Using Machine Learning to Predict the Winning Score of Professional Golf Events on the PGA Tour

Oisín Wiseman
School of Computing
National College Of Ireland
Dublin, Ireland
Email: Oisin.Wiseman@student.ncirl.ie

Dr. Arghir-Nicolae Moldovan (Supervisor)
School of Computing
National College Of Ireland
Dublin, Ireland
Arghir.Moldovan@ncirl.ie

***Abstract*—Online sports betting is big business, particularly in-play betting. It is a competitive market with bookmakers constantly looking for new types of bets to attract customers. Betting on the winning score of a golf event is not something offered by bookmakers today. In this paper linear regression and feature selection are applied to uncover a novel set of features that can predict the winning score of a golf event once the first round is complete. Various machine learning algorithms are evaluated using these features to determine which ones can accurately predict the winning score. Using Azure Machine Learning, applications are built to predict the winning score of an event based on data from the first round. This research would be of interest to online bookmakers looking to gain a competitive advantage by adding to their portfolio of in-play bets. In addition, the outcomes of this paper would be beneficial to golfers who could adjust their tactics during the event based on the predicted score. The final applications are validated against completed events on the 2016 PGA Tour. The machine learning models outperform the best known method of predicting the winning score in existence today by 50% for predictions within one shot of the final score. The Bayesian linear regression algorithm is the most accurate predicting the exact score in 22% of the events and 67% to within 3 shots of the winning score.**

***Keywords*—Golf, PGA TOUR, ShotLink, Machine Learning.**

I. Introduction

Sports betting is a growing industry, with the gross gambling yield from the sports betting market forecasted to exceed $70 billion per year by the end of 2016 [1]. The advance of the Internet and new technology has significantly changed the industry. The move to online gambling has forced bookmakers to adapt and offer new ways to bet. In-play betting, that is placing bets while an event is underway, has seen large growth in recent years accounting for up to 80% of turnover on all sports betting [2], [3].

Golf lends itself to in-play betting, given the length of time each event takes. A professional golf event is normally played over 4 rounds of 18 holes. There is an overnight break between each round so it allows bookmakers and punters time to assess and adjust their positions. As a result, new ways to bet on golf events are constantly being added. Where once a punter placed a bet on the outright winner before the event started and checked it at the end, there is now the ability to bet on each round throughout the event. Bookmakers now offer a wide variety of bets during a golf event. Some examples include, who will be leading at the end of each round, who will finish in the top 10, will there be a playoff, which player will finish highest from a given country etc. A typical golf event can have around 30 different types of bets available, with a number of 'specials' added for the major events. Bookmakers are constantly evaluating new types of in-play bets to add to their portfolio so they can attract new customers and gain a competitive advantage.

This paper discusses a potential new in-play type of bet, that is to bet on the winning score of a golf event after round 1 is complete. There is no accurate way of predicting the winning score in existence today, hence this type of bet is not available in the market. Occasionally pundits and commentators may offer predictions, however it seems like it is mainly guesswork or based on domain knowledge on the score from previous years. They are generally not accurate, as the example from [4] illustrates, the actual winning score for the 2016 US open was 276(-4), nine shots better than predicted by the experts in the article. In conversation with professional golfers, they suggested that a good rule-of-thumb method they use to predict the winning score is to double the leading score from round 1 and add 2 shots. This paper set out to determine the accuracy of this educated guess and investigate if applying statistical analysis and machine learning could improve on it.

To the best of our knowledge, there is no evidence of machine learning being applied on golf data to date. Using data supplied by the PGA TOUR [5] through their ShotLink [6] system, this paper sets out to identify a novel set of factors that could be used to predict the winning score once round 1 is complete. The hypothesis is that it is possible to predict the winning score of any PGA Tour event using data from round 1. Various machine learning algorithms specifically, 'Boosted Decision Tree', 'Bayesian Linear Regression', 'Decision Forest', 'Neural Network' and 'Linear Regression' are explored for accuracy of predictions. Bayesian linear regression and linear regression were the two top performing algorithms selected to build web applications using Azure Machine Learning. These applications were validated against 27 completed events in the 2016 season. The results show that for predictions within one shot of the actual final score, the machine learning models outperform the 'educated guess' by 50%. The Bayesian model performed best in predicting the exact winning score in 6 out of the 27 events.

Predicting the winning score would be of interest to both the betting industry and golfers participating in professional events. Bookmakers could use the findings to open up new

types of in-play bets to bring to market. Golfers would use this information to assess their current score and decide what tactics may be required to win the event. In-play predictions could also be used by sports broadcasters and media to enhance broadcasts for golf fans so it would appeal to a wide audience outside those interested in betting [7].

This paper is organised as follows. Section II discusses background on golf events, the PGA TOUR and ShotLink. Section III discusses the related literature. Section IV outlines the methodology applied in this research. Section V explains the results. Section VI provides the conclusions while section VII discusses areas for future research.

II. Background

A. Golf scoring and Events

The United States Golf Association defines par as: "the score that an expert player would be expected to make for a given hole" [8], so essentially if a hole is specified as a par-4 then a professional golfer would be expected to complete the hole in four strokes (a shot is referred to as a stroke in golf). There are 3 types of holes on a championship golf course. They are par-3, par-4 and par-5. It is distance in yards that determines the par for the hole so shorter holes are all par-3's with holes over 470 yards in length a par-5. Each course on the PGA Tour has a course par score defined for it, this is the sum of the par scores for each of the 18 holes in the course. The majority of courses on tour are par-72 consisting of four par-threes, ten par-fours, and four par-fives. There are events that deviate from this and are played on par-70 and par-71 courses. A golfer's score is always compared to the par score. If a course has a par of 72 and a golfer takes 75 strokes to complete the course, the reported score is +3, or "three-over-par". If a golfer takes 70 strokes, the reported score is -2, or "two-under-par". Par for a professional event is calculated by multiplying the par for the course by the number of rounds in that event. Therefore, on a par-72 golf course, par for a four-round tournament would be 288. In order to work out a players score in relation to par then subtract the event total from the total of the players 4 rounds, so if a players four round total for an event is 286 on a par 72 course then their event score is said to be -2 (286 - 288).

This paper focuses on stroke-play events that are played over 4 rounds. Typically, each event starts on a Thursday and finishes on the Sunday of each week. Each round consists of 18 holes and the number of strokes taken at each hole are combined and totalled to give the round score. At the end of the event, the four round scores are added together and the player who has taken the fewest strokes is deemed the winner. The majority of events on the PGA Tour start with 156 players in round 1. At the end of the second round the field is reduced to just the top 70 players and ties who go on to play the remaining two rounds and compete for the prize money on offer for that event. The process of cutting the field after round two is called the 'halfway cut' or more commonly just 'the cut' with the score that determines the top 70 players and ties referred to as the 'cut line'. The players on scores better than the cut line are said to have 'made the cut'. Every player that makes the cut will earn money for the event. How much they earn varies depending on their final position on the leaderboard. The winner will usually earn 18% of the overall money available for the event and reduces on a sliding scale as you drop down positions on the leaderboard [5]. The four major events, namely 'The Masters', 'The U.S. Open', 'The (British) Open Championship', and 'The PGA Championship' have the strongest fields of competitors, made up of the elite players from around the world hence the prize money on offer is much greater compared to a typical PGA Tour event.

B. The PGA TOUR

The PGA TOUR (Officially rendered as PGA TOUR) is the organisation with responsibility for running the main golf tournaments played by professional male golfers throughout the United States and North America. They organise a series of tournaments held on an almost weekly basis throughout the USA, this collection of tournaments is referred to as the 'PGA Tour'. While the season has been extended over the years with new tournaments added, on average there are 40 official PGA Tour events run by the PGA TOUR each season. The 2016 season started in October 2015 and will run through to the end of September 2016. Each event on the PGA Tour has a limited field that varies from 130 to 156 players that meet the specified eligibility requirements for that event. Most professional golfers who play full time on the PGA Tour play between 20 and 30 events on average per year [5].

C. ShotLink Research

The data being used for this paper has been made available by the PGA TOUR through their ShotLink platform. ShotLink is a platform for collecting data on every shot hit by every player on the PGA Tour in real-time. The vision of the system is to "Turn data into information, information into knowledge, and knowledge into entertainment" [6]. Each golf course is mapped prior to the event so a digital image of each hole is captured. In addition, static laser-guided shot tracking systems are installed on each hole that record how far each shot was hit and the distance left to the hole. Each match is followed by dedicated ShotLink volunteers with handheld devices to enter additional scoring data and other characteristics of each shot. Both the lasers and the handheld data are beamed back to central ShotLink servers on the course so all the data is updated and available real-time. The ShotLink system went live in 2004 and the dataset has grown into an incredibly rich dataset. Players and coaches scrutinise the data to identify areas of their game that need improvement and TV networks use it to enhance their sports broadcasts for golf fans.

In 2005, the PGA TOUR began sharing this data with academic institutions, establishing a formal process for academics to gain not-for-profit access to the data collected by ShotLink for experimentation and study. In 2012, the PGA TOUR introduced the "ShotLink Intelligence prize" [9] which offers academics the opportunity to compete by submitting research papers that best utilise ShotLink data in a new way. This competition has resulted in many new papers contributing to the overall body of knowledge. The PGA TOUR publishes these papers on their website [9]. Some of these papers will be discussed in section III which reviews the related work on the ShotLink data.

III. Literature Review

This section provides an overview of the related work carried out using the ShotLink dataset, golf performance and the use of machine learning in other sports.

A. ShotLink

The majority of the existing research on golf based on the ShotLink data has been very much exploratory and statistical based. However, there have been many novel and varied uses of the ShotLink data that will be covered in this section.

Professor Mark Broadie is recognised as the "godfather of golf analytics" [10], he invented the 'strokes gained' method of measuring golfers performance. He has published many papers of golf analytics over the years that are brought together in his 2014 book "Every Shot Counts" [11]. His background is in finance and as a result he uses mostly the mathematical techniques of simulation and dynamic programming [10] in his research. Broadie recognised that the existing statistics used to measure golfer performance were outdated and flawed. He invented the 'stokes gained' approach to address this, first discussed in papers published in 2008 [12] and 2012 [13].

The novel aspect of the 'strokes gained' approach was to establish a benchmark that all golfers could be compared against, for a golf pro this is the tour average (the average of all the players). Broadie in his book [11] was then able to compare tour pros and also compare professionals to amateur golfers and in this way determine the areas of the game that separates amateurs from average pros and average pros from the best pros in the world. Today the 'strokes gained' statistics are widely used by players to gain a detailed analysis of where their strengths and weaknesses lie. They are also widely quoted by media and TV broadcasts and give fans greater insight in to the players performance.

In addition to simulation, regression analysis is another technique that has been used to develop new stats. Sen [14] proposed using a regression model to predict annual player rankings based on previous tour earnings and average weighted scores. A new numerical metric was proposed called KCS (Key Criterion of Success), it was argued that this single statistic could capture the overall performance of golfers during a PGA season based on adjusted values for earnings and scoring average. While this study was novel in its use of predictive models its findings may be a little simplistic when compared to the 'strokes gained' metrics.

Other research examples that demonstrate the varied analysis offered by the ShotLink data include Fearing et al. [15] who built on Broadies early work and applied it to putting, using a 'Markov Model' to define a new putting metric, 'putts gained per round'. Yousefi & Swartz [16] looked to extend the putting metrics further by not only looking at distance but also accounting for the difficulty of the greens. To do this, they developed a novel spatial statistics model that used a Markov model for computations. Hickman & Metz [17] studied the impact of pressure on performance. They narrowed the research to focus only on the final putts on the final hole in each event. They then applied a regression model to estimate if the player made the put or not. Their findings suggest that there is definitive evidence of choking under pressure in golf events and as the financial stakes go up the performance goes down. Fried & Tauer [18] explored the relationship between age and ability. They concluded that while experience can be a factor in winning events that golfers peak around 36 years of age and after that the ability to perform under pressure diminishes.

The research referred to in this section demonstrates the value and richness of the ShotLink data. Players and coaches who may have been sceptical in the early days due to the credibility of the data are now seeing the proven potential offered by the research. New stats provide golf fans with real-time insights to a players performance during events. The research carried out to date has not focused on scoring or the relationship between each of the four rounds in an event. This will be the focus of this paper. The other novel approach that this paper will bring is the use of machine learning techniques to produce a web application to predict the winning score of a PGA Tour event.

To the best of our knowledge there is no definitive evidence of machine learning techniques being applied using the ShotLink data. There are examples from other sports where machine learning has been used to predict results. Looking at these examples from other sports will ensure that any lessons learnt can be considered in the context of golf for this paper.

B. Machine Learning in Other Sports

Predicting the outcome of soccer matches using machine learning has been the subject of a number of research papers. Machine learning was used in an attempt to predict the results of soccer matches in the Champions League over the course of a season [19]. This was a classification problem and they explored the most popular machine learning algorithms for classification. The authors set out to select the best set of features to be used with the top performing algorithms to ensure the most accurate predictions of the games possible. Its not clear exactly how the features were selected, it was largely down to domain knowledge of the authors rather than statistical analysis. Of all the algorithms tested, Artificial Neural Networks came out as the most accurate. The output of the research was a piece of software developed using WEKA [20] that resulted in around 60% accuracy in predicting the correct results. The dataset used in this research was small in that it only looked at 96 matches for one season and lacked historical data for previous seasons. Huang & Chang [21], researched the use of neural networks to predict the results of matches in the 2006 soccer World Cup. Again feature selection was based on domain knowledge of the sport. The research only explored neural networks and the results showed a 76.9% accuracy. However, the dataset was very small over just 13 matches. Drawn games were excluded as the neural network could only predict the winner or the loser.

Turning to basketball, Zimmermann et al. [22] discussed lessons learnt based on applying machine learning techniques to predict college basketball matches. An important finding from their research is that there seems to be "an upper limit to predictive quality" using machine leaning techniques. They state that "there seems to be a 'glass ceiling' of about 74% predictive accuracy that cannot be exceeded by machine learning or statistical techniques" [22]. The results of their

machine learning models were disappointing or at least did not improve on statistical based solutions already in place. However, another lesson they highlight is that selecting the right features "can make or break success". It is not only the machine learning models that are important, it is more to do with ensuring the right predictive features are selected. Given the sheer number of observations in the ShotLink data, feature selection is an area that needs to be addressed in this paper.

In both these examples WEKA [20] was used to build the machine learning models. In all the papers researched for this project there were none that used Azure Machine Learning [23]. A novel aspect of this paper will be to use Azure Machine Learning to develop a predictive Web application.

IV. Methodology

As this research is exploratory in nature, the Knowledge Discovery and Data mining (KDD) process was selected to guide this project. KDD defines the overall process of "extracting high-level knowledge from low-level data" [24]. Fig. 1 describes the KDD implementation used in this paper. This section discusses each of the steps in that process and how they were implemented. A more detailed end-to-end workflow covering all aspects of this papers methodology is outlined in Fig. 2.

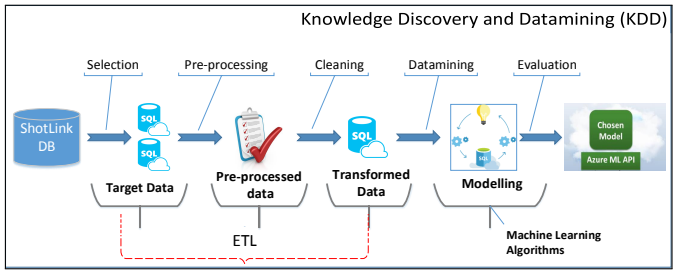

Fig. 1. KDD workflow for this research

A. ShotLink Dataset

There are four sets of data that ShotLink provides for offline analysis as listed in Table I. The data covers the twelve year period from when ShotLink was launched in 2004 through to the end of the 2015 season. Broadie refers to this period as "the ShotLink Era" [11].

One challenge of working with this data is the sheer volume. It contains detailed information on every shot hit by every golfer in every round since 2004. The dataset currently has 451 statistical categories, this leads to a large number of columns as outlined in Table I. When this level of granularity is multiplied by the twelve years the volume of data becomes big very quickly. As a result data selection can be challenging and time consuming.

The Event level dataset provides a comprehensive summary of every event played. It is an aggregation of the round and Hole level data and contains one row per tournament, per player. Examples of the type of data included within the event dataset are: 'Tournament Name', 'Course Name', 'Player Age', 'Round Scores', 'Finish Position' and 'Prize Money' etc. In addition there are specifics on individual shots such as 'Driving Distance', 'Putting Distances' and 'Approach Shot Accuracy'. The volume of data increases moving from event to round, hole and stroke level data. The round level dataset breaks down the event into each of the four rounds and provides additional data such as 'Course Name', 'Course Par' and 'Tee-off Time' for each player in each round of the event. The hole level dataset breaks the round down further into specifics on each of the 18 holes. The stroke level dataset is by far the largest as it contains the specifics on each individual stroke. There are two additional datasets, namely Radar Launch and Radar Trajectory that were not considered for this research. The focus of this paper is on the event itself so specific detail on shot locations are not required. The level of granularity in these datasets would not add additional insight to the research question.

TABLE I. ShotLink Datasets: Source Data

ShotLink Dataset Name	Total Columns	Total Rows
Event Level	190	68,807
Round Level	173	214,757
Hole Level	50	3,736,220
Stroke Level	38	12,212,043

While much of the data required for this research is available from the PGA TOUR website [5], it is only shown for the current season. The historical data required for this research is only available through ShotLink.

B. Pre-processing: Data Extraction and Cleaning

The ShotLink system is accessible through a secure website and all users are authenticated with a user-name and password supplied by the PGA TOUR. The system provides a GUI interface for golfers to query their data and review their performance statistics on an ad-hoc basis. To facilitate more in-depth research the system provides a mechanism to build custom bulk queries that can be packaged and exported as CSV files. Once these packages are extracted from ShotLink they can then be used with various analytical platforms and tools for further analysis. This is a once off operation and no further interaction with ShotLink is required once the data is extracted. All four datasets listed in Table I were extracted in full from ShotLink for the years 2004 to 2015. The next step in the extraction process was to create an Azure SQL [25] Database using the Azure Portal and then import these CSV files so the target dataset was brought back together on SQL Azure. The Extract, Transform and Load (ETL) workflow is detailed in Fig. 2.

Some pre-processing tasks were required to ensure a clean and consistent dataset to get the maximum benefit from the data mining stage. These tasks included:

1) Numeric fields with Null Values: Some numeric fields had NULL values. In the vast majority of these cases the NULL value just meant that the data did not exist for that player. An example is the 'Earnings' field. Not all players earn money at an event, so players who did not make the cut had a NULL value for earnings. This may skew analysis when using these numeric fields for calculations. In almost all instances of NULL values, replacing them with zero was the right solution to enable calculations and not skew the data. A series of R

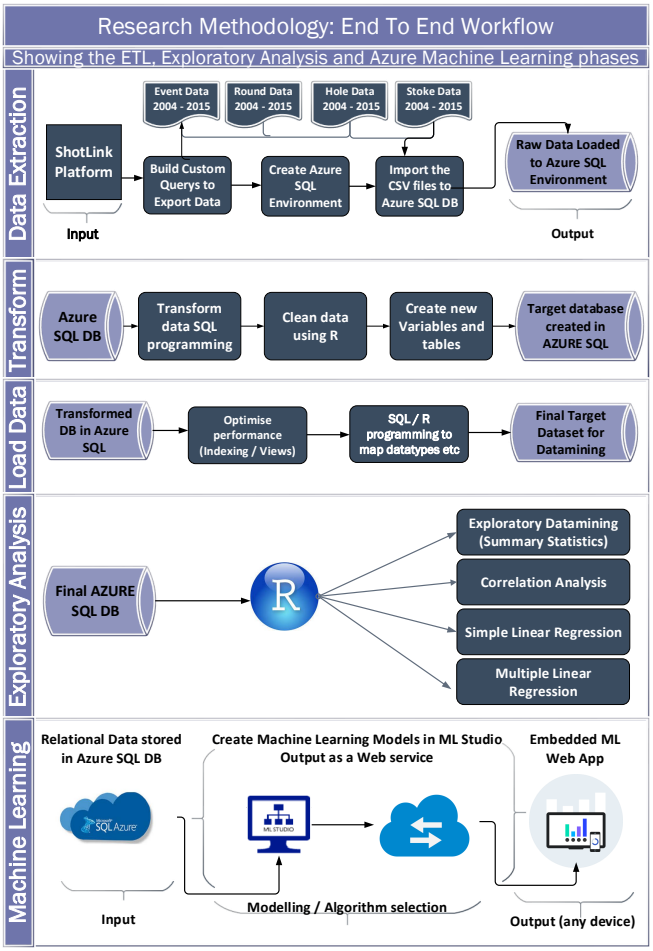

Fig. 2. End To End Workflow showing each phase of methodology

[26] scripts were written to perform the substitution of NULL values with zero.

2) Missing Values: In a small number of cases non-numeric fields were missing data. ShotLink already used a generic tag of '999' where data did not exist for non-numeric fields. Example for players who did not make the cut all text fields for round 3 and 4 were set to 999, in some cases this was missing and needed to be added particularly for older data. 'Mean replacement' was applied for some missing fields, specifically for missing players ages. To avoid removing the player and all associated data the blank age field was substituted with the mean player age for that event. This research does not make use of players age so this would not impact the results. some old events that were incomplete so these were removed completely which means that for the earlier years the number of events were smaller.

3) Event Clean-up: This paper only focuses on strokeplay events held over 4 rounds. A number of events on the PGA Tour are matchplay events, which uses different scoring mechanisms or are invitational pro-am events that do not follow the typical 4 round format. To ensure consistency and reduce noise these events were removed from the target dataset. In total there were approximately 15 events removed.

Finally, new and derived columns were created as defined in Table II. These will enable more in depth analysis during data mining. The 'Major' field easily identifies if an event is a major or not. The 'Event Final Score' and 'Rnd1 Leading Score' normalise the scoring in terms of par not strokes. The 'Rnd1 Avg Score' is tour average score for that round. 'Tour Average' the benchmark that all pros compare themselves against as discussed by Broadie [11].

TABLE II. NEWLY DERIVED AND CALCULATED FEATURES

Field Name	Data Type	Notes
Major	Binary	All events are categorised whether they are a major or not. This flag is as Major = 0 for regular events or Major = 1 for the 4 Major events.
Event Final Score	Numeric	A calculated field to record the winning score normalised to par. Calculated by working out the par for the event and subtracting the total strokes of the event winner $(coursePar \times 4) - MIN(totalStrokes)$
Rnd1 Lead Score	Numeric	The ShotLink Data only contains the score in terms of total strokes. This field stores the lowest score in comparison to par. $MIN(CoursePar - RND1TotalStroke)$
Rnd1 Avg Score	Numeric	Average Score of all players in the field for Round1. $(Total_Rnd1_Scores)/No_Of_Players)$

C. Data Transformation

The data required to address the research question is primarily at the Event Level. The transformation phase was focused on dimension reduction and creating a more consistent data set. As the Event data was so high level, some required information was not included such as specific data on the course the event was played on. The 'Course Name' and 'Course Par' were essential to calculate the fields listed in Table II. This course detail is part of the round dataset and was required in the final target dataset. SQL programming was required to isolate the necessary data and join the missing columns into one integrated view. The sheer volume of the SQL databases made these tasks difficult and error prone. As a result, stringent testing of the target data was required post transformation to ensure accuracy and consistency. The transformed view of the data was written back to Azure SQL to create the final target dataset described in Table III.

TABLE III. FINAL TARGET DATASET

Dataset Name	Total Columns	Total Rows
Event Level	40	69,359
Course Info	8	552

Dimension reduction was then applied to this database, given the extent of the data collected there was way more data than was required for this paper. The event data had many columns specific to players shots such as 'proximity of approach shots', 'putting performance' and 'proximity from the rough' which were not required. In total there were 154

columns removed that were focused on shot specific details that would not provide additional insights to the questions raised in this paper. As the focus of this research is at the event level, these columns were removed. This left a total of 40 columns and 69,359 rows in the final target database for this research. The final transformation work on this data was to clean-up the mapping of datatypes post the import into SQL. This work was completed using SQL Server Management Studio [27] and ensured that all fields were of the right datatype and had enough memory allocated. Table III describes the final target dataset. The course data for all the 515 events that was extracted from the hole level dataset is also available as a separate DB.

D. Data mining and Analysis

This section discusses the data mining techniques applied in this research, starting with exploratory data analysis right through to machine learning.

1) Exploratory Data Analysis: The initial phase of data mining was to apply Exploratory Data Analysis (EDA) techniques. EDA was introduced by John Tukey in the 1960's to better understand the structure and relationships within datasets [28]. Applying EDA in this research involved the plotting of different variables against each other and producing visualisations to help uncover deeper insights and patterns hidden within the data. While Tukey recommends the virtues of pen and paper for EDA, this research utilised R [26] and Microsoft Power BI Desktop [29] as the key technologies to explore the data.

Descriptive Statistics were run using R, this summarised the data through some key numbers such as the mean, median, max and min of each numeric field. It also gave counts of categorical fields broken into category. It helped identify variables that were not transformed correctly, some categorical variables such as 'major' or 'course par' needed to be transformed to factors instead of numeric. An important insight uncovered during EDA was that the leading round 1 scores were prone to outliers with players producing record scores. Based on this discovery it was decided to look at calculating the average round 1 score to account for the strength of the field and players producing one off incredible rounds. This led to the deeper analysis and graphs detailed in Simple Linear Regression Results.

2) Correlation Analysis: Correlation analysis was carried out across the dataset to investigate potential relationships between variables. Plotting the data can be helpful when determining if certain variables are related to each other. In addition to scatterplots, R code was written to generate correlation coefficients. 'Pearson's Product Moment' correlation coefficient is what this paper uses to measure the linear relationship between variables. Pearsons correlation coefficient is a measure of the strength and direction of the linear relationship between two variables, describing the direction and degree to which one variable is linearly related to another [30]. The Pearson coefficient metric runs on a scale from -1 to 1 where -1 indicates a strong negative correlation and 1 a strong positive correlation. Pearson's works well when variables are linear and normally distributed but is sensitive to outliers. Other measures such as the Spearman coefficient should be used if the data is skewed or non-linear.

The correlation analysis identified a set of independent variables that have significant relationship with the winning score. A correlation matrix was created using R to rank the variables with the strongest relationship. Correlation coefficients can only determine that there is a relationship between variables. It does not establish cause or determine whether a variable moves in response to another. Determining correlation is a first step, linear regression can add more certainty to the relationship.

3) Linear Regression: Correlation analysis and linear regression complement each other. While correlation signifies there is a potential relationship between the variables, regression analysis brings a degree of certainty to the relationship. It provides a mathematical method of determining the effect of the independent or predictor variable on the dependent variable. It is typically used to help prove or dispel working hypotheses. It was used in this research to select the variables that have the most impact on the dependent variable and to decide what variables should be discarded from the model. Two types of linear regression were applied in this paper, the first being simple linear regression which is simply comparing one variable against the other. The second is multiple linear regression, where multiple independent variables are explored in terms of their impact on the dependent variable. The dependent variable in this research is 'Winning Score'.

Based on the outcome of the correlation analysis multiple variables were fitted to regression models to measure the impact of the combination of the predictor variables on the winning score. Feature selection was then applied to add or remove predictor variables to measure the impact on the dependent variable. R^2 was the metric used to determine the accuracy of the regression model. R^2 also called the coefficient of determination is defined as "the proportion of variance explained by the regression model" [31], which is why it is useful as a measure of predictive accuracy. Note when working with multiple regression models the adjusted R^2 was used. The adjusted R^2 is a modified version of R^2 that has been adjusted for the number of predictor variables in the model. The adjusted R^2 increases only if the new variable improves the model more than would be expected by chance. It is always lower than the R^2 value. Upon completion of the linear regression analysis the best subset of variables to explain the variance around Winning Score were selected. The variables in the final model were then ranked in order of importance of the contribution they made to final model.

4) Machine Learning: Machine learning uses statistical algorithms to discover patterns within the data. It learns from these patterns so that it can automatically make decisions when confronted with new data. Microsoft's Azure Machine Learning [23] was the platform used to create the predictive models in this paper. Azure Machine Learning is a cloud-based predictive analytics service. The main reason it was selected for this research was the seamless integration with Azure SQL [25] and the ability to deploy web services directly as 'REST API's' that are consumed by the applications created as a result of this research.

Azure Machine Learning provides 5 applicable regression algorithms that were explored as part of this research they are:

- Linear Regression: Creates a linear regression model

using either the ordinary least squares method or the online gradient descent method. It is quick to train and very accurate if the data is fairly linear. The more difficult it is to separate the data by a straight line the less accurate it will be. This paper used the ordinary least squares method.

- Neural Network Regression: Typically associated with complex problems such as character recognition. They can be adapted to regression problems and work particularly well when the data is not strictly linear. A good option when traditional regression algorithms may not fit. Neural networks are associated with being very accurate [32].
- Bayesian Linear Regression: Probability based algorithm that is based around Bayes Theorem. Prior information about the parameters is combined with a likelihood function to generate estimates [32].
- Decision Forest Regression: Consists of an ensemble of decision trees. Each tree outputs a Gaussian distribution by way of prediction. An aggregation is performed over the ensemble of trees to find the distribution closest to the combined distribution of all trees in the model [32].
- Boosted Decision Tree Regression: Creates an ensemble of regression trees. Boosted implies that the tree is dependent on prior trees and learns by fitting the residual of the trees that preceded it. A boosted decision tree algorithm aims to improve accuracy but comes with a small risk of less coverage [32].

All of these algorithms were tested on target data. R^2 and the Mean Squared Error (MSE) were the metrics used to evaluate the accuracy of each algorithm. MSE is the average of absolute errors, lower error values typically mean the model is more accurate and the predictions closely match the actual values. The R^2 determines the accuracy of the predictions and how well the model fits the data.

The best performing algorithms were selected and deployed as web applications. These were validated against events in the 2016 that were completed during the research period for this paper. This was fresh new data not used to train and evaluate the models. The results will be discussed in the next section.

V. Results Analysis

A. Simple Linear Regression Results

Simple linear regression was used to explore the relationship between the 'Rnd1. Leading Score' and the final 'Winning Score' of the event. A simple linear regression model was fitted that consisted of the dependent variable 'Winning Score' and the predictor or independent variable 'Rnd1 Leading Score' covering all 515 events in the 'ShotLink era'. These were plotted in a simple scatterplot as can be seen in Fig. 3.

It is clear from looking at Fig. 3 that there is a linear relationship, the higher the 'Rnd1 Leading Score' the higher the 'Winning Score' of the event is. The results in Table IV show that a significant regression equation was found (R^2 = 0.3686, F(1,513) = 299.4, p <.000). The R^2 indicates that only 36.86% of the variation around the average Wining Score can

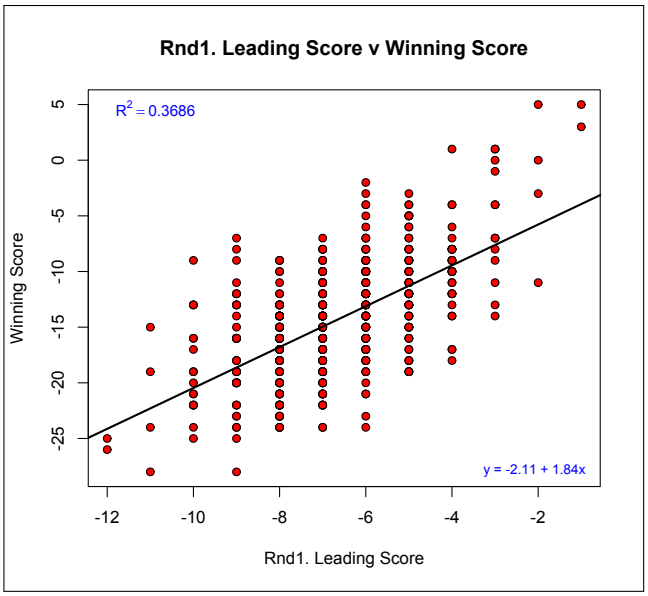

Fig. 3. Relationship between Rnd1 Leading Score and Winning Score for each event from 2004 to 2015. A total of 515 events plotted in total

TABLE IV. Linear Regression results for Wining Score verses Round 1 Leading Score

Predictor	Coefficient	Std. Error	t-value	p-value	
Intercept	-2.108	0.735	-2.868	0.004	**
Rnd1 Leading Score	1.835	0.106	17.304	<.000	***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
R^2 = 0.3686, Adj R^2 = 0.3673
F-statistic: 299.4 on 1 and 513 DF, p-value: <0.000

be explained by the 'Rnd1 Leading Score'. As there is just over a third of the variance explained by the leading score on round 1, this by itself would not make for an accurate prediction model.

While the 'Rnd1 leading score' is useful as a predictor variable there is too much variability. There are many potential factors that may account for the variability. Course difficulty, course conditions and strength of the field may be a few, but also on certain days any player is likely to have a perfect day when they outperform the field by a distance. Score comparison to the field is very important when it comes to scoring in golf and must to be factored into any prediction model.

To normalise for field strength it was decided to look at the average score for the field in round 1 (Rnd1 Avg Score), this is the sum of all the scores from round 1 divided by the number of players. This accounts for outliers where a player significantly outperforms the field and also minimises the variation due to the strength of the field and course difficulty.

A second simple linear regression was built to predict the Winning Score based on the 'Rnd1 Avg Score'. Fig. 4 shows the plot of this model, notice that regression line is steeper and the event winning scores are closer to the line. A look at the statistics in Table V reveal that a significant regression equation was found (R^2 = 0.5534, F(1,513) = 635.8, p < .000). This indicates that the 'Rnd1 Average Score' has more predictive

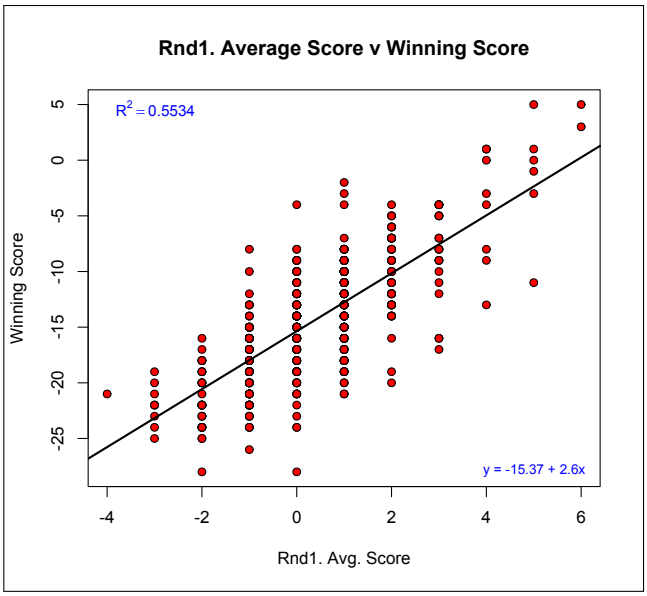

Fig. 4. Relationship between Rnd1 Average Score and Event Winning Score for each event from 2004 to 2015. A total of 515 events plotted in total

TABLE V. LINEAR REGRESSION RESULTS FOR WINNING SCORE VERSES ROUND 1 AVERAGE SCORE

Predictor	Coefficient	Std. Error	t-value	p-value	
Intercept	-15.365	0.159	-96.650	<0.000	***
Rnd1 Avg Score	2.602	0.103	25.210	<0.000	***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
R^2 = 0.5534, Adj R^2 = 0.5526
F-statistic: 635.8 on 1 and 513 DF, p-value: <0.000

power. The higher R^2 indicates that 55.34% of the variance of the Wining Score can be explained by the round 1 average score. In addition the lower standard error and high t-statistic indicates a highly significant relationship.

The results from the simple linear regression identified two variables that have a significant linear relationship with the winning score. These need to be considered when building a predictive machine learning model. To further strengthen the model more variables were required to optimise the model and reduce the variance. In the next section correlation analysis will be carried out to explore more potential variables that may have a relationship to the winning score.

B. Correlation Analysis Results

Correlation analysis was carried out to identify other potential variables that are related to the dependent variable 'Winning Score'. In total 10 variables were selected for the correlation analysis from the target set. Descriptive variables, primarily strings such as 'Event Name', 'Player Name' 'Event Year' for example were all removed as were player specific variables such as 'Finish Position' and 'Rankings'. This left 10 numeric event specific variables that could potentially be a factor in predicting the winning score of an event.

Table VI shows a matrix of the Pearson Product Moment correlation coefficients for all 10 variables. The sample correlation coefficient, denoted 'r' is listed for each pair of variables along with the significance. The results show that all the variables listed have a significant correlation with 'Winning Score'. The highest correlation coefficient is 0.90 and the lowest is 0.14. 'Rnd3 leading Score' is strongly related to Winning Score (r = 0.90, p <.001). On the other end 'Course Yardage' has a weaker relationship to Winning Score (r = 0.14, p <.001). The 'Rnd1 Avg Score' is also strongly related to 'Winning Score' (r = 0.74, p <0.001), which is consistent to what was discovered in the simple linear regression.

It is worth noting from that the two variables 'Rnd2 Leading Score' and 'Rnd3 leading score' are strongly correlated to the 'Winning Score'. This is to be expected as the certainty of predicting the winning score should increase after each round. The focus of this research is on the round 1 score so these will not be considered further for this paper. It is a potential area for future research as it will be required to re-set the in-play odds after each round.

Fig. 5 shows the final list of variables selected based on the correlation analysis that need to be explored further through multiple linear regression. It shows the Pearson Correlation coefficient for each of the 7 variables selected in ascending order. The variables selected for the linear regression analysis were 'Rnd1 Leading Score', 'Rnd1 Avg Score', 'Course Par', 'Major Event', 'Course Yardage' and 'Total Prizemoney'.

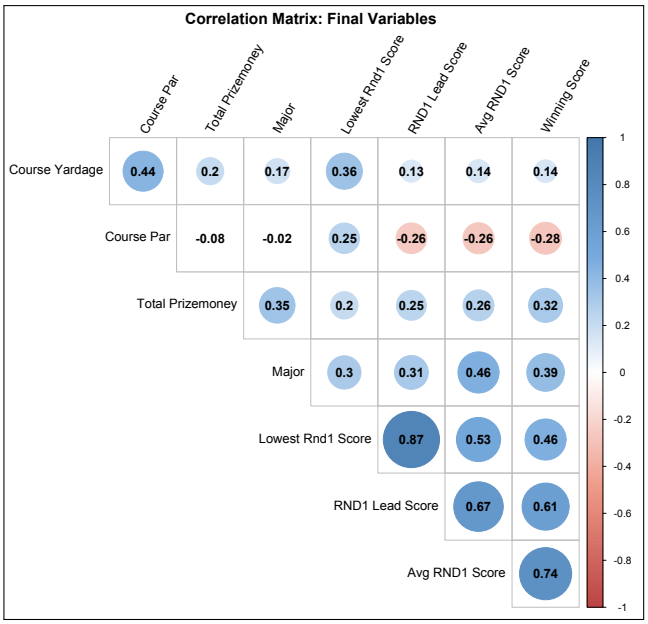

Fig. 5. The Pearson's coefficient score for each of the 7 variables selected as a result of correlation analysis

C. Multiple Linear Regression Results

Multiple Linear Regression allows for the consideration of multiple independent or predictor variables and how they account for variance in a single dependent variable, in this case 'Winning Score'. A multiple linear regression model was created to predict the winning score using the predictor variables selected as a result of the correlation analysis (see Fig. 5).

TABLE VI. PEARSON CORRELATION MATRIX FOR THE 10 VARIABLES SELECTED FOR CORRELATION ANALYSIS

	Variable	M	SD	1	2	3	4	5	6	7	8	9	10
1	Major	0.09	0.29	-									
2	Course Par	71.13	0.90	-0.02	-								
3	Course Yardage	7,241	200.39	0.17 ***	0.44 ***	-							
4	Total Prizemoney	5939936	1452104	0.35 ***	-0.08	0.20 ***	-						
5	Rnd1 Lowest Score	64.43	1.74	0.30 ***	0.25 ***	0.36 ***	0.20 ***	-					
6	Rnd1 Avg Score	0.36	1.50	0.46 ***	-0.26 ***	0.14 ***	0.26 ***	0.53 ***	-				
7	Rnd1 Lead Score	-6.71	1.73	0.31 ***	-0.26 ***	0.13 ***	0.25 ***	0.87 ***	0.67 ***	-			
8	Rnd2 Lead Score	-9.94	2.97	0.35 ***	-0.26 ***	0.14 ***	0.31 ***	0.58 ***	0.71 ***	0.71 ***	-		
9	Rnd3 Lead Score	-12.54	4.10	0.38 ***	-0.28 ***	0.13 ***	0.34 ***	0.5 ***	0.73 ***	0.65 ***	0.86 ***	-	
10	Winning Score	-14.43	5.24	0.39 ***	-0.28 ***	0.14 ***	0.32 ***	0.46 ***	0.74 ***	0.61 ***	0.79 ***	0.90 ***	-

Notes:
N = 515
For Major, 0 = No, 1 = Yes.
Signif. codes: *p <.05, **p <.01, ***p < .001
M = Mean, SD = Standard Deviation.

The results in Table VII show that a significant regression equation was found (Adj. R^2 = 0.5896, F(7,507) = 106.5, p <0.000), In addition 'Rnd1 Avg Score', 'Course Yardage' and 'Total Prizemoney' were significant predictors of Winning Score.

TABLE VII. INITIAL MULTIPLE LINEAR REGRESSION RESULTS WITH ALL 7 VARIABLES CONSIDERED

Predictor	**Coefficient**	**Std. Error**	**t-value**	**p-value**	
Intercept	19.300	12.410	1.555	0.121	
Major Event	0.613	0.607	1.010	0.313	
Course Par	-0.760	1.287	-0.591	0.555	***
Course Yardage	0.002	0.001	1.770	0.077	
Total Prizemoney	0.000	0.000	3.207	0.001	.
Rnd1 Lowest Score	0.132	1.278	0.103	0.918	
Rnd1 Lead Score	0.360	1.287	0.280	0.780	
Rnd1 Avg Score	1.944	0.146	13.287	<0.000	**

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
Multiple R^2: 0.5952, Adjusted R^2: 0.5896
F-statistic: 106.5 on 7 and 507 DF, p-value: <0.000

All variables in the linear model were checked for multi-collinearity. This occurs in a regression model when predictor variables are highly correlated to each other. Reviewing the Table VI indicates that 'Rnd1 Lead Score' is strongly correlated to 'Rnd1 Lowest Score' (r=.87, p < 0.001). Variance Inflation Factor (VIF) is one method of checking for collinearity using the 'VIF' function [33] in R. When the VIF statistic is greater than 10 it is considered a problem for multcollinearity [34]. The variable 'Rnd1 Lowest Score' fails this check with a VIF of 225. This is to be expected as 'Rnd1 Lowest Score' and 'Rnd1 Lead Score' are the same, the only difference is the unit of measurement. The 'Rnd1 Lowest Score' is the leading score in terms of strokes taken where 'Rnd1 Lead Score' is the strokes normalised to par. All other variables are below the VIF threshold of 10. 'Rnd1 Lowest Score' needs to be removed from the final model.

While the overall model is statistically significant, not all the variables are significant predictors of winning score. Further analysis was required to ensure all the variables were contributing to the model and identify the best subset of variables that fully explain the data. Stepwise regression was applied to identify the best subset of variables that represent the optimum set of predictors of the winning score. While stepwise regression methods have their critics [35] they provide a way for this research to identify the best subset of variables for the machine learning models. Specifically the method of selection applied in this paper was 'backward elimination stepwise regression'. The steps to manually carry out backward elimination step regression were:

- Start with all the predictors in the model
- Remove the predictor variable with highest p-value greater than 0.05
- Refit the model with the remaining variables and repeat
- Stop when all p-values are significant below 0.05

It is called backward elimination since it starts with all the predictors in the model and eliminates variables one by one. The predictor variables listed in Table VII was the starting list of variables. The first variable removed was 'Rnd1 Lowest Score' as it had the highest p-value. The model was refitted with the remaining variables, this time 'Major' had the highest p-value above 0.05 so it was removed and the model refitted. The remaining 5 variables were all significant below 0.05, no further action was required. R offers an alternative feature selection method that uses Akaikes Information Criterion (AIC) [36] as the metric for selection as opposed to p-values. AIC is a measure of the relative quality of a model and compares multiple models looking for the best performing one. It returns the model with the lowest AIC value. This method was applied in R using the 'step' function [37] for comparison purposes. The final results showed the exact same variables were selected as in the manual method.

The results of the final model can be seen in Table VIII. They show that a significant regression equation was found (Adj. R^2 = 0.5904, F(5,509) = 149.2, p < .000). In the final model all variables are significant predictors of Winning Score. The Adj. R^2 indicates that the variables in the final model now account for 59.04% of the variance in winning score.

In the next section the results of the machine learning experiments using the final variables selected from the linear analysis will be discussed.

D. Relative Importance

When reviewing the final multiple regression model, this paper looked at the relative importance of each of the 5

TABLE VIII. FINAL MULTIPLE LINEAR REGRESSION RESULTS, SHOWING THE FINAL 5 VARIABLES SELECTED FOR MACHINE LEARNING MODEL

Predictor	Coefficient	Std. Error	t-value	p-value	
Intercept	17.630	12.250	1.439	0.151	
Course Par	-0.608	0.201	-3.021	0.003	**
Course Yardage	0.002	0.001	1.782	0.075	.
Total Prizemoney	0.000	0.000	3.594	0.000	***
Rnd1 Lead Score	0.491	0.117	4.191	0.000	***
Avg Rnd1 Score	1.995	0.137	14.564	<0.000	***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
Multiple R^2: 0.5944, Adjusted R^2: 0.5904
F-statistic: 149.2 on 5 and 509 DF, p-value: <0.000

variables (see Table VIII). Relative importance is a method of quantifying what each of the variables are contributing to a multiple regression model. Johnson and Lebreton define relative importance as "the proportionate contribution each predictor makes to R^2, considering both its direct effect (i.e., its correlation with the criterion) and its effect when combined with the other variables in the regression equation" [38].

Using the relaimpo package [39] in R, a graphical representation of the relative importance of the variables was produced as per Fig. 6. The graph shows five metrics measuring relative importance. The main one of interest is the 'LMG' as it is the most widely recommended when decomposing R^2 is the objective [39]. The others are included for comparison purposes only. The results show that the 'Rnd1 Avg Score' variable contributes by far the most to the model at 57% of the R^2, with the 'Rnd1 Leading Score' accounting for 28%. This is in line with all the analysis we have done on this variable. It is clear that the 'Rnd1 Avg Score' is what makes the real impact to the model.

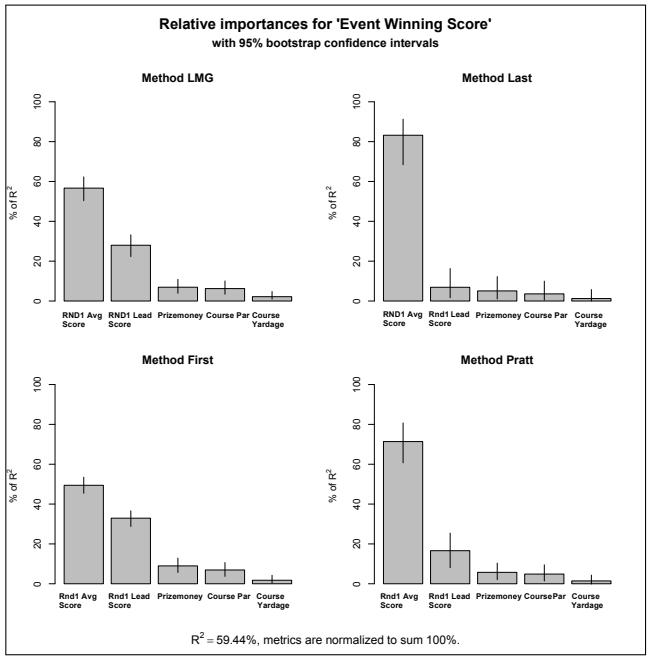

Fig. 6. Relative Importance of each of the 5 variables selected in the final regression model.

TABLE IX. COMPARING THE ACCURACY OF THE MACHINE LEARNING MODELS

Algorithm	R2	Mean Absolute Error	Root Mean Squared Error	Relative Squared Error
Boosted Decision Tree	0.4234	3.38	4.32	0.58
Neural Network	0.5723	2.86	3.72	0.43
Decision Forest	0.5774	2.93	3.70	0.42
Linear Regression	0.5800	2.87	3.69	0.42
Bayesian Linear	0.5867	2.85	3.66	0.41

E. Machine Learning Results

The output of this research is an Azure Machine Learning Application that will predict the winning score of a PGA Tour Event after round 1 is completed. This section discusses the testing and results used to determine the most accurate machine learning algorithm to build the final applications.

The Azure Machine Learning platform offers 5 regression algorithms to choose from. These were discussed in detail in section Section IV-D4. A machine learning algorithm makes predictions based on identifying patterns in historical data where the outcomes are already known. The predictions are then evaluated against the known result to determine the accuracy of the predictions. Five machine learning experiments were built, one for each of the regression algorithms. The target dataset was split randomly into a training and a test set, 70% was used to train each model with 30% held back for testing the accuracy of the predictions. The total number of events in the dataset is 515 so 360 events were used to train the models with the remaining 155 used to test the predictions. All 5 experiments were evaluated using this holdout method [40].

The main metric for measuring the accuracy of the predictions is R^2. Other metrics used are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Relative Squared Error (RSE). Lower error values indicate the model is more accurate and the predictions closely match the actual values. The results can be seen in Table IX, the algorithms are sorted in order from the least accurate to the most accurate.

The 'Bayesian Linear Regression' algorithm came out on top in terms of the highest R^2 and the lowest error values. Apart from the 'Boosted Decision Tree' algorithm the other 4 were quite close with the 'Linear Regression' and the 'Bayesian Linear regression' almost identical. It was decided that based on these results that two Azure applications would be built for comparison. One using the 'Bayesian Linear Regression' algorithm and the other the standard 'Linear Regression' algorithm.

Before deploying the models as web services they were re-trained using the entire dataset to ensure broader coverage. The parameters for each of the algorithms were optimised using the 'Tune Model Hyperparameters' [41] module in Azure ML and final evaluation was carried out using 10-fold cross validation [42]. The results of the optimised models can be seen in Table X. Note the R^2 values have slightly increased as a result of this optimisation to 0.5928 and 0.5944.

The fully optimised models were deployed live directly from Azure Machine Learning. These applications are live on the web today for anyone to use [43], [44].

TABLE X. FINAL ACCURACY RESULTS AFTER OPTIMISING THE MODELS

Algorithm Name	R2	MAE	RMSE	RSE
Bayesian Linear Regression	0.5928	2.64	3.34	0.41
Linear Regression	0.5944	2.64	3.34	0.41

F. Validation on 2016 Season Data

In order to test the true predictive power of the Azure applications, they were validated against the events in the 2016 season. There were 27 events from the 2016 season that were completed during the time frame of this research. The Azure applications were launched and the required variables entered for each event in turn as per Fig. 7. The 2016 dataset was new data that the was not exposed to the machine learning models during training.

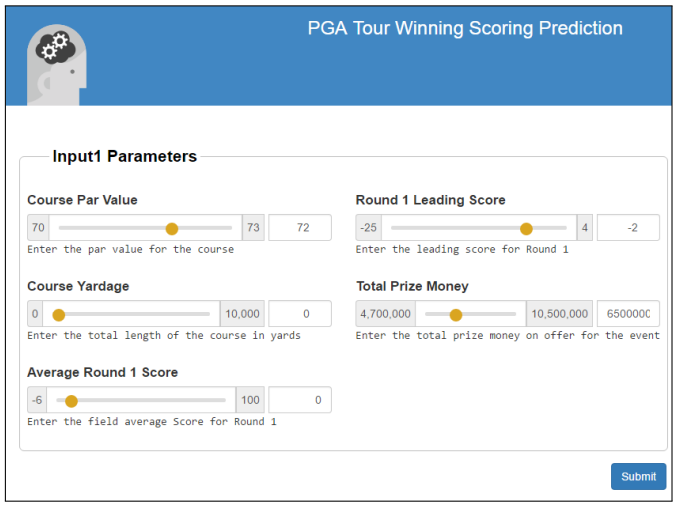

Fig. 7. The final Azure Web app used to validate the 2016 events. This App accepts as input the 5 features identified in this paper and returns the predicted winning score

The full results can be seen in Table XI. The 27 events are listed in sequential order from the start of the season. Each event has two predictions, one for the Linear Model and one for the Bayesian model. The 'guess' column refers to the rule-of-thumb method of predicting the event score that was identified from discussions with professional players. This 'best guess' is computed as per Eq. (1).

$$Guess = (Rnd1LeadScore \times 2) + 2 \quad (1)$$

The actual winning score for each event is also listed in the table.

Fig. 8 visualises the results split into five categories, exactly right, within one shot, within two shots, within 3 shots and over 3 shots. It shows that the both the Bayesian regression model predicted the exact winning score in 22% of the events, the Linear model followed with 19%, both of which were much more accurate than the 'best guess' which only predicted 7% exactly right. The Bayesian model gets 3 times as many exact predictions than the 'best guess' method that is used today. When including the predictions that are within one shot, the machine learning models perform 50% better than the 'best guess' which is a significant improvement. Fig. 9 shows the percentage of predictions as a cumulative chart through the 5 categories.

TABLE XI. PREDICTED RESULTS FOR EVENTS IN THE 2016 SEASON FOR BOTH MODELS. BEST GUESS IS CALCULATED AS PER EQ. (1)

Event Name	Winning Score	Best Guess	Linear Regression Prediction	Bayesian Regression Prediction
Frys.com Open	-15	-20	-19	-19
The RSM Classic	-22	-14	-17	-17
Sony Open in Hawaii	-20	-16	-17	-17
Farmers Insurance Open	-6	-14	-14	-14
Waste Management Phoenix Open	-14	-14	-14	-14
Northern Trust Open	-15	-18	-15	-15
The Honda Classic	-9	-12	-8	-9
Valspar Championship	-7	-10	-10	-9
Arnold Palmer Invitational	-17	-14	-17	-17
Shell Houston Open	-15	-18	-18	-17
Masters Tournament *	-5	-14	-9	-9
RBC Heritage	-9	-12	-12	-12
Wells Fargo Championship	-9	-16	-13	-13
The Players Championship	-15	-20	-17	-17
At&T Byron Nelson	-15	-16	-16	-17
DEAN & DELUCA Invitational	-17	-14	-12	-12
The Memorial Tournament	-15	-18	-17	-17
FedEx St. Jude Classic	-13	-12	-12	-12
U.S. Open *	-4	-10	-4	-4
Quicken Loans National	-17	-16	-18	-18
WGC Bridgestone Invitational	-6	-14	-9	-9
Barbesol championship	-18	-14	-18	-18
The (British) Open Championship *	-20	-18	-12	-12
RBC Canadian Open	-12	-14	-13	-13
PGA Championship *	-14	-12	-8	-8
Travelers Championship	-14	-14	-15	-15
John Deere Classic	-22	-16	-18	-18

Notes:
Events Listed in order from the start of the season until the week of August 8th
* Denotes event is a major championship event

While the 'best guess' may not do as well predicting the score exactly right it catches up when predictions within 3 shots are taken into account. This may be good enough for players to have an idea what they should aim for to have a chance of winning the event. However when betting on the outcome it would not be accurate enough. Overall looking at all predictions within 3 shots of the winning score, the machine learning models perform approximately 8% better than the 'best guess' prediction but crucially they are much more accurate to within 1 shot. Predictions over 3 shots from the winning score are too far out to be of any relevance.

The least accurate machine learning prediction was for 'The (British) Open Championship'. The predicted score from both applications were 8 shots out. The winning score of -20 was a record score for a major event. Two players avoided the worst of the weather and produced great golf to significantly outscore the field. It was a clear outlier and something that could not have been predicted. Predicting the winning score will always be prone to extraordinary events, where players over perform or weather dictates the scoring. However based on this small sample from the 2016 season the best performing machine learning model will predict the winning score to within one shot 41% of the time.

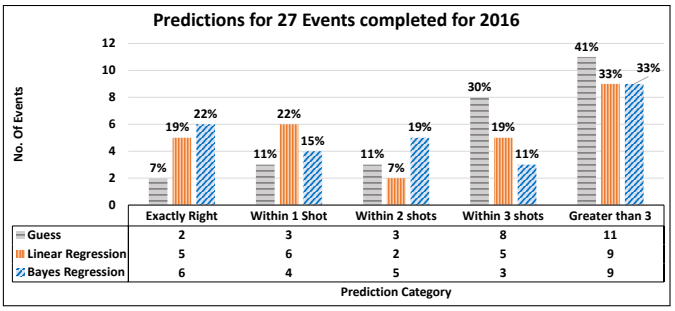

Fig. 8. The final results from testing the applications on the 2016 events. The table contains the actual number of predictions in each category with the percentages on the columns.

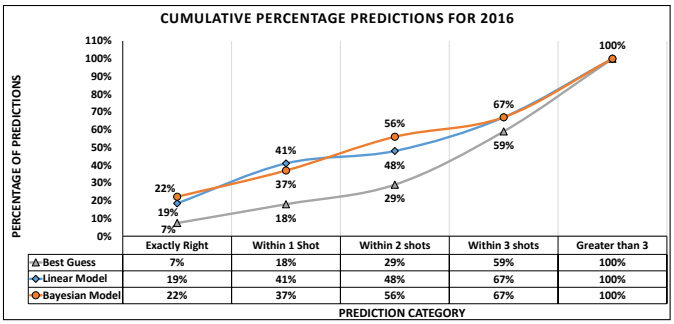

Fig. 9. A cumulative percentage chart showing the percentage of correct predictions through the different categories for each of the 3 models.

VI. Conclusion

This research has demonstrated that machine learning can be used to predict the winning score in a PGA Tour event. Feature selection is the key to success and this paper selected a novel set of features to ensure accurate predictions. The breakthrough in making the predictions more accurate came when introducing the average score of the field rather than the leading score after round 1. Features were selected both through domain knowledge and statistical analysis. All the available machine learning algorithms were tested for accuracy. The top two performing algorithms, 'Bayesian Linear Regression' and 'Linear Regression' were selected to produce two working web applications. When these were validated against the events in the current 2016 season, the results showed that the models could predict the winning score to within 3 shots 67% of the time. The results show that the machine learning models out-perform the 'best guess' when predicting the winning score by 50% for predictions to within one shot of the final score. This represents a significant improvement.

The objective of this research was not necessarily to predict the exact winning score for every event. It was more to add to the body of knowledge and apply machine learning to the ShotLink data. The methodology applied in this paper is something future research can build upon. There are opportunities for greater use of machine learning to be applied when working with he ShotLink dataset.

This research demonstrated that machine learning models can be of use to sports bookmakers to potentially offer a new in-play market to bet on the winning score. The results show that this is viable to open up an in-play market for the Winning Score. Professional golfers and coaches may also find this helpful when deciding tactics during an event. They can assess their own score against the predicted score and decide how to approach subsequent rounds to give them the best chance of winning the event. The Azure Machine Learning apps that have been produced can be used anywhere on any device and if a market does open up punters could use these to educate themselves on how to bet and increase their chances of winning.

VII. Future Research

This section discusses variations and new areas of further research that would enhance this project.

Predictions prior to the Event: In discussions with bookmakers while researching this paper, they highlighted that enticing customers in with an "opening bet" on the winning score prior to the start of an event would expedite adoption of the in-play offering. More research would be required to determine what features could be used to build a similar model without the details about round 1. Data is available on past events that would help but not all events are held at the same course each year.

Weather Conditions: The weather (and consequently course conditions) during an event has a big impact on the winning score. While weather data is becoming more accessible, the big challenge is the granularity required for golf events. Events are not held in the same locations each year and players can experience very different conditions to others in the field depending on what time they play. It is possible for some players to play the entire event with no rain and others play in very wet conditions depending on the draw. Modelling the weather for golf would be difficult but as more data becomes available it may be possible in future years.

Round 2 and 3 predictions: This research was focused on building a model based on round 1. Future research could expand this for round 2 and round 3. It would be expected that the predictions get easier and more certain after each round. The research should be extended on not only what the winning score may be but other in-play bets such as how by how many shots will the winner win by.

Acknowledgment

The author would like to thank:

My Supervisor **Dr. Arghir-Nicolae Moldovan**, for his guidance, support and valuable insights throughout this research.

Dr. Barry Haycock for providing valuable feedback and support when doubt set in!

My wife Noelle for her unyielding support and holding the fort while I ran away to the circus.

My 3 year daughter Allie for laughs, hugs and keeping things in perspective.

Thanks to my classmates for all their help and encouragement. Special thanks to my colleagues and management in the Microsoft European Development Centre in Dublin for their flexibility and support throughout the past two years.

REFERENCES

[1] Gbgc, "Global sports betting the state of play," 2013. [Online]. Available: http://www.gbgc.com/global-sports-betting-the-state-of-play/

[2] B. O'Halloran, "Boylesports founder bets on british expansion," 2013. [Online]. Available: http://www.irishtimes.com/business/retail-and-services/boylesports-founder-bets-on-british-expansion-1.1637718

[3] Bookmakersreview.com. (2013) Unibet generates 70 per cent of its sports betting turnover from live betting. https://www.bookmakersreview.com/bookmaker-newsflash/unibet-generates-70-cent-its-sports-betting-turnover-live-betting/49562. [Accessed: 2016-7-25].

[4] B. Porath, K. Robbins, E. Kay, T. Reaske, and M. Sandritter. (2016) U.s. open 2016 picks and predictions: Can jordan spieth go back-to-back at oakmont? http://www.sbnation.com/golf/2016/6/15/11943916/2016-us-open-golf-picks-predictions-oakmont-jordan-spieth. [Accessed: 2016-7-26].

[5] PGA TOUR, "official home of the pga tour," 2016. [Online]. Available: http://www.pgatour.com/

[6] shotlink.com, "About shotlink," 2014. [Online]. Available: http://www.shotlink.com/

[7] T. Barnett, D. O'Shaughnessy, and A. Bedford, "Predicting a tennis match in progress for sports multimedia," *OR insight*, vol. 24, no. 3, pp. 190–204, 2011.

[8] USGA, "Usga handicap system manual," 2016. [Online]. Available: http://www.usga.org/content/usga/home-page/Handicapping/handicap-manual.html#!rule-14370

[9] PGA TOUR, "The shotlink intelligence prize." [Online]. Available: http://www.pgatour.com/stats/shotlinkintelligence/prize.html

[10] S. Martin, "Interview with the godfather of golf analytics." [Online]. Available: http://www.pgatour.com/link-to-the-future/2015/07/22/mark-broadie.html

[11] M. Broadie, *Every Shot Counts: Using the Revolutionary Strokes Gained Approach to Improve Your Golf Performance and Strategy*. New York: Gotham, 2014.

[12] ——, "Assessing golfer performance using golfmetrics," in *Science and golf V: Proceedings of the 2008 world scientific congress of golf*, 2008, pp. 253–262.

[13] ——, "Assessing golfer performance on the pga tour," *Interfaces*, vol. 42, no. 2, pp. 146–165, 2012.

[14] K. C. Sen, "Mapping statistics to success on the pga tour: Insights from the use of a single metric," *Sport, Business and Management: An International Journal*, vol. 2, no. 1, pp. 39–50, 2012.

[15] D. Fearing, J. Acimovic, and S. C. Graves, "How to catch a tiger: Understanding putting performance on the pga tour," *Journal of Quantitative Analysis in Sports*, vol. 7, no. 1, 2011.

[16] K. Yousefi and T. B. Swartz, "Advanced putting metrics in golf," *Journal of Quantitative Analysis in Sports*, vol. 9, no. 3, pp. 239–248, 2013.

[17] D. C. Hickman and N. E. Metz, "The impact of pressure on performance: Evidence from the pga tour," *Journal of Economic Behavior & Organization*, vol. 116, pp. 319–330, 2015.

[18] H. O. Fried and L. W. Tauer, "The impact of age on the ability to perform under pressure: golfers on the pga tour," *Journal of Productivity Analysis*, vol. 35, no. 1, pp. 75–84, 2011.

[19] J. Hucaljuk and A. Rakipović, "Predicting football scores using machine learning techniques," in *MIPRO, 2011 Proceedings of the 34th International Convention*. IEEE, 2011, pp. 1623–1627.

[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten , "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[21] K.-Y. Huang and W.-L. Chang, "A neural network method for prediction of 2006 world cup football game," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2010, pp. 1–8.

[22] A. Zimmermann, S. Moorthy, and Z. Shi, "Predicting college basketball match outcomes using machine learning techniques: some results and lessons learned," *arXiv preprint arXiv:1310.3607*, 2013.

[23] microsoft.com, "Azure machine learning homepage," 2016. [Online]. Available: https://azure.microsoft.com/en-us/services/machine-learning/

[24] U. Feyyad, "Data mining and knowledge discovery: making sense out of data," *IEEE Expert*, vol. 11, no. 5, pp. 20–25, 1996.

[25] Microsoft.com, "Sql database cloud database as a service — microsoft azure," 2016. [Online]. Available: https://azure.microsoft.com/en-us/services/sql-database/

[26] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016. [Online]. Available: https://www.R-project.org/

[27] Microsoft.com, "Download sql server management studio 2016 (ssms)," 2016. [Online]. Available: https://msdn.microsoft.com/en-us/library/mt238290.aspx

[28] J. W. Tukey, *Exploratory data analysis*. Reading, Mass: Addison-Wesley, 1977.

[29] Microsoft.com, "Power BI Homepage." [Online]. Available: https://powerbi.microsoft.com/en-us/

[30] S.-D. Bolboaca and L. Jäntschi, "Pearson versus spearman, kendalls tau correlation analysis on structure-activity relationships of biologic active compounds," *Leonardo Journal of Sciences*, vol. 5, no. 9, pp. 179–200, 2006.

[31] N. J. Nagelkerke, "A note on a general definition of the coefficient of determination," *Biometrika*, vol. 78, no. 3, pp. 691–692, 1991.

[32] Microsoft.com, "More info on regression algorithms in azure machine learning." [Online]. Available: https://msdn.microsoft.com/en-us/library/azure/dn905922.aspx

[33] D. Lin, D. P. Foster, and L. H. Ungar, "Vif regression: a fast regression algorithm for large data," *Journal of the American Statistical Association*, vol. 106, no. 493, pp. 232–247, 2012.

[34] Y.-K. Tu, M. Kellett, V. Clerehugh, and M. S. Gilthorpe, "Problems of correlations between explanatory variables in multiple regression analyses in the dental literature," *British dental journal*, vol. 199, no. 7, pp. 457–461, 2005.

[35] L. L. Nathans, F. L. Oswald, and K. Nimon, "Interpreting multiple linear regression: A guidebook of variable importance," *Practical Assessment, Research & Evaluation*, vol. 17, no. 9, 2012.

[36] M. R. Symonds and A. Moussalli, "A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using akaikes information criterion," *Behavioral Ecology and Sociobiology*, vol. 65, no. 1, pp. 13–21, 2011.

[37] B. D. Ripley, "R: Choose a model by aic in a stepwise algorithm," 2015. [Online]. Available: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/step.html

[38] J. W. Johnson and J. M. LeBreton, "History and use of relative importance indices in organizational research," *Organizational Research Methods*, vol. 7, no. 3, pp. 238–257, 2004.

[39] U. Grömping *et al.*, "Relative importance for linear regression in r: the package relaimpo," *Journal of statistical software*, vol. 17, no. 1, pp. 1–27, 2006.

[40] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2, 1995, pp. 1137–1145.

[41] Microsoft.com, "Tune model hyperparameters in azure machine learning," 2016. [Online]. Available: https://msdn.microsoft.com/en-us/library/azure/dn905810.aspx

[42] ——, "Understanding cross-validation in azure machine learning," 2016. [Online]. Available: https://msdn.microsoft.com/en-us/library/azure/dn905852.aspx

[43] O. Wiseman, "Pga event winning score predictor using linear regression algoritthm," 2016. [Online]. Available: http://scorepredictionappfinal2.azurewebsites.net/

[44] ——, "Pga event winning score predictor using bayesian linear regression," 2016. [Online]. Available: http://scorepredictionappfinal1.azurewebsites.net/Default.aspx