate skills and tools.

7.3 Model Interpretability

The original rationale behind the predictive models of this experiment was to aid the decision making process of a professional golfer. If the model predicted that the golfer would miss the cut then that golfer could decide to skip the next tournament and instead practice of their game. This would potentially be a less expensive alternative than entering the tournament.

In order to persuade the golfer that the model is accurate, a number of steps would have to be followed. Firstly, it would be necessary to explain why the model chose to predict that the golfer would miss the cut. For decision tree models, this is quite straight forward as the tree itself can be converted into an English translation.

An exception to this would be when the dimension reduction technique of Principal Components Analysis is used. In SAS Enterprise Miner there is no easy way to determine which attributes where used to create any of the components so unless PCA based models proved to be highly accurate, a golfer will not be convinced of its merits.

k-Nearest Neighbour does provide details as to which neighbours the new data is most similar too. Depending on the value of k (how many neighbours the new data should be compared with), dissemination of these details would require mapping of the nearest neighbour data point values to the training dataset rows. In turn this could be translated into a prediction stating, for example: “Based on the golfer’s previous performance, this is most similar to training row 1267 where the number of birdies scored and the number of three putts are most similar. The outcome from the training row was a missed cut.”

Neural Networks are very difficult to interpret. Similarly to using PCA, any Neural Network model would have to prove itself over a time period before it could be trusted on blind faith by a golfer.

Logistic Regression offers reasonably easy interpretability of its predictions. The significant explanatory factors are provided as the output of the model. In cases were the prediction is that the golfer will miss the cut, the explanatory factors can be offered by way of explanation. In this way the golfer could agree that, for example, their scrambling performance has actually been in decline and this is suggesting that they'll miss the cut.

The second step to convince a golfer to use the model is simply to demonstrate the accuracy of the model over time. If it can save them money then they will use it.

7.4 Value of Derived Attributes

The derived attributes as detailed in Section 5.1.2 appear to have had limited utilisation by any of the models produced. The Form Score attribute was a significant explanatory factor in some of the Logistic Regression models produced for Experiment 4, as noted in Table 6.13.

While they may not have been as useful as hoped, their development aided the understanding of the data and may be useful in future analysis which utilises the other levels of data available through Shotlink.

7.5 Evaluation of Statistical Experiments

The results from the statistical experiments have been very encouraging. It has proven that there is strong evidence to suggest the importance of the caddy and the impact a change "on the bag" can have. For Adam Scott, it was very beneficial to employ Steve Williams as his earnings practically doubled and he is now a major winner. Lee Westwood has experienced the opposite fortunes since Billy Foster left his side. Not only did Westwood's earning halve, but he is still without the elusive major win.

Of course, there is always the chance that other factors apart from the change of a caddy that could account for the differences in fortunes for Scott and Westwood. It is still possible that Scott would have won the Masters regardless of who his caddy was.

Recently, Steve Williams has recently announced his semi-retirement and Foster has re-paired with Westwood. This now presents an opportunity for further analysis. Given time and new data, it will be possible to re-run the experiments to test whether their performances have suffered, or gained as a result of these further changes.

The results of the Jordan Spieth and Rory McIlroy comparison essentially mean that if a person compared their anonymous scorecards for a given number of tournaments, they would not be able to discern which golfer was which. Based on this evidence, given the four year age gap between the two players, it would be not unreasonable to suggest the Jordan Spieth will win numerous major championships in the very near future.

Finally, the finding that a golfer's age has a bearing on whether they'll make the cut or not would seem to be intuitive. Analysis of the "Cell Chi-Square" values (available in Appendix A) show that that the golfers over 45 years of age contribute most to this finding as they have the highest Chi-Square values[26].

Interestingly though, golfers in the age category of 40-45 contribute less than golfers in the 35-40 age category with Chi-Square values of 0.3152 and 2.1948 respectively. Golfers in the 40-45 age category missed the cut 43% of the time in 2013 vs 38% of the time for golfers in the 35-40 age category. This would imply that age is not a significant reason for missing the cut for those golfers in the 40-45 age category. There is simply is not a strong enough association for this age category. This suggests that other factors may be involved for the 40-45s.

7.6 Software Evaluation

SAS Enterprise Miner is a reputable tool. It is used by many companies throughout the globe to aid data mining and predictive analytics. In Ireland the Revenue Commissioners use it to predict which companies are likely to yield results from a tax audit (Cleary, 2011). Primarily used in a business context, it has limitations when predicting the cut line for golf tournaments, especially from a model evaluation standpoint (as outlined in Sub-chapter 6.1).

[26] The higher the Chi-Square value, the stronger the association is between the two variables.

For experiments 1,2 and 3, SAS Enterprise Miner provided an interface which allowed for easy implementation of these models. It was also essential for a simple approach to segmenting the golfers into different clusters.

Experiment 4 was not particularly suited to SAS Enterprise Miner as it was not as suited to the design of a simple, iterative approach based on one training dataset file which could be filtered according to specific golfers and tournaments. It may have been possible to implement an automated solution in Enterprise Miner using SAS code directly but the author did not have time to research this.

SAS Enterprise Guide proved to be an invaluable substitute tool for Experiment 4. Its ability to select data from separate worksheets within the same Excel spreadsheet, along with simple tools to filter the data allowed a “one file suits all” approach. Once set up correctly, it was a case of setting filters, running the Logistic Regression model, recording the predictions and then adjusting the filters to allow the extra data from the tournament which had just been predicted into the training data. The additional output provided by Enterprise Guide in relation to goodness of fit tests etc. was also beneficial when compared with output from Enterprise Miner.

The statistical experiments were also very straightforward to design and implement using SAS Enterprise Guide. The tests used in this project are well documented and allow for swift, effective interpretation.

7.7 Conclusion

This chapter has discussed the outcomes and interpretations of both the predictive and statistical experiments performed. It has outlined the potential of the predictive models to determine which golfers will miss the cut in a given tournament. It has presented an argument that the set of individual methodology experiments may offer the greatest potential for models of high accuracy, though further refinement is required.

The evaluation and success of the statistical experiments would suggest that the Shotlink dataset is ideally suited to answer many more questions and follow up experiments can be easily run to determine the validity of their results.

Chapter Eight, the final chapter will discuss the conclusions of project, how it as contributed to the body of knowledge and whether it has successfully achieved its stated objectives. Potential areas for future research will also be outlined.

8. CONCLUSION

This chapter will discuss how the dissertation has achieved its stated goals. It will summarise the initial aims and objectives of the project and assess the projects results against these. The projects contribution to the body of knowledge will be discussed along with potential areas for future research for which this project can be the foundation.

8.1 Research Definition & Research Overview

Prior to this dissertation, Shotlink data had been primarily used to create new ways of measuring the past. New statistics had been created or proposed which are of limited value unless they are combined with several other measures of performance. The use of Machine Learning to preform predictive experiments had not been utilised as part of the previously published research.

This dissertation sought to determine if Data Analytical techniques could be implemented using the Shotlink dataset. Its aims were to determine whether it was possible to predict professional golfer performance using the available data, and if suitable statistical analysis could be applied in order to test a number of hypotheses.

Research was carried out to determine the value and inspiration that could be drawn from previously published literature. The topics of this research centred around the headings of Performance Assessment, Performance Analysis and the external influences on a golfer’s performance.

The objectives achieved by this dissertation were:

- Previous research from different aspects of golf science reviewed
- Statistical tests reviewed for suitability to secondary research problem
- Shotlink data loaded into MySQL database and subsequently assessed and transformed as required for the needs of the experiments

- Adaptable “Golfer Analytical Records” were successfully created which could be used by the competing methodologies for training purposes
- Predictive models and statistical experiments where successfully designed and implemented
- Evaluate of the aforementioned predictive models and statistical experiments to determine their success or failure and to interpret these results.

8.2 Contributions to the Body of Knowledge

This dissertation has contributed to the body of knowledge by showing that the use of Data Analysis techniques on the PGA Tour’s Shotlink dataset can predict a professional golfer’s performance relatively successfully. It has demonstrated that the Shotlink dataset is ideally suited for answering questions relating to golfer comparisons, either of the same golfer at different points of time, or comparing two different golfers.

It has established the potential for prediction of the cut line using two separate methodologies utilising either collective or individual past performance data. The collective methodology was based on the theory that if a golfer is experiencing similar performance to that of previously recorded performances from other golfers, then their tournament outcome should be predictable. Accuracy rates of up to 60% were achieved using this methodology.

The individual methodology demonstrated the significant potential for developing multiple predictive models using iterative training data. This methodology predicted a golfer’s performance based solely on that individual golfer’s previous performances. It did not use other golfer’s data to train. With accuracy rates of up to 75% achieved for some golfers using this model there is a large amount of potential for this methodology.

Separately the statistical analysis of this dissertation has contributed the following outcomes:

- There does indeed appear to be a relationship between the caddy a golfer uses and that golfer’s performance in terms of scoring averages and prize money earned.

- The comparison of Jordan Spieth to Rory McIlroy in the media is fair as both their performances from early tournaments played on the PGA Tour are very similar.
- The age of a golfer does influence their ability to make the cut or not.

8.3 Experimentation, Evaluation and Limitation

The experimentation performed in this dissertation utilised numerous models and techniques in order to achieve its goals. Two methodologies where designed for the predictive modelling experiments, one based on the use of collective golfer data and the other based on use of individual golfer data only.

Within the collective methodology, experiments attempted to automatically reduce the number of dimensions and create clusters of similar golfers before being deployed into the four separate classification models, Decision Trees, k-Nearest Neighbour, Neural Networks and Logistic Regression. The entry list for Northern Trust Open was then used to score the models on their predictions. Logistic Regression was the sole technique used to create models based on the individual methodology.

The statistical experiments utilised in this dissertation consisted of t-tests to test claims of various hypotheses and table analysis which uses a Chi - Square test to determine associations between two categorical variables.

An evaluation of the experiments performed suggests that there is merit and potential to utilising Shotlink data in order to predict professional golfer performance. The SQL scripts coded to create the Golfer Analytical Records which trained the models was coded in a flexible way. The code was written in a modular fashion which allowed for features to included or excluded easily, or for simple changes in the number of tournaments to aggregate data for.

Overall, the experiment results would suggest that the use of the individual methodology has the best chance of being deployed into the “real world” of professional golf. They provided the highest accuracy for some of the trial golfers and they are fairly straightforward to interpret.

The dissertation and its experiments were limited by the following factors:

- Potential issues regarding dual tour members and the likelihood of not have the most up-to-date data available for them
- Lack of data availability from other worldwide tours
- Lack of time / appropriate tools to create individual models for every golfer who competed on the PGA Tour
- Complex psychological issues such as determining if a golfer is suffering from the Yips could not be brought into the data
- Only event level detail was used to throughout the project.

8.4 Future Work & Research

With the vast amount of data available via Shotlink there are numerous interesting and exciting areas and ideas on which data analysis can be carried out to further the work carried out in this project.

- **Data from Non Shotlink sources**

One of the potential drawbacks for the models created for this project was the lack of data from additional sources. As mentioned previously, golfers who are members of other professional tours such as the European or Asian tours are competing in tournaments for which there is no Shotlink data. This means that there could be gaps in the knowledge of the model so it is unaware that a golfer has started to come back to their best, for example. If data from these additional tours could be made available in a format similar to the Shotlink datasets then the models could become more powerful predictive tools.

Practice round data would also be a useful addition if it could be collected appropriately and applied to the models. In this way a golfer could supplement his tournament play with this extra data which would allow the model to re-train and re-evaluate. A smartphone app could be developed which allows the golfer's caddy to input the data of the practice round.

- **Fatigue**

The effect of fatigue on a golfer could also be considered for future research. This would need knowledge of a golfer's tournament schedule and perhaps knowledge of other external activities such as gym training, or their penchant for socialising etc. With this knowledge it would be possible to monitor golfers to establish if they were liable to "burn out" and how this could affect their performance.

- **Golfer Interactions**

A particularly interesting area for future work could involve the study of how golfers interact with each other when paired together in either stroke play or match play competitions. Animosity between golfers, or indeed, friendly rivalries tend to attract significant media coverage. This would allow for simple discovery of which golfers any research should focus on.

Potential avenues of research could revolve around the effects on scoring averages, total number of putts etc. to determine if there is any difference in golfer behaviour when paired with certain other golfers. The pairings for each tournament are readily available through the Shotlink website but is not directly available from the downloadable datasets. It can however be derived from this data.

The effect of playing in the same group as a golfer who is known for slow play is another area of golfer interaction which could yield interesting results. Slow play is a cause of frustration to both golfers paired with the slow playing golfer and those golfers who may be held up behind them.

Shotlink provides details of the start time and completion time for each golfer, for hole, so it would be possible to deduce which golfers are slower than average. Multiple areas of analysis could then be performed based on this knowledge.

- **Competitor Analysis**

The Shotlink data lends itself to the creation of competitor analysis reports which could be used for match play competitions or team competitions such as the Ryder Cup. Reports on the performance statistics of the rival golfers could be used to highlight the strengths and weaknesses of the competition. This could be used to gain advantage over the opposition.

For example, a Ryder Cup captain could use this data to aid in their determination of which golfers to choose for their wildcard picks. By analysing the performance metrics of the golfers who will automatically qualify for the team, the captain will be able to ascertain if there is any imbalance in the team and search for golfers who have the necessary strengths to restore the skills equilibrium to the team.

During competition, data from these reports could also be used to pick the best partnerships for the fourball and foursomes matches[27]. At the start of each day, the opposing captains submit their team lists and the order in which they will play, separately. This means that neither captain can be sure which golfers from either team will actually compete against each other, as neither has knowledge of the other's picks. However, the use of performance metrics together with a captain anticipating the pairs from the opposition could allow for successful player partnerships to be formed.

In knockout matchplay events, a golfer could compare their performance metrics against their competitor. While this may seem of limited value, as the golf course provides the main opposition, relevant data could still be useful. Take for instance a golfer who knows that their competitor is quite weak for sand saves. If, during the match their competitor finds themselves in the bunker, then the golfer knows that their competitor is in a position of weakness. This may allow them to relieve any potential pressure the may be experiencing.

- **Extension of research into the LPGA and Champions tours**

All the literature encountered for research of this dissertation was based on male professional golfers who played on the PGA Tour. Certainly, all the research which used Shotlink data as its foundation did not utilise the data from either the Ladies PGA (LPGA) tour or the Champions Tours. This would imply that there is a "gap in the

[27] Fourball and foursomes consist of two two-man teams competing against each other. In Fourball matches, each golfer plays with their own ball. In foursomes matches, each team has only one ball so team members play alternate shots,

market” per say, for previous studies on the game of golf to be extended to both these tours.

The research carried out by this dissertation could be quite easily transitioned to the other tours for which Shotlink data is captured. It would be interesting to see whether the predictive models outlined in this dissertation would produce results of similar or different accuracy. Until this research is carried out, it would be difficult to hypothesise was to whether any differences would be expected.

A potential output from these transitioned experiments could be a study in the similarity and differences between the games of professional male and female golfers.

8.5 Conclusion

This chapter draws to a conclusion the research and experimentation conducted to determine the predictability of professional golfer performance using the PGA Tour’s Shotlink dataset. A summary of the research, its objectives and the experiments designed and implemented have been discussed. An evaluation of the experiment and limitations encounter was also presented. Finally a discussion on the areas of future work which could carry on from this dissertation has also been outlined.

The potential offered by the predictive models of this dissertation offer an exciting glimpse towards the power and capability of Data Analytics combined with the Shotlink dataset. Further refinement of the models created, or finding answers to correctly phrased questions could eventually do for golf what SABERMETRICS has done for baseball. That is, to bring a completely new perspective on how data is used to add value to the game and its participants.

BIBLIOGRAPHY

Agresti, A., 2007. An Introduction to Categorical DataAnalysis. Wiley.

Alexander, D.L., Kern, W., 2005. Drive for Show and Putt for Dough? An Analysis of the Earnings of PGA Tour Golfers. J. Sports Econ. 6, 46–60. doi:10.1177/1527002503260797

Broadie, M., 2008. Assessing Golfer Performance Using Golfmetrics.

Broadie, M., 2011a. Putts Gained - Measuring Putting on the PGA Tour.

Broadie, M., 2011b. Assessing Golfer Performance on the PGA TOUR.

Cleary, D., 2011. Predictive Analytics in the Public Sector: Using Data Mining to Assist Better Target Selection for Audit. Electron. J. E-Gov. Vol. 9 Issue 2 2011 132–140.

Coate, D., Toomey, M., 2012. Do Professional Golf Tour Caddies Improve Player Scoring? J. Sports Econ. 1527002512458799. doi:10.1177/1527002512458799

Connolly, R., Rendleman Jr., R., 2008. Dominance, Intimidation, and “Choking” on the PGA Tour.

Cosgrove, B., 2014. Discussion of factors which influence a professional golfers ability from one tournament to the next.

D’Arcy, A., 2012. Rregression Analysis.

Der, G., Everitt, B., 2007. Basic Statistics Using SAS Enterprise Guide: A Primer. SAS Institute.

Fearing, D., Acimovic, J., Graves, S., 2010. How to Catch a Tiger: Understanding Putting Performance on the PGA Tour (SSRN Scholarly Paper No. ID 1538300). Social Science Research Network, Rochester, NY.

Fried, H.O., Tauer, L.W., 2011. The impact of age on the ability to perform under pressure: golfers on the PGA tour. J. Product. Anal. 35, 75–84. doi:10.1007/s11123-009-0151-9

Gray, W., 2014. Harrington’s caddie collapses at Volvo Champions [WWW Document]. Golf Channel. URL http://www.golfchannel.com/news/golftalkcentral/harringtons-caddie-collapses-volvo-champions/ (accessed 3.23.14).

Han, J., Kamber, M., Pei, J., 2012. Data Mining Concepts and Techniques, Third. ed. Morgan Kaufmann.

Heiny, E.L., 2008. PGA Tour Pro: Long but Not so Straight.

Huguenin, M., 2013. Jason Day credits caddy for easing World Cup of Golf nerves | Sportal Australia [WWW Document]. URL http://www.sportal.com.au/golf/news/jason-day-credits-caddy-for-easing-world-cup-of-golf-nerves/1azt579yqb03v1wryno2gpckvw (accessed 2.28.14).

Iso-Ahola, S.E., Mobily, K., 1980. “PSYCHOLOGICAL MOMENTUM”: A PHENOMENON AND AN EMPIRICAL (UNOBTRUSIVE) VALIDATION OF ITS INFLUENCE IN A COMPETITIVE SPORT TOURNAMENT. Psychol. Rep. 46, 391–401. doi:10.2466/pr0.1980.46.2.391

Klämpfl, M.K., Lobinger, B.H., Raab, M., 2013. Reinvestment – the Cause of the Yips? PLoS ONE 8, e82470. doi:10.1371/journal.pone.0082470

Lavalle, D., Bruce, D., Gorely, T., 2004. The Golfer-Caddie Partnership: An Exploratory Investigation into the Role of the Caddie [WWW Document]. URL http://www.athleticinsight.com/Vol6Iss1/GolfCaddieRole.htm#References (accessed 2.27.14).

Lewis, M., 2003. Moneyball.

Mac Namee, B., Kelleher, J., 2012. Information Based Learning.

Meyers, L.S., Gamst, G., Guarino, A.J., 2009. Data Analysis Using Sas Enterprise Guide.

Montgomery, D.C., Runger, G.C., 2011. Applied Statistics and Probabilty for Engineers.

Noer, M., 2012. On the Edge: Money, Life and Loneliness on the Fringe of the PGA Tour [WWW Document]. Forbes. URL http://www.forbes.com/sites/michaelnoer/2012/02/08/ben-martin-pga-tour/ (accessed 2.28.14).

Peters, A., 2008. Determinants of Performance on the PGA Tour.

PGA Tour Strokes Gained Putting Stats [WWW Document], 2014. URL http://www.pgatour.com/stats/stat.02564.html (accessed 5.15.14).

Riccio Ph.D, L., 2012. THE BEST FAIRWAY BALL STRIKER ON TOUR !

Ridenoure, W., 2005. Regression Analysis of Golf Statistics and their Relationship to PGA Tour Performance.

Rotella, D.B., 2008. Your 15th Club: The Inner Secret to Great Golf.

Rumsey, D.J., 2007. Intermediate Statistics For Dummies. John Wiley & Sons.

Savage, K.J., 2012. Making the Cut: Psychological Momentum on the PGA Tour.

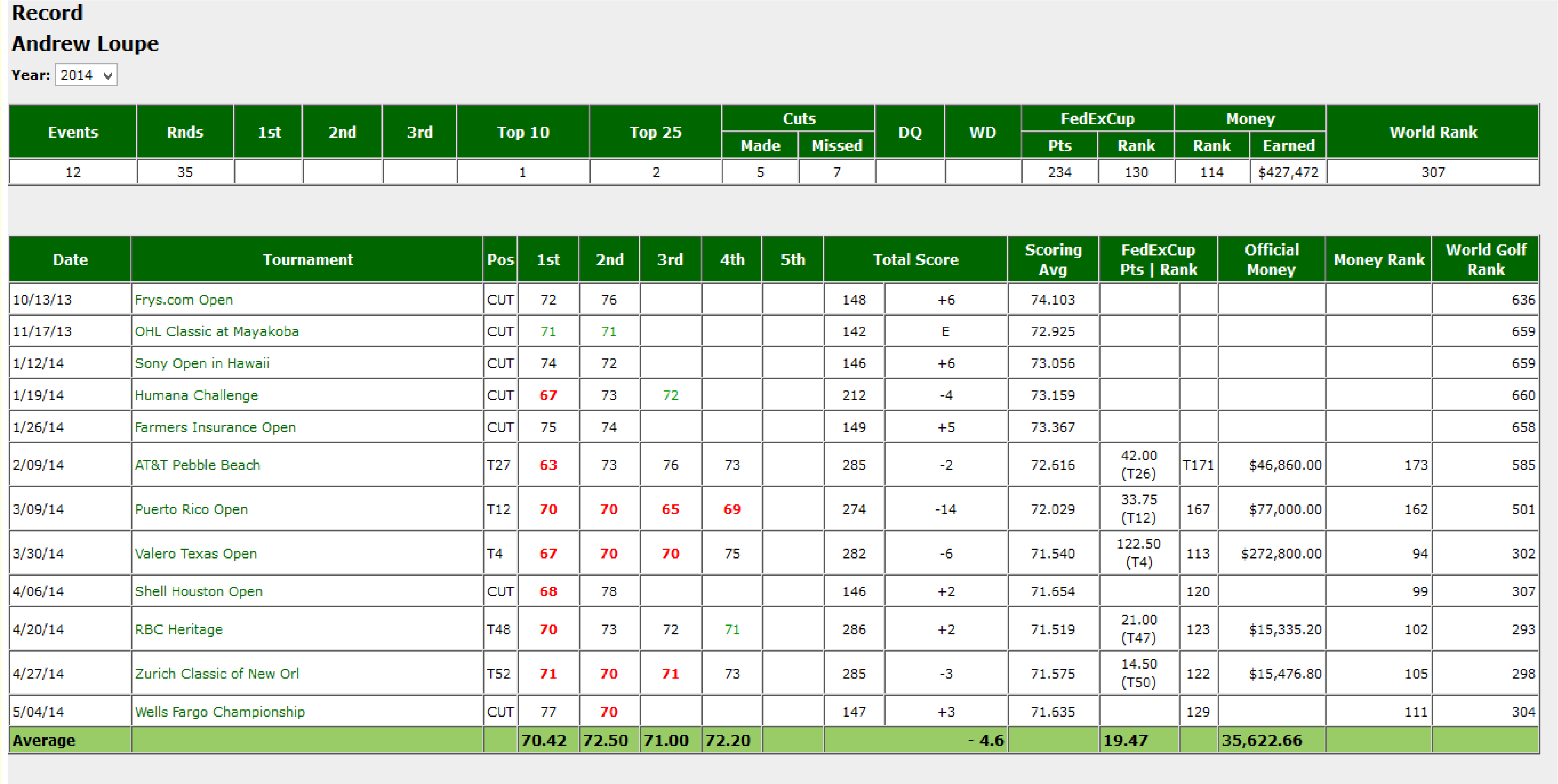
Schulz, R., Curnow, C., 1988. Peak Performance and Age Among Superathletes: Track and Field, Swimming, Baseball, Tennis, and Golf. J. Gerontol. 43, P113–P120. doi:10.1093/geronj/43.5.P113

Sen, K.C., 2012. Mapping statistics to success on the PGA Tour: Insights from the use of a single metric. Sport Bus. Manag. Int. J. 2, 39–50. doi:10.1108/20426781211207656

Yau, N., 2008. How to Read (and Use) a Box-and-Whisker Plot.

APPENDIX A – TOURNAMENT RECORDS & CHI-SQUARE OUTPUT

Andrew Loupe Tournament Record 2014 (as at 15/05/2014)



Record

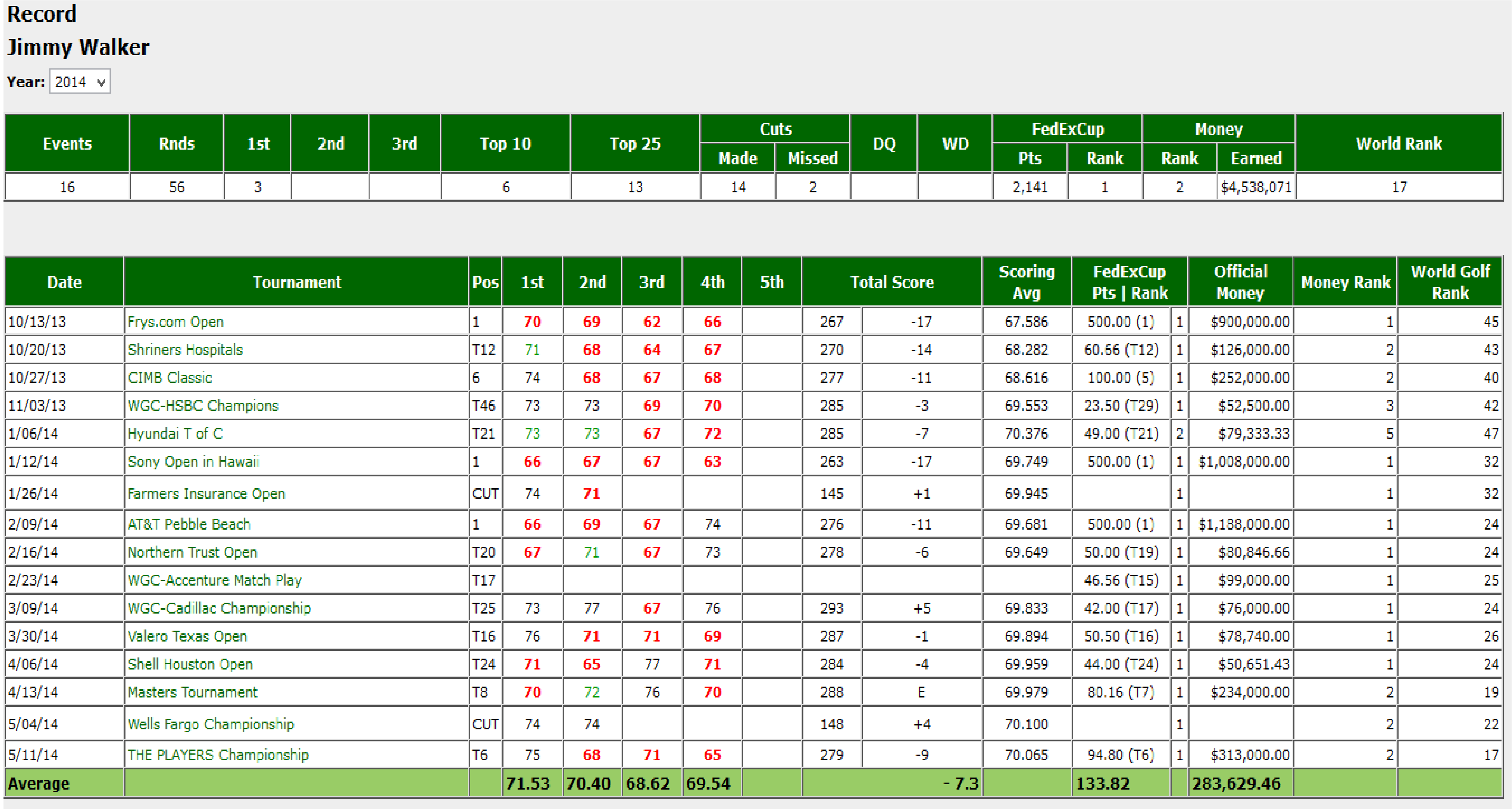
Andrew Loupe

Year: 2014

Events	Rnds	1st	2nd	3rd	Top 10	Top 25	Cuts		DQ	WD	FedExCup		Money		World Rank
							Made	Missed			Pts	Rank	Rank	Earned	
12	35				1	2	5	7			234	130	114	$427,472	307

Date	Tournament	Pos	1st	2nd	3rd	4th	5th	Total Score		Scoring Avg	FedExCup Pts | Rank		Official Money	Money Rank	World Golf Rank
10/13/13	Frys.com Open	CUT	72	76				148	+6	74.103					636
11/17/13	OHL Classic at Mayakoba	CUT	71	71				142	E	72.925					659
1/12/14	Sony Open in Hawaii	CUT	74	72				146	+6	73.056					659
1/19/14	Humana Challenge	CUT	67	73	72			212	-4	73.159					660
1/26/14	Farmers Insurance Open	CUT	75	74				149	+5	73.367					658
2/09/14	AT&T Pebble Beach	T27	63	73	76	73		285	-2	72.616	42.00 (T26)	T171	$46,860.00	173	585
3/09/14	Puerto Rico Open	T12	70	70	65	69		274	-14	72.029	33.75 (T12)	167	$77,000.00	162	501
3/30/14	Valero Texas Open	T4	67	70	70	75		282	-6	71.540	122.50 (T4)	113	$272,800.00	94	302
4/06/14	Shell Houston Open	CUT	68	78				146	+2	71.654		120		99	307
4/20/14	RBC Heritage	T48	70	73	72	71		286	+2	71.519	21.00 (T47)	123	$15,335.20	102	293
4/27/14	Zurich Classic of New Orl	T52	71	70	71	73		285	-3	71.575	14.50 (T50)	122	$15,476.80	105	298
5/04/14	Wells Fargo Championship	CUT	77	70				147	+3	71.635		129		111	304
Average			**70.42**	**72.50**	**71.00**	**72.20**			**- 4.6**		**19.47**		**35,622.66**		

Jimmy Walker Tournament Record 2014 (as at 15/105/2014)



Record

Jimmy Walker

Year: 2014

Events	Rnds	1st	2nd	3rd	Top 10	Top 25	Cuts		DQ	WD	FedExCup		Money		World Rank
							Made	Missed			Pts	Rank	Rank	Earned	
16	56	3			6	13	14	2			2,141	1	2	$4,538,071	17

Date	Tournament	Pos	1st	2nd	3rd	4th	5th	Total Score		Scoring Avg	FedExCup Pts | Rank		Official Money	Money Rank	World Golf Rank
10/13/13	Frys.com Open	1	70	69	62	66		267	-17	67.586	500.00 (1)	1	$900,000.00	1	45
10/20/13	Shriners Hospitals	T12	71	68	64	67		270	-14	68.282	60.66 (T12)	1	$126,000.00	2	43
10/27/13	CIMB Classic	6	74	68	67	68		277	-11	68.616	100.00 (5)	1	$252,000.00	2	40
11/03/13	WGC-HSBC Champions	T46	73	73	69	70		285	-3	69.553	23.50 (T29)	1	$52,500.00	3	42
1/06/14	Hyundai T of C	T21	73	73	67	72		285	-7	70.376	49.00 (T21)	2	$79,333.33	5	47
1/12/14	Sony Open in Hawaii	1	66	67	67	63		263	-17	69.749	500.00 (1)	1	$1,008,000.00	1	32
1/26/14	Farmers Insurance Open	CUT	74	71				145	+1	69.945		1		1	32
2/09/14	AT&T Pebble Beach	1	66	69	67	74		276	-11	69.681	500.00 (1)	1	$1,188,000.00	1	24
2/16/14	Northern Trust Open	T20	67	71	67	73		278	-6	69.649	50.00 (T19)	1	$80,846.66	1	24
2/23/14	WGC-Accenture Match Play	T17									46.56 (T15)	1	$99,000.00	1	25
3/09/14	WGC-Cadillac Championship	T25	73	77	67	76		293	+5	69.833	42.00 (T17)	1	$76,000.00	1	24
3/30/14	Valero Texas Open	T16	76	71	71	69		287	-1	69.894	50.50 (T16)	1	$78,740.00	1	26
4/06/14	Shell Houston Open	T24	71	65	77	71		284	-4	69.959	44.00 (T24)	1	$50,651.43	1	24
4/13/14	Masters Tournament	T8	70	72	76	70		288	E	69.979	80.16 (T7)	1	$234,000.00	2	19
5/04/14	Wells Fargo Championship	CUT	74	74				148	+4	70.100		1		2	22
5/11/14	THE PLAYERS Championship	T6	75	68	71	65		279	-9	70.065	94.80 (T6)	1	$313,000.00	2	17
Average			**71.53**	**70.40**	**68.62**	**69.54**			**- 7.3**		**133.82**		**283,629.46**		

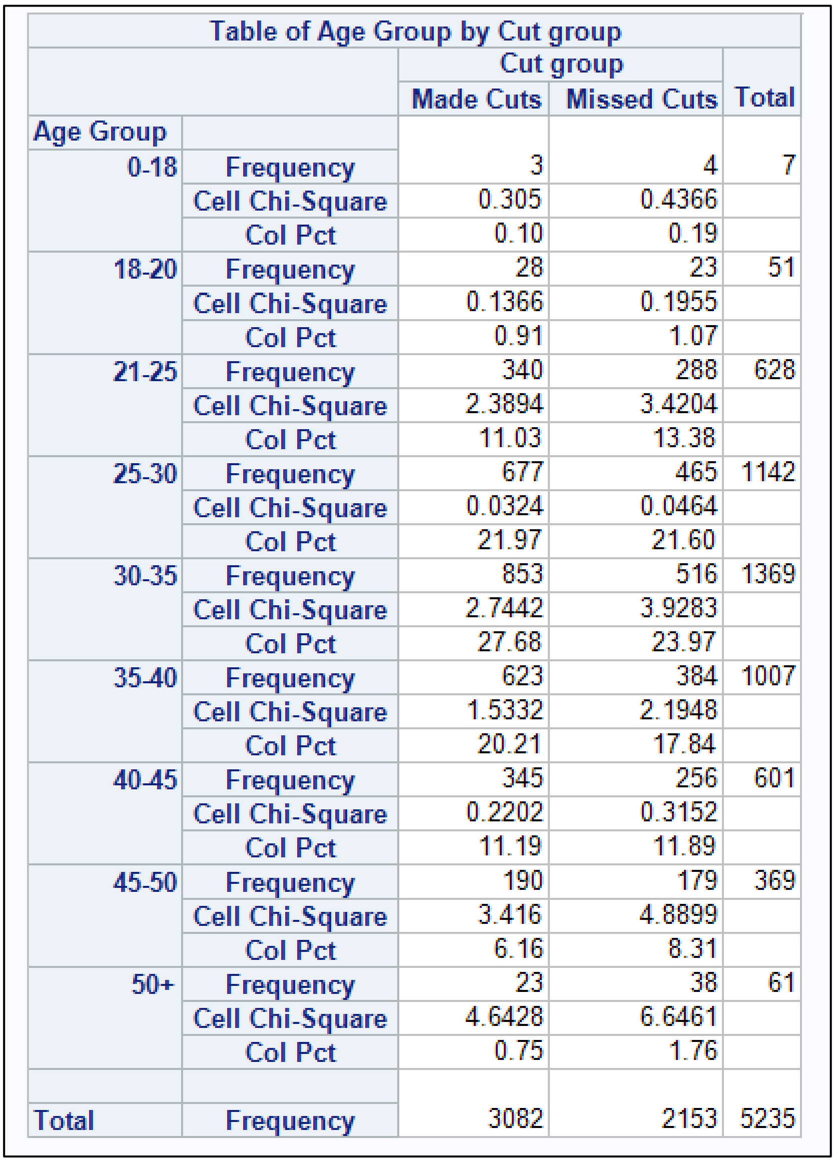


Table of Age Group by Cut group				
		Cut group		
		Made Cuts	Missed Cuts	Total
Age Group				
0-18	Frequency	3	4	7
	Cell Chi-Square	0.305	0.4366	
	Col Pct	0.10	0.19	
18-20	Frequency	28	23	51
	Cell Chi-Square	0.1366	0.1955	
	Col Pct	0.91	1.07	
21-25	Frequency	340	288	628
	Cell Chi-Square	2.3894	3.4204	
	Col Pct	11.03	13.38	
25-30	Frequency	677	465	1142
	Cell Chi-Square	0.0324	0.0464	
	Col Pct	21.97	21.60	
30-35	Frequency	853	516	1369
	Cell Chi-Square	2.7442	3.9283	
	Col Pct	27.68	23.97	
35-40	Frequency	623	384	1007
	Cell Chi-Square	1.5332	2.1948	
	Col Pct	20.21	17.84	
40-45	Frequency	345	256	601
	Cell Chi-Square	0.2202	0.3152	
	Col Pct	11.19	11.89	
45-50	Frequency	190	179	369
	Cell Chi-Square	3.416	4.8899	
	Col Pct	6.16	8.31	
50+	Frequency	23	38	61
	Cell Chi-Square	4.6428	6.6461	
	Col Pct	0.75	1.76	
Total	Frequency	3082	2153	5235

Detailed Chi Squared Association Analysis of Age Group and Performance

APPENDIX B – EVENT LEVEL ATTRIBUTES

Full List of fields available in Shotlink Event Level Detail Dataset

COLUMN	DATA TYPE
TOUR	VARCHAR(5)
TOURNAMENT_YEAR	INT(11)
TOURNAMENT_NUMBER	INT(11)
PERMANENT_TOURNAMENT_NUMBER	INT(11)
TEAM_ID	VARCHAR(255)
TEAM_NUMBER	INT(11)
PLAYER_NUMBER	INT(11)
PLAYER_NAME	VARCHAR(25)
PLAYER_AGE_YEARS,_MONTHS,DAYS	VARCHAR(12)
EVENT_NAME	VARCHAR(255)
OFFICIAL_EVENTY/N	VARCHAR(5)
FEDEXCUP_POINTS	DECIMAL(65,30)
MONEY	DECIMAL(12,2)
FINISH_POSITIONNUMERIC	INT(11)
FINISH_POSITIONTEXT	VARCHAR(5)
ROUND_1_SCORE	INT(11)
ROUND_1_POS	INT(11)
ROUND_2_SCORE	INT(11)
ROUND_2_POS	INT(11)
ROUND_3_SCORE	INT(11)
ROUND_3_POS	INT(11)
ROUND_4_SCORE	INT(11)
ROUND_4_POS	INT(11)
ROUND_5_SCORE	INT(11)
ROUND_5_POS	INT(11)
ROUND_6_SCORE	INT(11)
ROUND_6_POS	INT(11)
LOWEST_ROUND	INT(11)
TOTAL_STROKES	INT(11)
TOTAL_ROUNDS	INT(11)
STROKE_AVERAGE_RANK	INT(11)
SCORING_AVGTOTAL_ADJUSTMENT	DECIMAL(65,30)
SCORING_AVGTOTAL_ADJUSTMENT_-_RANK	INT(11)
EAGLES	INT(11)
EAGLES_RANK	INT(11)
BIRDIES	INT(11)
BIRDIES_RANK	INT(11)
PARS	INT(11)
BOGEYS	INT(11)
BOGEYS_RANK	INT(11)
DOUBLES	INT(11)
OTHERS	INT(11)
TOTAL_HOLES_OVER_PAR	INT(11)
BOGEY_AVOIDANCE_RANK	INT(11)
BIRDIE_OR_BETTER_CONV_%#_BIRDIES	INT(11)
BIRDIE_OR_BETTER_CONV_%#_GREENS_HIT	INT(11)
LONGEST_DRIVE	INT(11)

LONGEST_DRIVE_RANK	INT(11)
DRIVING_DISTANCETOTAL_DISTANCE	INT(11)
DRIVING_DISTANCETOTAL_DRIVES	INT(11)
DRIVING_DISTANCE_RANK	INT(11)
DRIVING_DIST._-_ALL_DRIVESTOT._DIST.	INT(11)
DRIVING_DIST._-_ALL_DRIVESRANK	INT(11)
DRIVES_OVER_300_YARDS_#_OF_DRIVES	INT(11)
DRIVING_ACC._%FAIRWAYS_HIT	INT(11)
DRIVING_ACC._%POSSIBLE_FAIRWAYS	INT(11)
DRIVING_ACCURACY_RANK	INT(11)
TOTAL_DRIVING_RANK	INT(11)
LEFT_ROUGH_TENDENCYTOTAL_LEFT_ROUGH	INT(11)
RIGHT_ROUGH_TENDENCYTOTAL_RIGHT_ROUGH	INT(11)
APP.__50-75_YDSFT	DECIMAL(65,30)
APP.__50-75_YDSATTEMPTS	INT(11)
APP.__75-100_YDSFT	DECIMAL(65,30)
APP.__75-100_YDSATTEMPTS	INT(11)
APP.__100-125_YDSFT	DECIMAL(65,30)
APP.__100-125_YDSATTEMPTS	INT(11)
APP.__50-125_YARDSFT	DECIMAL(65,30)
APP.__50-125_YARDSATTEMPTS	INT(11)
APPROACHES__125-150_YARDSFT	DECIMAL(65,30)
APPROACHES___125-150_YARDSATTEMPTS	INT(11)
APPROACHES___150-175_YARDSFT	DECIMAL(65,30)
APPROACHES___150-175_YARDSATTEMPTS	INT(11)
APPROACHES___175-200_YARDSFT	DECIMAL(65,30)
APPROACHES___175-200_YARDSATTEMPTS	INT(11)
APPROACHES___>200_YARDSFT	DECIMAL(65,30)
APPROACHES___>200_YARDSATTEMPTS	INT(11)
APP.___50-75_YDSFT_-_ROUGH	VARCHAR(12)
APP.___50-75_YDSATTEMPTS_-_ROUGH	INT(11)
APP.___75-100_YDSFT_-_ROUGH	DECIMAL(65,30)
APP.___75-100_YDSATTEMPTS_-_ROUGH	INT(11)
APP.___100-125_YDSFT_-_ROUGH	DECIMAL(65,30)
APP.___100-125_YDSATTEMPTS_-_ROUGH	INT(11)
APPROACHES___50-125_YARDSFT_-_ROUGH	DECIMAL(65,30)
APPROACHES___50-125_YARDSATTEMPTS_-_ROUGH	INT(11)
APPROACHES___125-150_YARDSFT_-_ROUGH	DECIMAL(65,30)
APPROACHES___125-150_YARDSATTEMPTS_-_ROUGH	INT(11)
APPROACHES___150-175_YARDSFT_-_ROUGH	DECIMAL(65,30)
APPROACHES___150-175_YARDSATTEMPTS_-_ROUGH	INT(11)
APPROACHES___175-200_YARDSFT_-_ROUGH	DECIMAL(65,30)
APPROACHES___175-200_YARDSATTEMPTS_-_ROUGH	INT(11)
APPROACHES___>200_YARDSFT_-_ROUGH	DECIMAL(65,30)
APPROACHES___>200_YARDSATTEMPTS_-_ROUGH	INT(11)
TOTAL_HOLES_PLAYED	INT(11)
TOTAL_GREENS_IN_REGULATION	INT(11)
GIR_RANK	INT(11)
TOTAL_DISTANCEFT_PROX_TO_HOLE	DECIMAL(65,30)
#_OF_ATTEMPTS_PROX_TO_HOLE	INT(11)
PROXIMITY_TO_HOLE_RANK	VARCHAR(5)
FAIRWAY_PROXATTEMPTS	INT(11)
FAIRWAY_PROXDISTANCE_IN_FT	DECIMAL(65,30)
FAIRWAY_PROX_RANK	INT(11)
ROUGH_PROXATTEMPTS	INT(11)
ROUGH_PROXDISTANCE_IN_FT	DECIMAL(65,30)

ROUGH_PROX_RANK	INT(11)
LEFT_ROUGH_PROXATTEMPTS	INT(11)
LEFT_ROUGH_PROXDISTANCE_IN_FT	DECIMAL(65,30)
RIGHT_ROUGH_PROXATTEMPTS	INT(11)
RIGHT_ROUGH_PROXDISTANCE_IN_FT	DECIMAL(65,30)
GOING_FOR_GREENATTEMPTS	INT(11)
GOING_FOR_GREENNON-ATTEMPTS	INT(11)
GOING_FOR_THE_GREENSUCCESSES	INT(11)
SCRAMBLING_PAR_OR_BETTER	INT(11)
SCRAMBLING_MISSED_GIR	INT(11)
SCRAMBLING_RANK	INT(11)
SCRAMBLING_PROXIMITY_TOTAL_DISTANCE	DECIMAL(65,30)
SCRAMBLING_PROXIMITY_#_OF_SHOTS	INT(11)
SCRAMBLING_PROXIMITY_RANK	INT(11)
SCRAMBLING_FROM_THE_ROUGHSUCCESSES	INT(11)
SCRAMBLING_FROM_THE_ROUGHATTEMPTS	INT(11)
SCRAMBLING_FROM_THE_FRINGESUCCESSES	INT(11)
SCRAMBLING_FROM_THE_FRINGEATTEMPTS	INT(11)
SCRAMBLING___>30_YARDSSUCCESSES	INT(11)
SCRAMBLING___>30_YARDSATTEMPTS	INT(11)
SCRAMBLING___20-30_YARDSSUCCESSES	INT(11)
SCRAMBLING___20-30_YARDSATTEMPTS	INT(11)
SCRAMBLING___10-20_YARDSSUCCESSES	INT(11)
SCRAMBLING___10-20_YARDSATTEMPTS	INT(11)
SCRAMBLING___<_10_YARDSSUCCESSES	INT(11)
SCRAMBLING___<_10_YARDSATTEMPTS	INT(11)
SAND_SAVE_%#_SAVES	INT(11)
SAND_SAVE_%#_BUNKERS	INT(11)
SAND_SAVE_RANK	INT(11)
PROX_TO_HOLE_FROM_SANDTOTAL_DISTANCE	VARCHAR(12)
PROX_TO_HOLE_FROM_SAND#_OF_SHOTS	INT(11)
TOTAL_HOLE_OUTS	INT(11)
LONGEST_HOLE_OUTYARDS	INT(11)
OVERALL_PUTTING_AVG#_OF_PUTTS	INT(11)
PUTTING_AVGGIR_PUTTS	INT(11)
ONE-PUTT_%#_OF_ONE_PUTTS	INT(11)
3-PUTT_AVOIDTOTAL_3_PUTTS	INT(11)
APPROACH_PUTT_PERFORMANCEATTEMPTS	INT(11)
APPROACH_PUTT_PERFORMANCEFT	DECIMAL(65,30)
AVG_DISTANCE_OF_PUTTS_MADETOTAL_DISTANCE_OF_PUTTS	INT(11)
TOTAL_ROUNDS_PLAYED	INT(11)
PUTTING___3'ATTEMPTS	INT(11)
PUTTING___3'PUTTS_MADE	INT(11)
PUTTING___4'ATTEMPTS	INT(11)
PUTTING___4'PUTTS_MADE	INT(11)
PUTTING___5'ATTEMPTS	INT(11)
PUTTING___5'PUTTS_MADE	INT(11)
PUTTING___6'ATTEMPTS	INT(11)
PUTTING___6'PUTTS_MADE	INT(11)
PUTTING___7'ATTEMPTS	INT(11)
PUTTING___7'PUTTS_MADE	INT(11)
PUTTING___8'ATTEMPTS	INT(11)
PUTTING___8'PUTTS_MADE	INT(11)
PUTTING___9'ATTEMPTS	INT(11)
PUTTING___9'PUTTS_MADE	INT(11)
PUTTING___10'ATTEMPTS	INT(11)

PUTTING___10'PUTTS_MADE	INT(11)
PUTTING_INSIDE_5'_PUTTS_MADE	INT(11)
PUTTING_INSIDE_5'_ATTEMPTS	INT(11)
PUTTING_INSIDE_5_FEET_RANK	INT(11)
PUTTING___5'_-_10'_PUTTS_MADE	INT(11)
PUTTING___5'_-_10'_ATTEMPTS	VARCHAR(5)
PUTTING___5'_-_10'_RANK	INT(11)
PUTTING___4'-8'ATTEMPTS	INT(11)
PUTTING___4'-8'PUTTS_MADE	INT(11)
PUTTING___4'_-_8'_RANK	INT(11)
PUTTING-INSIDE_10'ATTEMPTS	INT(11)
PUTTING-INSIDE_10'PUTTS_MADE	INT(11)
PUTTING-INSIDE_10'RANK	INT(11)
PUTTING___10'-15'ATTEMPTS	INT(11)
PUTTING___10'-15'PUTTS_MADE	INT(11)
PUTTING___10'-15'RANK	INT(11)
PUTTING___15-20'ATTEMPTS	INT(11)
PUTTING___15'-20'PUTTS_MADE	INT(11)
PUTTING___15'-20'RANK	INT(11)
PUTTING___20'-25'ATTEMPTS	INT(11)
PUTTING___20'-25'PUTTS_MADE	INT(11)
PUTTING___20'-25'RANK	INT(11)
PUTTING___>25'ATTEMPTS	INT(11)
PUTTING___>25'PUTTS_MADE	INT(11)
PUTTING___>25'RANK	INT(11)
PUTTING___>_10'_PUTTS_MADE	INT(11)
PUTTING___>_10'_ATTEMPTS	INT(11)
PUTTING___>_10'_RANK	INT(11)
TOTAL_PUTTS_GAINED	DECIMAL(65,30)
TOTAL_ROUNDS_PLAYEDPUTTS_GAINED	INT(11)
PUTTS_GAINED_RANK	INT(11)

APPENDIX C – SUPPLEMENTARY RESULTS

Example of differences between actual predictions of Decision Tree & K-Nearest-Neighbour models for Experiment 2 (2 tournament aggregation). The reader is referred to Table 6.2 for details.

	Decision Tree Result		K-NN Results	
PLAYER_NAME	Probability for level Y of MISSED_CUT_IND	Prediction	Probability for level Y of MISSED_CUT_IND	Prediction
Adams, Blake	0.4270063	N	0.25	N
Allenby, Robert	0.4270063	N	0.416666667	Y
Appleby, Stuart	0.305305305	N	0.416666667	Y
Austin, Woody	0.4270063	Y	0.166666667	N
Baddeley, Aaron	0.4270063	N	0.5	Y
Bae, Sang-Moon	0.4270063	N	0.5	Y
Baird, Briny	0.4270063	Y	0.416666667	Y
Barnes, Ricky	0.4270063	Y	0.583333333	Y
Beljan, Charlie	0.2	N	0.25	N
Bowditch, Steven	0.305305305	N	0.25	N
Bradley, Keegan	0.4270063	N	0.333333333	N
Brown, Scott	0.4270063	N	0.166666667	N
Byrd, Jonathan	0.4270063	Y	0.5	Y
Cabrera, Angel	0.4270063	N	0.25	N
Chalmers, Greg	0.371546149	N	0.5	Y
Chappell, Kevin	0.4270063	N	0.5	Y
Choi, K.J.	0.4270063	N	0.166666667	N
Cink, Stewart	0.305305305	N	0.416666667	Y
Clarke, Darren	0.4270063	N	0.333333333	N
Colsaerts, Nicolas	0.371546149	N	0.416666667	Y
Compton, Erik	0.371546149	N	0.416666667	Y
Couples, Fred	0.371546149	N	0.416666667	Y
Crane, Ben	0.4270063	N	0.5	Y
Curtis, Ben	0.4270063	N	0.333333333	N
Davis, Brian	0.305305305	N	0.083333333	N
de Jonge, Brendon	0.4270063	Y	0.333333333	Y

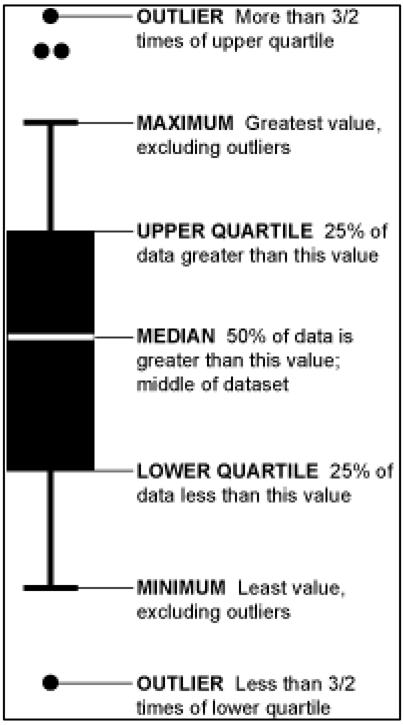
Table to reconcile the derived field “TID” (Tournament ID) with the name of the tournament.

TID	EVENT_NAME
2014010	Frys.com Open
2014020	Shriners Hospitals for Children Open
2014030	CIMB Classic
2014040	World Golf Championships-HSBC Champions
2014050	The McGladrey Classic
2014060	OHL Classic at Mayakoba
2014110	Hyundai Tournament of Champions
2014120	Sony Open in Hawaii
2014130	Humana Challenge in partnership with the Clinton Foundation
2014140	Farmers Insurance Open
2014150	Waste Management Phoenix Open
2014160	AT&T Pebble Beach National Pro-Am
2014170	Northern Trust Open
2014180	World Golf Championships-Accenture Match Play Championship
2014190	The Honda Classic
2014200	Puerto Rico Open presented by seepuertorico.com
2014210	World Golf Championships-Cadillac Championship
2014220	Valspar Championship
2014230	Arnold Palmer Invitational presented by MasterCard
2014240	Valero Texas Open
2014250	Shell Houston Open

APPENDIX D – BOX PLOT INTERPRETATION

Box plots are an effective way to visualise the descriptive statistics of a distribution range at a glance. They illustrate the upper and lower quartiles of the data, the median of the data, as well as the minimum and maximum values. Note that the minimum and maximum values must be within 1.5 times the value of the lower or upper quartiles. Any values outside the 1.5 limit are considered outliers.

All box plots used in this project have been created using SAS Enterprise Guide. In addition to the points illustrated in the figure below, Enterprise Guide also shows the mean as a diamond shape. This allows the user to visually observe the distance between the mean and the median.


OUTLIER More than 3/2 times of upper quartile
MAXIMUM Greatest value, excluding outliers
UPPER QUARTILE 25% of data greater than this value
MEDIAN 50% of data is greater than this value; middle of dataset
LOWER QUARTILE 25% of data less than this value
MINIMUM Least value, excluding outliers
OUTLIER Less than 3/2 times of lower quartile

Box Plot Explanation (Yau, 2008)

APPENDIX E – INTERVIEW TRANSCRIPT

The following is a partial transcript of an interview on the 4th of May 2014, between the author Brian Leahy (BL) and Brian Cosgrove (BC), resident PGA Pro at Killeen Golf Club, Kill, Co. Kildare, Ireland. Only sections of the interview which were cited throughout the dissertation are transcribed here. A full electronic recording of the interview is available.

Section 1

BL: Yeah, I've seen that mentioned in the research - like the course difficulty - and a lot of them say that the length of the course seems to be the main indicator of course difficulty.

BC: One of the main things, and I think when you're on about stats and what's more important you hear different things, like, and different coaches will have different views on what's the most important area like, for example a putting coach may be more inclined to say it's all about putting and they'll give you some research to suggest that. And then a short game coach might say it's all about your chipping and pitching and your distance wide shots or whatever, but like some of the newer research seems to be coming out is that it's...

BL: The long game...

BC: Yeah, the separation between the really good tour player and the average journeyman is from 175 to 250 yards because they can all really, you know, they can all hit a decent long ball, obviously some of them hit it a lot further than others - look at Bubba at the masters and that...

The difference between putting isn't massive - ok there still is a difference there but it's that difference between you know 175... 170 and 175 yards to 250, that's when the really good players are, they're hitting the greens a lot more precisely... highly...

BL: Yeah, that's something that I've seen as well that they're saying that it's not so much driving distance but driving accuracy and that's - what they mean by that is it's staying on the fairway from your tee shot so that your approach shot is on the fairway but then getting on to the...

BC: But driving accuracy also has an implication for driving distance because if you're hitting the fairway it's going to improve your distance anyway like as your flying it all the way and pitching it which... is gonna be at an optimum landing angle which will help it run out.

BL: It's your accuracy.... what they've been saying is you can go longer, but you're not going to be as accurate, whereas if you're a shorter hitter you're more likely to get on the fairway but by the same token the closer you are to the green for your second shot that the more accurate your approach shot is going to be for getting closer to the pin.

BC: Personally I find, give me a wedge in my hand, give me one of my Vokey wedges compared to say an 8 iron, I'm really thinking of getting my wedges in really tight to the pin, whereas my 8 iron, you're thinking of getting it in tight to the pin but you'll be happy with 15 feet like. The difference, even though it's only a few clubs, it's huge. That's 30 yards like.

BL: Is that psychological? Or it's just a fact that the wedge gives distance?

BC: Ah no...I think it's both. The head is tougher, there's less loft, it is longer and further away but I think the psychology of the wedge as well you build up your confidence, you definitely feel more confident with the wedges and you do it more often... so I think it's a combination of both.

Section 2

BL: Yeah it's the same culture, its the same food, essentially so you're not so bad.

No that's grand. You don't have any... Is there anything in your mind in terms of insight?

BC: I go through notions of looking at the stats, as in what's the key indicators but you know you hear different things. I think it varies so much from course to course. It's a lot more variable in the stats I would say than some places inland which have a very mundane climate, year in year out, whereas you can play Pebble Beach one day and it's calm and it's a great day play it the next day and it's....

BL: Yeah so any links course...

BC: Even though it's not a links course it very much has a links feel to it, so the variability there is a lot greater.

BL: That's fair enough, that's grand. I don't see anything... You've kind of covered that. Just ideas in terms of, if you have a bad first round followed by a second great round? Like you said it very much depends how you feel.

BC: I think the feeling, the psychology of it. But I think a lot of that depends on how a player is learning and adjusting as well. Like you'll hear someone saying about it takes a few years out on the circuit or on the tour to adjust to all that, like, and it's your rate of development and how quick you learn because everyone can play great golf but it's when you have the bad rounds how quick are you learning from it as opposed to just go on a downward spiral. It's like Jordan Spieth said. He said he learned more from his bad - from losing the Masters than winning it, like.

BL: So it's like McIlroy when he bounced back after... You could do something on that. It's like Spades, another fella, one time he was 7 strokes from it and then the one time I tried to predict it he missed the cut.

But yeah... I was looking at the previous 5 tournaments - I was looking at the previous 10 but I was thinking that's just too much of a variation to even get an idea of somebody if they are improving and getting there I think five is roughly - but do you think 5 might even be too much?

BC: Five is max I would say - that's nearly 20 rounds of golf like so when you think of all that can go on on the golf course in 20 rounds, you know with changes in your swing, changes mentally and all that like - again depending on the quality of your mind set but you know if you've a bad round - you know you often see commentators saying 'he's come off with a bogey that's going to, he'll be thinking of that all night' like - but you've 20 of those nights to put in after 20 rounds after 5 tournaments, which is 20 rounds so that can have a very powerful impact.

APPENDIX F – SAMPLE SQL CODE

```
/*GOLFER ANALYTICAL RECORD - AGGREGATE INFO*/
/*INCLUDES FORM SCORE*/
/*07/05/2014*/
/*CHANGED AVG TO SUM / TOT NO OF RNDS - NOT THE NUMBER OF TOURNAMENTS*/


/*AGGREGATED QUERY*/
SELECT TSF.PREDICT_TID
                ,CASE
                    WHEN  TRIM(ED2.FINISH_POSITIONTEXT)  =  'CUT'
THEN 'Y'
                ELSE 'N'
                END AS MISSED_CUT_IND /* Indicates if the player missed the
cut for the PREDICT_ID tournament*/
                #,ED. PLAYER_NUMBER
                ,ED.PLAYER_NAME
                ,MAG.AVG_GIR_PER_RND
                ,MAG.AVG_DIST_TO_HOLE_AFTER_APP
                ,MA_DRIVE_ACC
                ,ED2.FINISH_POSITIONNUMERIC AS PREDICT_FINISHPOS
                ,MAG.MA_FINISH_POS  AS  MOV_AVG_FINISH_POS  /*LAST
5 TRNS*/
                ,COUNT(ED.TOURNAMENT_NUMBER)                    AS
NO_OF_TOURNAMENTS_PLAYED
                ,
                SUM
                (
                CASE
                        WHEN TRIM(ED.FINISH_POSITIONTEXT) <>
'CUT'  THEN 1
                        ELSE 0
                    END
                )AS NO_OF_CUTS_MADE
                ,
                SUM
                (
                CASE
                    WHEN TRIM(ED.FINISH_POSITIONTEXT) = 'CUT' OR
ED.FINISH_POSITIONTEXT IS NULL THEN 1
                    ELSE 0
                END
                ) AS NO_OF_CUTS_MISSED
                ,FS.Moving_Avg  AS  FORM_SCORE  /*HOW  FAR  AWAY  (OR
NOT) A PLAYER HAS BEEN FROM THE CUT LINE IN THE LAST 5 TOURNAMENTS*/
                ,
                CASE
                    WHEN FRM.TID IS NULL THEN 0
                    ELSE 1
                END AS LAST_TRN_4TH_MELT
                ,
                SUM(
                    CASE
                        WHEN JMDE.TID IS NOT NULL THEN 1
                        ELSE 0
                    END
                    ) AS JUST_MADE_CUT_CNT
                ,
                SUM(
                    CASE
                        WHEN JMISS.TID IS NOT NULL THEN 1
                        ELSE 0
                    END
                    ) AS JUST_MISSED_CUT_CNT

                , SUM(CASE WHEN TRIM(ED.FINISH_POSITIONTEXT) =
'CUT' OR ED.FINISH_POSITIONTEXT IS NULL THEN 1  ELSE 0 END        )        /
COUNT(ED.TOURNAMENT_NUMBER)  AS CUTS_MISSED_RATIO
            /*  ,SUM(MONEY) AS TOT_MONEY_EARNED
                ,AVG(MONEY) AS AVG_MONEY_EARNED*/
                ,SUM(TOTAL_GREENS_IN_REGULATION)                    /
SUM(TOTAL_ROUNDS) AS TOT_GREENS_IN_REGULATION
                ,SUM(EAGLES)  / SUM(TOTAL_ROUNDS) AS AVG_EAGLES
                ,SUM(BIRDIES)  / SUM(TOTAL_ROUNDS) AS AVG_BIRDIES
                ,SUM(PARS)  / SUM(TOTAL_ROUNDS) AS AVG_PARS
                ,SUM(BOGEYS)  / SUM(TOTAL_ROUNDS) AS AVG_BOGEYS
                ,SUM(DOUBLES)      /      SUM(TOTAL_ROUNDS)      AS
AVG_DOUBLES
                ,SUM(OTHERS)  / SUM(TOTAL_ROUNDS) AS AVG_OTHERS
                ,SUM(TOTAL_GREENS_IN_REGULATION)                    /
SUM(TOTAL_ROUNDS) AS AVG_TOTAL_GREENS_IN_REGULATION
                ,SUM("DRIVING_ACC._%FAIRWAYS_HIT")                  /
SUM(TOTAL_ROUNDS) AS AVG_DRV_ACC
                /*RANKS*/
                ,RANK.AVG_STROKE_AVERAGE_RANK
                ,RANK.AVG_EAGLES_RANK
                ,RANK.AVG_BIRDIES_RANK
                ,RANK.AVG_BOGEYS_RANK
                ,RANK.AVG_BOGEY_AVOIDANCE_RANK
                ,RANK.AVG_GIR_RANK
                ,RANK.AVG_SCRAMBLING_RANK
                ,RANK.AVG_PUTTS_GAINED_RANK
                ,RANK.AVG_SAND_SAVE_RANK
                ,SUM("Overall_Putting_Avg#_of_Putts") AS NO_OF_PUTTS
                ,SUM("Overall_Putting_Avg#_of_Putts")               /
SUM(TOTAL_ROUNDS)AS AVG_NO_OF_PUTTS
                ,SUM("One-Putt_%#_of_One_Putts")  AS NO_OF_ONE_PUTTS
                ,SUM("3-Putt_AvoidTotal_3_Putts")  AS  NO_OF_THREE_PUTTS
/* 3 OR MORE PUTTS ON A HOLE*/



FROM
```

```
SL_EVENT_DETAIL AS ED
/*IF RESTRICITING ON DATES - THIS WILL NEED TO BE BROUGHT INTO THE RANK TEMP
TABLE TOO*/

INNER JOIN
TRN_START_FINISH AS TSF
ON ED.PLAYER_NUMBER = TSF.PLAYER_NUMBER
#AND ED.TID = TSF.PREDICT_TID

INNER JOIN
(
SELECT TID
                                        ,PLAYER_NUMBER
                                        ,FINISH_POSITIONTEXT
                                        ,FINISH_POSITIONNUMERIC
FROM
SL_EVENT_DETAIL
) AS ED2
ON TSF.PREDICT_TID = ED2.TID
AND TSF.PLAYER_NUMBER = ED2.PLAYER_NUMBER

LEFT JOIN
(
/* RANKS*/
          SELECT ED.PLAYER_NUMBER,TSF.PREDICT_TID
          ,AVG(STROKE_AVERAGE_RANK) AS AVG_STROKE_AVERAGE_RANK
          ,AVG(EAGLES_RANK) AS AVG_EAGLES_RANK
          ,AVG(BIRDIES_RANK) AS AVG_BIRDIES_RANK
          ,AVG(BOGEYS_RANK) AS AVG_BOGEYS_RANK
          ,AVG(BOGEY_AVOIDANCE_RANK) AS AVG_BOGEY_AVOIDANCE_RANK
          ,AVG(GIR_RANK) AS AVG_GIR_RANK
          ,AVG(SCRAMBLING_RANK) AS AVG_SCRAMBLING_RANK
          ,AVG(PUTTS_GAINED_RANK) AS AVG_PUTTS_GAINED_RANK
          ,AVG(SAND_SAVE_RANK) AS AVG_SAND_SAVE_RANK
          FROM
          SL_EVENT_DETAIL AS ED

          INNER JOIN
          TRN_START_FINISH AS TSF
          ON ED.PLAYER_NUMBER = TSF.PLAYER_NUMBER

          WHERE /*ED.PLAYER_NUMBER = 20070
          AND*/ ED.FINISH_POSITIONNUMERIC <> 999
          AND TRIM(ED.FINISH_POSITIONTEXT) NOT IN ('DNS','DQ','W/D')
          AND TRIM(ED.FINISH_POSITIONTEXT) IS NOT NULL
          AND BIRDIES_RANK <> 999
          AND ED.TID BETWEEN TSF.START_TID AND TSF.END_TID
          GROUP BY 1,2
)
AS RANK
ON TSF.PLAYER_NUMBER = RANK.PLAYER_NUMBER
AND TSF.PREDICT_TID = RANK.PREDICT_TID
```

```
INNER JOIN
FS_MOVE_AVG3 AS FS
ON TSF.PREDICT_TID = FS.PREDICT_TID
AND TSF.PLAYER_NUMBER = FS.PLAYER_NUMBER

/*DETERMINE IF THE PLAYER HAD A BAD 4TH ROUND IN THE LAST TOURNAMENT*/
LEFT JOIN
FOURTH_RND_MELTDOWN AS FRM
ON TSF.END_TID = FRM.TID
AND TSF.PLAYER_NUMBER = FRM.PLAYER_NUMBER


LEFT JOIN
(
        /*who just made the cut?*/
        SELECT BSE1.TID,BSE1.PLAYER_NUMBER
        FROM
        sl_event_detail AS BSE1

        INNER JOIN
        (
                SELECT TID,MAX(ROUND_2_POS) AS MAX_ROUND_2_POS
                FROM
                sl_event_detail
                WHERE FINISH_POSITIONNUMERIC <> 999
                GROUP BY 1
        )
        AS BSE2
        ON BSE1.TID = BSE2.TID
        AND BSE1.ROUND_2_POS = BSE2.MAX_ROUND_2_POS
)
AS JMDE
ON ED.PLAYER_NUMBER = JMDE.PLAYER_NUMBER
AND  ED.TID = JMDE.TID

LEFT JOIN
(
/*WHO JUST MISSED THE CUT?*/
        SELECT  BSE1.TID,BSE1.PLAYER_NUMBER
        FROM
        sl_event_detail AS BSE1

        INNER JOIN
        (
                SELECT TID,MIN(ROUND_2_POS) AS MAX_ROUND_2_POS
                FROM
                sl_event_detail
                WHERE TRIM(FINISH_POSITIONTEXT) = 'CUT'
                GROUP BY 1
        )
        AS BSE2
        ON BSE1.TID = BSE2.TID
        AND BSE1.ROUND_2_POS = BSE2.MAX_ROUND_2_POS
```

```
)
JMISS
ON ED.PLAYER_NUMBER = JMISS.PLAYER_NUMBER
AND  ED.TID = JMISS.TID

LEFT JOIN
(
        /*MOVING AVERAGE GIR*/
        SELECT TSF.PREDICT_TID
                                        ,TSF.PLAYER_NUMBER
                                        ,SUM(TOTAL_GREENS_IN_REGULATION)            /
SUM(TOTAL_ROUNDS) AS AVG_GIR_PER_RND

        ,CAST(SUM(TOTAL_DISTANCEFT_PROX_TO_HOLE)                                    /
SUM("#_OF_ATTEMPTS_PROX_TO_HOLE")       AS       DECIMAL       (13,2))        AS
AVG_DIST_TO_HOLE_AFTER_APP
                                        ,SUM("DRIVING_ACC._%FAIRWAYS_HIT")          /
SUM("DRIVING_ACC._%POSSIBLE_FAIRWAYS") AS MA_DRIVE_ACC
                                        ,
                                        SUM(
                                                                CASE
                                                                        WHEN
FINISH_POSITIONNUMERIC = 999 THEN 80
                                                                        ELSE
FINISH_POSITIONNUMERIC
                                                                END
                                                        ) / 5 AS MA_FINISH_POS
        FROM
        TRN_START_FINISH AS TSF

        INNER JOIN
        SL_EVENT_DETAIL AS SED
        ON TSF.PLAYER_NUMBER = SED.PLAYER_NUMBER
        AND SED.TID BETWEEN TSF.START_TID AND TSF.END_TID
        #WHERE TSF.PLAYER_NUMBER = 28237
        #AND TSF.PREDICT_TID = 2011260
        GROUP BY 1,2
)
MAG
ON ED2.PLAYER_NUMBER = MAG.PLAYER_NUMBER
AND ED2.TID = MAG.PREDICT_TID

WHERE /*ED.PLAYER_NUMBER = 20766
AND*/ TRIM(ED.FINISH_POSITIONTEXT) NOT IN ('DNS','DQ','W/D')
AND TRIM(ED.FINISH_POSITIONTEXT) IS NOT NULL
AND ED.TID BETWEEN TSF.START_TID AND TSF.END_TID
AND TSF.PREDICT_TID = 2014170
/*ONLY BRING BACK DETAILS FOR THE GOLFERS WHO MISSED THE CUT AT LEAST 40%
OF THE TIME*/
/*AND PLAYER_NAME IN
(
        SELECT PLAYER_NAME
```

```
            FROM
            CUT_DETAILS_2013
            WHERE CUTS_MISSED_RATIO >=0.4
)*/
GROUP BY 1,2,3,4,5,6,7,8,FS.Moving_Avg,LAST_TRN_4TH_MELT
                                                ,RANK.AVG_STROKE_AVERAGE_RANK
                                                ,RANK.AVG_EAGLES_RANK
                                                ,RANK.AVG_BIRDIES_RANK
                                                ,RANK.AVG_BOGEYS_RANK
                                                ,RANK.AVG_BOGEY_AVOIDANCE_RANK
                                                ,RANK.AVG_GIR_RANK
                                                ,RANK.AVG_SCRAMBLING_RANK
                                                ,RANK.AVG_PUTTS_GAINED_RANK
                                                ,RANK.AVG_SAND_SAVE_RANK
                                                #,TOTAL_ROUNDS_PLAYED
;
/*PREDICT TOURNAMENT REFERENCE*/
/*SEE LAST 10 TOURNAMENTS.SQL TO UPDATE THIS*/

/*27/04/2014 - modified to bring back last  5 tournaments*/


SELECT *
FROM
LAST10
WHERE PLAYER_NUMBER = 23778
LIMIT 20
;

DROP TABLE TRN_START_FINISH
;

RENAME TABLE TRN_START_FINISH TO TRN_START_FINISH_290414 /* 2 trns*/
;


RENAME TABLE TRN_START_FINISH_290414 TO TRN_START_FINISH

CREATE TABLE TRN_START_FINISH
(
PLAYER_NUMBER INT
,PREDICT_TID INT
,START_TID INT
,END_TID INT
,PRIMARY KEY(PLAYER_NUMBER,PREDICT_TID)
)
;

DELETE FROM
TRN_START_FINISH
;

INSERT INTO TRN_START_FINISH
```

```
SELECT LT1.PLAYER_NUMBER
                                        ,LT1.TID AS PREDICT_TID
                                        ,LT3.TID AS START_TID
                                        ,LT2.TID AS END_TID
FROM
LAST10 AS LT1

LEFT JOIN
LAST10 AS LT2
ON LT1.PLAYER_NUMBER = LT2.PLAYER_NUMBER
AND LT1.RANK + 1 = LT2.RANK


LEFT JOIN
LAST10 AS LT3
ON LT1.PLAYER_NUMBER = LT3.PLAYER_NUMBER
#AND LT1.RANK + 5= LT3.RANK
AND LT1.RANK + 2 = LT3.RANK /* last 2 tournaments*/

#WHERE LT1.PLAYER_NUMBER = 20766

;

DELETE FROM TRN_START_FINISH
WHERE START_TID IS NULL
;
```