**Physical Activity for Health / School of Psychological Sciences and Health**

**Electronic Submission Cover Sheet**

**This page must be included at the start of any electronic document submitted for assessment to Myplace, Turnitin, or via email.**

|  |  |
| --- | --- |
| **Class Code (e.g. B1701):** | B1706 |
| **Student Registration Number:** | 202362797 |
| **Tutor/Lecturer:** | Dr Allan Hewitt |
| **Supervisor:**  **(Yr 4 Dissertation only)** | Dr Allan Hewitt |
| **Title of Assignment:** | Dissertation |
| **Title of Project or Topic Area:**  **(Dissertation only)** | A Model Performance: Which classification model performs best at predicting the outcome of 5-foot putts in golf? |
| **Date:** | 27 June 2025 |
| **Word Count:** | 8,698 |

Plagiarism is defined as the submission or presentation of work in any form which is not one's own, **without acknowledgement of the sources**. Plagiarism includes inappropriate collaboration with others. Special cases of plagiarism can arise from a student using their own previous work (termed auto-plagiarism or self-plagiarism). Auto-plagiarism includes using work that has already been submitted for assessment at this University or for any other academic award.

In submitting this electronic assignment, I confirm that:

* It represents my own original work and that, where the work of others or my own previous work is used, this is clearly indicated to the reader and is referenced accurately;
* I have read and understood the guidance on plagiarism in the Course Handbook and in the University of Strathclyde’s guidance on plagiarism for students ([link](http://www.strath.ac.uk/media/ps/cs/gmap/plagiarism/plagiarism_student_booklet.pdf));
* I have clearly referenced, in both the assignment text **and** the bibliography and/or references section, all sources used in the assignment (including my own previous work). Note: the course referencing style is APA6;
* I have fully and accurately referenced (including page numbers), and used inverted commas for, all text **directly** **quoted** from books, journals, web pages etc.;
* I have provided sources for, and accurately referenced, all tables, figures, numerical data etc. that are not my own work (i.e. which I have not produced or collected myself);
* I have not made use of the work of any other student(s) past or present without acknowledgement. This includes any of my **own** work that has previously, or concurrently, been submitted for assessment, at this or any other educational institution;
* I have not sought or used the services of any professional agencies to produce this work;
* I understand that any false claim in respect of this work will result in disciplinary action in accordance with University regulations.



A Model Performance: Which classification model performs best at predicting the outcome of 5-foot putts in golf?

A dissertation submitted in fulfillment of the requirements for the degree of Master of Science (MSc)

Scott Dunn

202362797

Department of Psychological Sciences & Health

Supervisor: Dr Allan Hewitt

Word Count: 8,698

# Abstract

Golf is a sport that has been played as early as the 15th century. There have been many changes in that time. In recent years, the use of data and technological advancement such as Machine Learning (ML) are increasingly impacting the way golf is played and consumed.

Previous research shows 40%-45% of golf shots are putts meaning putting is a key aspect of the game. However, there has been comparatively limited use of advanced data analytics for optimising putting performance.

The aim of this project is to identify the best performing classification model that predicts the outcome of 5-foot putts.

Primary data was collected on 3,000 putts from a single participant using a Blast motion sensor. The outcomes of the putts were recorded by using a camcorder and a cross-reference approach was used to enable a supervised learning approach.

The data collected was partitioned into two categories, training and testing data consisting of 70/30% split respectively. Each model used the training data to establish patterns; predictions were made using these patterns on the test data. A comparison of the predictions and actual results were then reviewed to establish which model performed best.

Four classification models were investigated in this project: Logistic Regression, Support Vector Machines (SVM), Random Forest and XGBoost. These models were evaluated using a confusion matrix measuring: True Positive (TP), True Negative (TN), False Positive (FP) and false negative (FN). These evaluations were aggregated, it was found that XGBoost accuracy performed best with a score of 92.89%, which was closely followed by Random Forest which scored 92.78%.

This research provided a new insight how the use of machine learning can be applied in predicting the outcome of putts. It also highlighted a need for further research of the use of these models within the sport.

# Acknowledgements

After 24 years away from academia, studying the MSc Sport Data Analytics course at Strathclyde University was an honour and a privilege. The course has taught me a lot of new skills, perspectives and has developed me as a person. I would not have been able to complete this course and dissertation without the teachings and guidance of Dr Xanne Janssen and Dr Allan Hewitt.

Thank you, I wish you both a continued success.

Table of Contents

[Abstract i](#_Toc201859339)

[Acknowledgements ii](#_Toc201859340)

[List of Figures v](#_Toc201859341)

[List of Tables vi](#_Toc201859342)

[Chapter 1 – Introduction 1](#_Toc201859343)

[Background information 1](#_Toc201859344)

[Golf 1](#_Toc201859345)

[Machine Learning 1](#_Toc201859346)

[Models Reviewed 2](#_Toc201859347)

[Problem statement 2](#_Toc201859348)

[Research Objectives 2](#_Toc201859349)

[Research Question 3](#_Toc201859350)

[Defining “best” 3](#_Toc201859351)

[Report Structure 3](#_Toc201859352)

[Chapter 2 – Literature Review 4](#_Toc201859353)

[Introduction 4](#_Toc201859354)

[Use of Data in Golf 4](#_Toc201859355)

[Data Collection 6](#_Toc201859356)

[Machine Learning Approach 7](#_Toc201859357)

[Models 8](#_Toc201859358)

[Logistic Regression 8](#_Toc201859359)

[Support Vector Machines 8](#_Toc201859360)

[Random Forest 9](#_Toc201859361)

[XGBoost 10](#_Toc201859362)

[Model Evaluation 10](#_Toc201859363)

[Research Gap 11](#_Toc201859364)

[Chapter 3 – Methods 12](#_Toc201859365)

[Introduction 12](#_Toc201859366)

[Data Gathering 12](#_Toc201859367)

[Data Analysis 13](#_Toc201859368)

[Models 14](#_Toc201859369)

[Logistic Regression 14](#_Toc201859370)

[Support Vector Machines 14](#_Toc201859371)

[Random Forest 15](#_Toc201859372)

[XGBoost 15](#_Toc201859373)

[Model Evaluation 15](#_Toc201859374)

[Chapter 4 – Results 17](#_Toc201859375)

[Introduction 17](#_Toc201859376)

[Confusion Matrix 17](#_Toc201859377)

[True Positive 17](#_Toc201859378)

[True Negative 18](#_Toc201859379)

[False Positive 18](#_Toc201859380)

[False Negative 19](#_Toc201859381)

[Performance Matrix 19](#_Toc201859382)

[Accuracy 20](#_Toc201859383)

[Precision 20](#_Toc201859384)

[Recall 20](#_Toc201859385)

[F1 Score 20](#_Toc201859386)

[Chapter 5 – Discussion 21](#_Toc201859387)

[Introduction 21](#_Toc201859388)

[Key findings and insights 21](#_Toc201859389)

[Limitations 23](#_Toc201859390)

[Future Work 24](#_Toc201859391)

[Chapter 6 – Conclusion 26](#_Toc201859392)

[Background & Aim 26](#_Toc201859393)

[Methods 26](#_Toc201859394)

[Key findings 26](#_Toc201859395)

[Contribution to the field 27](#_Toc201859396)

[Future research 27](#_Toc201859397)

[Final remarks 27](#_Toc201859398)

[References 28](#_Toc201859399)

[Appendix 31](#_Toc201859400)

[Appendix A 31](#_Toc201859401)

[Appendix B 31](#_Toc201859402)

[Appendix C 32](#_Toc201859403)

# List of Figures

[Figure 1a - Driving range bay screens displaying shot data. Figure 1b – Television coverage using data visualizations at The Open Championship. 5](#_Toc201845880)

[Figure 2 - Logistic regression using the Sigmoid Function for a binary outcome. 8](#_Toc201845881)

[Figure 3a - SVM use of a hyperplane separating decision points. Figure 3b - Mis-classification possibilities using Linear (flat) Kernels. Figure 3c - Radial basis function curved decision boundary. 8](#_Toc201845882)

[Figure 4 - Random Forest using multiple trees and aggregation to achieve a final classification. 9](#_Toc201845883)

[Figure 5 - XGBoost using multiple trees in sequence where each new tree corrects the errors of previous trees. 10](#_Toc201845884)

[Figure 6a - Blast motion sensor on top of the putter grip. Figure 6b - Session segmentation used in data collection. 12](#_Toc201845885)

[Figure 7 - Logistic regression code. 14](#_Toc201845886)

[Figure 8 - Support Vector Machine code. 14](#_Toc201845887)

[Figure 9 - Random Forest code. 15](#_Toc201845888)

[Figure 10 - XGBoost Code. 15](#_Toc201845889)

[Figure 11 - True Positive results taken from the confusion matrix. 17](#_Toc201845890)

[Figure 12 - True Negative results taken from the confusion matrix. 18](#_Toc201845891)

[Figure 13 - False Positive results taken from the confusion matrix. 18](#_Toc201845892)

[Figure 14 - False Negative results taken from the confusion matrix. 19](#_Toc201845893)

# List of Tables

[Table 1 - Confusion Matrix breakdown 10](#_Toc201845894)

[Table 2 - Confusion Matrix results 17](#_Toc201845895)

[Table 3 - Performance Matrix results 19](#_Toc201845896)

Chapter 1 – Introduction

Background information

Golf

Golf is a club and ball sport which originated in St Andrews, Scotland around the 16th century[1]. Golf is an outdoor sport played on a course, which usually consists of 9 or 18 holes. Unlike many other sports, such as football and tennis, the playing area in golf has no fixed boundaries.

Each hole includes a tee box, fairway, rough, hazards and a green. A hole begins from a tee box where the ball can be placed on a tee peg for the first shot. The objective is to play the ball from the tee box into the hole (located on the green) in the least number of shots.

When the ball has been hit onto the green, a putter is used to roll the ball on the greens surface with the intention for the ball to fall into the hole, this action is known as putting. Putting is often referred to as a most important part of the game due to its impact of a player’s score. Previous research indicated putting represents 40%-45% of all shots played[2, 3]. Putting often determines the players score as it is frequently the final shot of a hole and the common phrase “drive for show, putt for doe” is often used by golfers[3].

Hasegawa et al stated tour professional golfers have a 75% success rate of holing 5-foot putts[4]. This is an important aspect of this research, as an achievable task with a high but not dominating success rate is required to make the predictions achievable. This means that if the success rate of the putts is either too high or too low, the variance in outcomes may be insufficient to produce meaningful variation in predictions.

Putting sensors are devices that record mathematical data for each putt, the data generated would be unattainable without such technology. This project aims to gather primary data and use a few *Machine Learning* methods to determine which method works best at predicting the outcome of putts. The length of the putts under review are 5-foot long due to its high percentage rate of putts holed and the impact this may have on the area of study.

Machine Learning

There are many options available when using machine learning, this study uses a supervised learning approach to make predictions on the outcome of putts. Supervised learning requires the model being provided with variables also known as features, and outputs also known as labels[5]. The machine is then trained on these details enabling patterns to be identified and predictions to be made from these patterns.

The machine learning approach used in this research project is supervised learning, due to the dataset containing both variables (features) and outputs(labels). The alternative to supervised learning is unsupervised learning, although this approach was considered, it was not required as we have the necessary outputs supervised learning requires.

This dataset was separated into two sections called training and testing data. The training data contains the outcome of these putts (which was covered in the supervised section above) which the classification model scans and learns patterns which will be used to predict testing data outcomes. This project predicts the outcome of 900 putts based on the patterns of 2,100 putts, making a 30/70% split.

The outcome which we are evaluating these models are based on a binary output (missed or holed). The model provides a comparison of predicted and actual outcome, known as a confusion matrix. This matrix provides four permutations, which are true positive, true negative, false positive and false negative.

Each model analyses the relationship between those permutations which determines what model works best, commonly known as performance matrix. The benefit of using evaluation metrics is that results can be analysed from multiple angles aiming to avoid misleading results.

Models Reviewed

There are several machine learning techniques available which have the capability to predict the outcome of a binary nature such as a make/miss outcome in putting. This study reviewed the performance of the following models:

* Logistic Regression
* Support Vector Machines (SVM)
* Random Forest
* XGBoost (Gradient boost)

Problem statement

Accurately predicting the outcome of putts using machine learning classification models appears to be an unexplored area. It is unclear which machine learning method provides the most accurate predictions of 5-foot putts.

Research Objectives

The aim of this study was to evaluate which machine learning classification model performs best at predicting the outcome of 5-foot putts.

* Gather quantitative primary data and compare the results of four classification models predictions on a binary outcome.
* Scrutinize the results these models produce to establish which model performs best at predicting the outcome of 5-foot putts.
* Compare each model on a like-for-like basis.
* Contribute to the field of machine learning by addressing an area which current literature is in short supply.
* Identify areas for further research and provide a solid benchmark for future study.

This study has the potential to benefit a wide audience, including golfers, performance analysts, coaches, club manufacturers (design) and club fitters at the point of sale. Its scope is broad, and the methods used are transferrable to other areas within the game and to other sports.

Research Question

Which Machine Learning classification model performs best at predicting the outcome (make/miss) of 5-foot putts in golf?

Defining “best”

For clarification, the term “best” in the research question refers to the highest accuracy score achieved in the performance matrices. The accuracy result shows the number of correct predictions as a percentage of the total number of predictions made.

This study evaluates each model’s performance through the analysis of a confusion matrix[6, 7]. A comprehensive approach was used to evaluate the results, the strengths and potential areas of weakness, such as an in-balance of a models results.

Report Structure

A literature review is detailed in chapter 2 providing a deep insight into previous research approaches. This section reviews previous research within golf and machine learning fields. Gaps in the literature available have been highlighted which supports the relevance and need of this study.

Chapter 3 provides a step-by-step guide to the methods used in this project, enabling these methods to be repeatable by another user. The coding used in this project is also available free on Github which is an open-source repository.

The results of this research are shown in chapter 4 and a discussion of these methods, results and study are shown in chapter 5. Chapter 6 is the closing chapter which concludes the findings of this research. The study ends with recommendations of areas that would potentially benefit with further research.

# Chapter 2 – Literature Review

Introduction

This chapter explores how data has been used in golf, from its early application to modern use. Highlighting the significant role data plays in the sport, from performance analysis to broadcasting professional events as entertainment. The project utilised primary data, a comprehensive outline of the methods used in the data collection process has been provided. The chapter closes with an overview of the four machine learning models was used in making these predictions, including a breakdown of how these models have been evaluated.

Use of Data in Golf

Golf is a sport with a rich history, and previous research indicates that it can be traced back through acts of parliament in the 15th century[1]. The Open Championship, one of the sport’s most prestigious tournaments, was first played in 1860 showing this sport has had a longstanding history at a professional level[8]. Newspapers like *The Scotsman*, *Glasgow Herald*, and *Aberdeen Journal* were among the first to document professional competitions, highlighting the early use of data to report the outcomes and scores of events[8].

In the 20th century, advancements in golf technology included yardage booksand course maps (also known as course guides). These course guides provided golfers with a bird’s-eye view of each hole, allowing them to visualize the layout and plan their shots, giving them tools to make informed course management decisions. This data driven shift fundamentally transformed the way the game was played.

In recent years wearable devices such as GPS watches became increasingly popular, providing golfers with real-time updates on distance to specific targets on the course[9-12]. These devices provide granular updates with each movement the golfer makes, giving precise data on distance, enabling golfers to make more data-driven decisions on every shot of their game.

Shot tracers became widely available, having the ability to measure metrics such as, clubhead speed, ball speed and many more[13]. These measurements provided accurate data on the distance and dispersion of each shot recorded. Every aspect of the shots detected provided golfers and coaches with insights into performance providing a golfer with the opportunity to make informed decisions.

The use if data has been transformed in the way golfers are coached, particularly with the arrival of shot tracking devices. Technologies such as this has enabled coached to adopt a more objective and data driven approach.

Where player performance can be analysed and quantified, in contrast to earlier methods, which relied on a coach’s experience and eye for talent. The data provided by these devices contain measurable insights with precise feedback that bespoke training packages can be created.

Driving ranges are places golfers can go to hit golf balls and practise for a small fee, usually purchasing either 50 or 100 balls per session. There are driving ranges that offer launch monitors in every bay which offers this service[14]. Allowing golfers to assess their practise session in real time providing up to date data allowing them to adjust as their practise, as shown in Figure 1a.

Indoor golf has allowed golfers to play virtual golf, where they can experience playing some of the most prestigious golf courses in an indoor environment. This invention has transformed the game and expanded the games reach, golf is now accessible in many new venues such as people’s homes, golf clubs and pubs[15, 16]. This highlights the impact technology and data have had on how golf is consumed by golfers and non-golfers alike.

A person standing at a golf course

AI-generated content may be incorrect.A map of golf course

AI-generated content may be incorrect.

Figure a - Driving range bay screens displaying shot data. Figure 1b – Television coverage using data visualizations at The Open Championship.

Data has played a significant role in how golf is broadcasted to the public. As mentioned earlier, the long history of golf was documented in newspapers, broadcasting the sport on television has become a competitive business[17]. As a result, the spectator benefits from the insights provided by data visualisations, as seen in Figure 1b.

Every shot can be analysed from multiple golfers, such as swing speed, ball speed and much more[13]. In major professional tournaments such as The Open, the volume of data is large, typically hosting 160 players each playing approximately 72 shots per round.

The role of data should not be understated; due to the depth and scale of insight it provides. Data has significantly influenced the design and manufacturing of golf equipment[18]. The design of the golf ball is a key example, where the dimples of the ball have been developed using a data driven approach[19].

A debate arose regarding the role of technology in the golf, where governing bodies such as the R&A (Royal & Ancient) and USGA (United States Golf Association) have discussed the possibility of limitations on equipment[20, 21].

As mentioned in the previous chapter, putting represents approximately 40%-45% of all shots in golf[2, 3], however current literature does not appear to reflect a similar focus on putting, current literature prioritises other aspects within the game, such as the longer shots with less focus on putting. There appears to be a lack of research regarding the prediction of the outcome of putts, emphasising a need for this research.

Data Collection

Careful consideration went into the planning of this project due to the problematic nature of collecting primary data[22-24]. By using primary data this research benefitting from maintaining full control over the collection process, ensuring the data remained both valid and reliable.

Multiple portable battery powered devices were used to collect the data for this research, including a putting sensor, a mobile phone, and a camcorder. Due to the battery power limitations of each device’s source, if one device failed it would compromise the data collected for that session[22]. Each session was consistent in length; this was shaped by the limitation of the power of the devices used.

Putting sensors record various putting variables, which includes rotation, face angle at impact, tempo and many other metrics[9, 13, 25, 26]. The blast motion sensor provides these measurements while being a wearable device[10]. An additional benefit is its affordability when compared to more expensive systems such as SAM Putt LAB[25].

Blast motion measures the mechanics of the putting strike, with data captured from the beginning of the strike through to the impact with the ball. This means its measurements are based on the initial setup of the stroke rather than alignment with a target.

In contrast, devices like SAM Putt lab require a calibration process which aligns the putter with a target, and measurements are made in reference to that target[25]. While this calibration facilitates the aimpoint of a putt, it limits the flexibility due to the requirement of recalibration every time the putt angle changes.

Blast motion was the preferred option in this study due to its flexibility. Which allowed a more varied and realistic representation of a typical practise session, where golfers attempt putts from different angles[27]. As opposed to repeating the same putt, leading the golfer to mastering a specific shot, potentially compromising the validity of the research by reducing the variability found in real-world practise conditions. The metrics collected by the blast motion sensor are consisted with previous research showing that the use of this monitor is suitable for a project of this nature[25]. The mobile phone is required to transfer the data from the sensor to the cloud.

It is important to note that the author of this paper is both the researcher and the golfer, meaning this project was completed as a sole venture. The use of a camcorder benefited both the golfing and research aspects of the project[28, 29].

The golfer could focus entirely on executing each putt as an independent event, without the need to record the outcome in real time[27]. The researcher had full confidence that the outcomes were being recorded accurately with an opportunity to review the footage multiple times if needed.

However, it should be acknowledged that a limitation of this study is its use of an intra operator approach[30], an inter operator approach may have added an improved layer of validity but would have required more than one individual to review the footage[31].

The r programming language is a well-established tool for data analysis and was the preferred option for this study due to its ease of use, online training resources, and open-source availability making it free to use[32, 33].

Greenbaum et al augured the use of data in sport is an elitist approach[34]. This study challenges that perspective, all equipment was sourced on a limited budget and all software used was open source.

The code and dataset have been uploaded to Github, providing anyone with the opportunity to replicate this study, see Appendix A. By doing this also means the code can be updated or modified by suggestions from the coding community, potentially making this project future proof[35, 36].

Machine Learning Approach

As discussed earlier, data was collected from multiple sources: the blast motion sensor captured the putting stroke (independent) variables, and the camcorder recorded the outcome of each putt (dependant variable)[5, 23]. This enabled the use of a supervised learning approach. Supervised learning is a machine learning method where an algorithm is trained to identify patterns and relationships between independent and dependant variables.

This is achieved by splitting the dataset into two tranches, known as the training data and the testing data. The patterns the machine learns from the training dataset are applied to test dataset to predict outcomes. Previous research has shown supervised learning is a widely used method and is suitable for this research[5]. The outcomes measured in this project is binary (make or miss), which is why a classification model is most appropriate.

Models

Logistic Regression

Logistic regression is a statistical method which can work with both continuous and categorical data[33, 37]. This method was chosen for this project as it works well as a classification model to predict the outcome of a binary dependent variable.

Logistic regression can work with several variables; however, it should be noted that the model is sensitive to overfitting. Overfitting occurs when the datasets contaminate training data with noise rather than identifying patterns. This has a negative impact on the learning within the training data, potentially resulting in less accurate predictions.

A graph with a line and a line

AI-generated content may be incorrect.

Figure - Logistic regression using the Sigmoid Function for a binary outcome.

The machine uses the odds of an event occurring and converting these odds into the probability of this event occurring[37]. This is achieved with the use of a sigmoid function, the sigmoid function converts the log odds into a probability estimation ranging between 0 & 1, in this instance 0 represents a missed putt and 1 represents a holed putt, as shown in figure 2.

Support Vector Machines

A diagram of a graph

AI-generated content may be incorrect.

Figure a - SVM use of a hyperplane separating decision points. Figure 3b - Mis-classification possibilities using Linear (flat) Kernels. Figure 3c - Radial basis function curved decision boundary.

The core concept of a Support vector machine is to create a decision boundary and a margin between decision points, often referred to as a hyperplane, as seen in Figure 3a. These boundaries aim to create the largest margin possible between decision points. The sample vectors are the data points which are located on the edge of the margin[33, 38].

Linear kernels use a decision boundary which is flat, this enables the machine to make prediction on a linear format. There is a limitation using this method due to potential inaccurate predictions when using high/multi-dimensional data, as shown in Figure 3b. It is not always possible to separate the categories of data points by using a straight line.

A 2-dimensional space could be measured using an x & y axis, whereas a 3-dimensional space could use x, y & z axis. Although higher dimensions are not easy to visualise, SVM works well in mapping multiple dimensions using polynomial kernels.

A polynomial kernel considers the interaction between the data points and measures their similarity between these points in multiple dimensions. The radial basis function (radial) uses a curved decision boundary; this is useful where the decision boundaries are multidimensional and therefore non-linear, as seen in Figure 3c.

Random Forest

Random Forest is a collection of decision trees which is commonly referred to as an ensemble algorithm[39-41]. In this case an ensemble algorithm combines multiple decision trees, the machine is instructed to the number of trees used and to the number of decision nodes within each tree.

A diagram of a tree

AI-generated content may be incorrect.

Figure - Random Forest using multiple trees and aggregation to achieve a final classification.

This process is referred to as bagging, which is a combination of both bootstrapping and aggregation of results, as seen in Figure 4. Bootstrapping is where a sample of the training dataset is taken for the machine to establish patterns, enabling predictions to be made on these patterns and aggregate the results of each tree. This aggregation determines the final classification.

XGBoost

Gradient boosting models such as XGBoost is also an ensemble method, unlike Random Forest there is no aggregation of results[36, 39, 41]. Each tree is trained sequentially and learns patterns from previous trees, as shown in Figure 5. For example, if 100 trees have been stipulated tree 50 would have benefitted from the learning of the previous 49 trees.

A diagram of a tree

AI-generated content may be incorrect.

Figure - XGBoost using multiple trees in sequence where each new tree corrects the errors of previous trees.

The benefit of this approach is that each subsequent tree attempts to fix the errors of previous trees to provide a higher level of accuracy. Although levels of accuracy are often high, it is a slower method to use and could be prone to overfitting.

An output is created once the algorithm has been through the last tree, this output is expressed as a probability figure between 0 and 1. Functioning in the same manner as logistic regression, where 0-0.49 is considered a negative result and 0.5 and over is as a positive result. Again, the threshold of a negative/positive result is subject to change, which is instructed by the researcher.

Model Evaluation

Confusion Matrix

Table - Confusion Matrix breakdown

A black and white text on a white background

AI-generated content may be incorrect.

Table 1 shows the outputs of this project were displayed in a confusion matrix, the benefit of using this approach is the details it provides beyond the overall accuracy of a model. It also provides an insight into an imbalance or bias of the results, as well as allowing to compute detailed measurements such as precision, recall and F1 scores. Previous research has shown this approach works well in a classification driven project[6, 7, 42].

Research Gap

Previous research shows that the methods used in this study have been applied and compared in the past[5, 33, 39, 41, 43]. However, while comparisons between these models have been explored the researcher was unable to identify any studies comparing these models to predict the outcome of putts in golf. This highlights a gap in literature currently available and supports the relevance of this research project.

# Chapter 3 – Methods

Introduction

This chapter outlines the processes applied in this research, including data collection and analysis. A breakdown of the code each models used and how each model can be used to evaluate its performance.

Data Gathering

There were no ethical concerns associated with this research, as the author served as both the researcher and the sole participating golfer. This study used a case study approach using quantitative data. The participant was a 44-year-old right-handed male golfer with over 30 years’ experience and a scratch handicap.

A close up of a golf club

AI-generated content may be incorrect.A golf course with a camera

AI-generated content may be incorrect.

Figure a - Blast motion sensor on top of the putter grip. Figure 6b - Session segmentation used in data collection.

The golfer performed 3,000 putts at The East Renfrewshire Golf Club practise putting green in April 2025. The data was gathered by using a “Blast Motion” sensor, which was attached to the top of the putter grip, as shown in Figure 6a[10, 44].

The data collection process had to be simple and easy to manage, this was achieved by gathering 150 putts per session and repeating this process 20 times. Each session was broken down into 5 segments to avoid golfer fatigue while maintaining a real-world practise session[45-47], as demonstrated in Figure 6b.

The golfer had 6 attempts in each putt before moving to the next segment, creating a total of 30 putts per loop. A series of 5 loops per session completed the requirement of 150 putts per session. This process was repeated 20 times over the course a month and the measurement of each boundary was key in maintaining consistency. This ensured the putts were the same length and the integrity of the data collected was valid[30, 48].

There were eleven metrics the sensor captured on each putt. A comprehensive definition of each metric is available from Blast Motion Sensors webpage[44], please see Appendix B. As previously stated in the literature review, the metrics ranged from speed, directional and rotation of the putting stroke characteristics[9, 25]. The data captured by the sensor for each putt was numbered in sequence, the outcome of these putts was cross referenced with recorded video footage taken at each session.

As explained earlier a supervised learning approach requires the outcomes, this was achieved by using a camcorder in each session, shown at the top of Figure 6b. After each session the researcher downloaded the data from the Blast Motion cloud onto a csv file. Where the golfer holed the putt the outcome metric would be awarded with a value of 1, if a putt were missed a value of 0 was awarded.

As previously discussed, the golfer and researcher are the same person, the use of video supported him perform both roles effectively, as the golfer could focus on the technique of each putt without the additional requirement of noting down the outcome of the putt. Additionally, the researcher had the luxury of being able to rely on video footage to log the outcomes onto the csv file previously mentioned[28]. Thus, adding another layer of robustness in maintaining the reliability and validity of the data collected.

Data Analysis

R programming language was used in this project along with its integrated environment R Studio. This was chosen due to its ability in the statistical field along with its ability to use machine learning techniques[32, 33]. R version 4.3.1 was used in this project, the full code containing step-by-step instructions can be viewed in a Github account which is an open-source developer platform[35], see Appendix A.

A crucial step in the data analysis process was scanning the data for any missing values, the is.na() function determined there was no missing values in the dataset. Another significant aspect was testing the significance of the metrics; a generalised linear model (GLM) function is a well-established approach which determines the contribution of each metric has on the outcome of a prediction.

After several attempts it was determined the most significant metrics were *impact stroke speed*, *face angle at impact*, *loft change* and *lie change*[26]. By reducing the metrics from 11 to 4 metrics the GLM process simplified the models, reduced the risk of overfitting and potentially enabling more accurate and consistent predictions.

After identifying the significant variables, the next step was to split the data into training and testing sets. This is an essential part of the supervised learning process, where the model learns patterns from the training data and uses those patterns to make predictions on the unseen test data.

It was determined the size of the sample was large enough to justify a 70% training and 30% test sample split. This was considered the best option due to an 80/20% split would reduce the test sample; this would limit the evaluation of the models and potentially have an adverse effect on the effectiveness of the predictions. In contrast a 60/40% ratio may limit the machine learning capabilities; patterns may be missed which could potentially lead to less accurate predictions.

Models

Coding is an intrinsic part of data analytics and in this section a brief comprehensive review of each method is provided. Abbreviations were used in the coding to make the administration easier, please see Appendix C for the list of abbreviations used.

As was mentioned in the literature review, these models are adjustable and have many distinctive features available. It is important to note that where there was similarity between these models, a consistent approach was used.

Logistic Regression

**A screenshot of a computer code

AI-generated content may be incorrect.**

Figure - Logistic regression code.

Binomial was the option chosen in the family section of the code; this instructs the machine to create a binary output, as shown in Figure 7. It also instructs the machine to make a prediction based on probabilities. The probability line was set at 0.5, which instructs the machine to convert any result 0.49 and below a negative award (0). A positive result (1) will apply to measurements equal and greater than 0.5.

Support Vector Machines

**A white background with black text

AI-generated content may be incorrect.**

Figure - Support Vector Machine code.

A key feature within the SVM code was the kernel instruction, “radial” was chosen as this represents radiable basis function (RBF - Gaussian), see Figure 8. As indicated in the literature review chapter, this approach works well in multiple dimensions. As there are 4 dimensions within the data this approach appeared to be the most appropriate.

Random Forest

**A screenshot of a computer code

AI-generated content may be incorrect.**

Figure - Random Forest code.

Figure 9 shows the ntree section of the code refers to the number of trees which are used for each prediction. The “mtry” sets the number of variables used in each tree, this has been set as 4 due to the strength of the variables when evaluating the metric significance test.

XGBoost

A screenshot of a computer code

AI-generated content may be incorrect.

Figure - XGBoost Code.

The “nrounds” refers to the number of trees used for each prediction, as shown in Figure 10. To be consistent with random forest 100 trees were used to make these predictions. “Verbose” identifies any warnings through this process, this was set to 0 as notification was not a requirement. The probability line was set at 0.5 to maintain consistency with Logistic Regression.

Model Evaluation

Model evaluation compares actual outcomes with predictions by using a confusion matrix, summarising the results into 4 categories. A performance matrix is derived from the confusion matrix this is to assess the model’s accuracy, precision, recall and f1 score. This additional step provides a deeper analysis of the results and helps identify any imbalances in the model’s predictions.

An award of “true positive” is given when the golfer holes the putt in question and the machine predicts this outcome. “True negative” represents when a putt was missed and the model predicted this correctly. A higher score is preferred in both measurements, which determines the better performing model.

A “false positive” is awarded when the model predicts the putt was holed, however the putt was missed. “False negative” occurs when the model predicted the putt was missed, but this prediction was incorrect as the putt was holed. The preference of these measurements is a lower score; this determines the model which performs best.

Accuracy measured the number of correct predictions were made as a percentage of all 900 test data putts. This was achieved by dividing the TP & TN by the total amount of predictions. Precision measures how well the model performed when predicting the actual holed putts to the predicted holed putts. By dividing the TP by the total of both TP & FP.

Recall measures the proportion of true positive predictions in proportion of all the actual positives (true positive and False negatives). F1 score measures a balance between the Precision and Recall, establishing if a model’s predictions are weighted in a certain measurement and unbalanced.

# Chapter 4 – Results

## Introduction

This section presents the results of the four classification models applied to the test dataset. The confusion matrix contains the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each model. The performance model contains the accuracy, precision, recall and F1-scores.

## Confusion Matrix

Table - Confusion Matrix results

| **Metric** | **XGBoost** | **Random Forest** | **SVM** | **Logistic Regression** | **Mean** | **SD** |
| --- | --- | --- | --- | --- | --- | --- |
| TP | 357 | 359 | 364 | 200 | 320.00 | 80.05 |
| TN | 478 | 477 | 431 | 393 | 444.75 | 40.88 |
| FP | 28 | 29 | 75 | 113 | 61.25 | 40.88 |
| FN | 37 | 35 | 30 | 194 | 74.00 | 80.05 |

### True Positive

A graph of a graph with a green and blue rectangle

AI-generated content may be incorrect.

Figure - True Positive results taken from the confusion matrix.

The Support Vector Machine (SVM) model achieved the highest True Positive prediction with 364. This was closely followed by Random Forest and XGBoost, which reached 359 and 357 respectively. The model which returned the lowest score was Logistic Regression that obtained an accuracy of 200, these results are shown in Table 2.

The True Positive mean was calculated as 320 ± 80.05. Three models scored better than the true negative average, were Support Vector Machine (SVM), XGBoost and Random Forest. Logistic Regression returned with a score that fell below average, as shown in Figure 11.

### True Negative

A green and blue rectangles

AI-generated content may be incorrect.

Figure - True Negative results taken from the confusion matrix.

The by XGBoost model both achieved the best True Negative prediction with 478. This was closely followed by Random Forest which scored 477. Support Vector Machine (SVM) returned with a score of 431. The model which returned the lowest score was Logistic Regression with an accuracy of 393, these results are shown in Table 2.

The True Negative mean was calculated as 444.75 ± 40.88. Two models scored better than the true negative average, which were XGBoost and Random Forest. Logistic Regression and Support Vector Machines returned with a score that fell below average, as shown in Figure 12.

### False Positive

A graph with a green and blue rectangle

AI-generated content may be incorrect.

Figure - False Positive results taken from the confusion matrix.

The by XGBoost model both achieved the best False Positive prediction with 28. This was closely followed by Random Forest which scored 29. Support Vector Machine (SVM) returned with a score of 75. The model which returned the lowest score was Logistic Regression with an accuracy of 113, these results are shown in Table 2.

The False Positive mean was calculated as 61.25 ± 40.88. Two models scored better than the false positive average, which were XGBoost and Random Forest. Logistic Regression and Support Vector Machines returned with a score that fell below average, as shown in Figure 13.

### False Negative

A graph with green and blue squares

AI-generated content may be incorrect.

Figure - False Negative results taken from the confusion matrix.

The by Support Vector Machine (SVM) model both achieved the best False Negative prediction with 30. This was closely followed by Random Forest which scored 35. XGBoost returned with a score of 37. The model which returned the lowest score was Logistic Regression with an accuracy of 194, these results are shown in Table 2.

The false negative mean was calculated as 74 ± 80.05. Three models scored better than the false negative average, which were Support Vector Machine (SVM), Random Forest and XGBoost. Logistic Regression returned with a score that fell below average, as shown in Figure 14.

## Performance Matrix

Table - Performance Matrix results

| **Metric** | **XGBoost** | **Random Forest** | **SVM** | **Logistic Regression** |
| --- | --- | --- | --- | --- |
| Accuracy | 92.78% | 92.89% | 88.33% | 65.89% |
| Precision | 92.73% | 92.53% | 82.92% | 63.9% |
| Recall | 90.61% | 91.12% | 92.39% | 50.76% |
| F1\_Score | 91.66% | 91.82% | 87.39% | 56.58% |

### Accuracy

The by Random Forest model both achieved the best Accuracy result with 92.89%. This was closely followed by XGBoost which scored 92.78%. Support Vector Machine (SVM) returned with a score of 82.92%. The model which returned the lowest score was Logistic Regression with an accuracy of 63.9%, these results are shown in Table 3.

The accuracy mean was calculated as 84.97%. Three models scored better than the false negative average, which were Random Forest, XGBoost and Support Vector Machine (SVM). Logistic Regression returned with a score that fell below average.

### Precision

The by XGBoost model both achieved the best Accuracy result with 92.73%. This was closely followed by Random Forest which scored 92.53%. Support Vector Machine (SVM) returned with a score of 82.92%. The model which returned the lowest score was Logistic Regression with an accuracy of 63.9%, these results are shown in Table 3.

The mean accuracy across all models was calculated as 84.97%. Three models scored better than the false negative average, which were Random Forest, XGBoost and Support Vector Machine (SVM). Logistic Regression returned with a score that fell below average.

### Recall

The by Support Vector Machine (SVM) model both achieved the best False Negative prediction with 92.39%. This was closely followed by Random Forest which scored 91.12%. XGBoost returned with a score of 90.61%. The model which returned the lowest score was Logistic Regression with an accuracy of 50.76%, these results are shown in Table 3.

The Recall mean was calculated as 81.22%. Three models scored better than the false negative average, which were Random Forest, XGBoost and Support Vector Machine (SVM). Logistic Regression returned with a score that fell below average.

### F1 Score

The by Random Forest model both achieved the best Accuracy result with 91.82%. This was closely followed by XGBoost which scored 91.66%. Support Vector Machine (SVM) returned with a score of 87.39%. The model which returned the lowest score was Logistic Regression with an accuracy of 56.58%, these results are shown in Table 3.

The F1Score mean was calculated as 82.11%. Three models scored better than the false negative average, which were Random Forest, XGBoost and Support Vector Machine (SVM). Logistic Regression returned with a score that fell below average.

Chapter 5 – Discussion

Introduction

This chapter discusses the results of the classification models used to predict the outcome of 5-foot putts. The results indicate that a marginal difference in the top performing models, therefore a deeper level of analysis was performed to establish the best performer. In doing so this resulted in finding limitations and an imbalance within the results, which highlights the need for future research in this area.

Key findings and insights

This research has shown Random Forest is the model that performs best at predicting the outcome of 5-foot putts. As it achieved the highest scores on both accuracy and F1 score, and it ranked second in the precision and recall.

Logistic regression consistently underperformed by ranking last in all evaluated metrics. The highest score achieved by the logistic regression model was 65.89%. To surpass the third-ranked model, an improvement of over 22.44% would have been required.

However, the results for Random Forest, XGBoost and Support Vector Machines were closely matched, and the differences were small. Therefore, an examination of these nuances between these three top performing models would better explain the similarities of these results.

A true positive and a false negative both refer to instances where the golfer successfully holed a putt, this occurred 394 times in total. The key distinction is that the true positive represents a correct prediction, whereas a false negative indicates an incorrect prediction.

This relationship shows a higher performing model will have a greater number of true positives and fewer false negatives[42]. The total amount of putts is fixed and an increase in true positives will have an equal and opposite decrease in false negatives, and vice versa.

Support Vector Machines (SVM) performed strongly in terms of true positives and false negatives, by scoring 364 and 30 respectively and ranking first in both categories. This indicates that the model was particularly effective at correctly identifying putts that the golfer holed. This is evident in the recall metric, where SVM achieved the highest score of 92.39%.

Although Random Forest categorically outperformed XGBoost to achieve second place, it done so by only two predictions for both the true positive and false negative metrics. Resulting in a total difference of only four predictions, out of 394 putts this totals a 1.15% margin.

It could be argued that such a margin is small enough to be considered negligible, (remembering the increase and equal and opposite decrease relationship), only two predictions would make both models joint second place performers in these metrics.

In the case of false positives and true negatives, these results represent the putts the golfer missed. A true negative occurs when the model correctly predicts a missed putt. Whereas a false positive is where the model incorrectly predicts a successful putt. Again, these putts were fixed and any increase in one result has an equal decrease in the other.

For both random forest and XGBoost the margin in false positive and true negative results were even smaller, as only one prediction separated these models in each metric. A total difference of only two predictions out of 506 putts, this totals a 0.39% margin. Meaning only a single prediction would be enough to make both models joint top performers in these metrics.

The SVM model performance produced a weaker result at predicting missed putts. As shown by its results in the false positive category, ranking third place with 75 incorrect predictions. This result has a clear impact on this models Accuracy, Precision and F1 score, significantly reducing the SVM model’s overall results.

It could be argued that if false positive results were excluded from this evaluation, the SVM model could become the highest-performing model. Therefore, with slight improvements in this area SVM may be capable of challenging Random Forest and XGBoost as the most effective overall model. Support Vector Machines could benefit by including additional features, which could potentially reduce the number of false positives.

However, to ensure a like for like comparison across all models a consistent approach was maintained, introducing inconsistencies would undermine the integrity of this analysis, such adjustments fall outside the scope of this research.

While XGBoost outperformed Random Forest in the precision metric by 0.2%, Random Forest outperformed XGBoost in accuracy by 0.11%, Recall by 0.59% and F1 by 0.16%. On that basis Random Forest was determined to be the best overall performing model.

Based on the results, both Random Forest and XGBoost models appear to have reached a near optimal performance. Additional improvement could come at the expense of overfitting; this is where the models are at risk of being overly tailored of the patterns within the training data. Although this area falls outside the scope of this study, it is fair to say that both models do not require additional information to improve their performance.

Limitations

This study was limited to a single participating golfer, which restricted the variability of the putts used for analysis. It is unclear to what extent the golfer’s unique technique may have compensated for their flaws when predicting the outcome of putts. These compensations are specific to the individual, and metric readings may not be transferrable to other golfers and false readings may occur as a result.

For example, a golfer might have a natural aiming tendency to the right but compensate with the putter having a left facing angle at impact. This type of adjustment is specific to the individual, which may not be generalisable to other golfers. This study was unable to measure the starting aim of a putt; however, it did record the face angle at impact. Such a limitation could introduce an error in the predictions for a different golfer who does not compensate in the same way. Despite having the same face angle at impact, the outcome would be different, and the prediction would be incorrect.

Consequently, the dataset may contain patterns which are misleading and as the machine learns compensated metrics, this could lead to incorrect predictions for other golfers.

As previously mentioned, the data collection process recorded putts from 5 different angles, allowing a variety of putts. Each angle featured a unique slope, independent of the other putting angles in subsequent sessions. Typically, each angle included on straight putt, two putts sloping left to right and two putts sloping right to left. On a left-to-right slope the ball curves from the left towards the right and vice versa for a right-to-left slope (referred to as the “break”). If the golfer does not consider the break, the putt is likely to miss due to the curvature of the balls path on the slope.

Unfortunately, these slope measurements were not included in the planning phase and as a result were not incorporated into the dataset used for predictions. This may have had a significant impact as each session contained 5 different slopes, as there were 20 sessions, a total of 100 measurement opportunities were missed. In hindsight, a digital spirit level would have been a sufficient tool to capture these angles[2, 49].

Another key factor in putting is the speed at which the ball rolls towards the hole. Stroke speed at impact was measured, however the speed of the green was not.

Moisture on the green increase’s friction (increased resistance) between the ball and the surface, causing the ball to roll slower and cover less distance, significantly influencing the speed of the green[2]. The green where the data was collected was equipped with a sprinkler system, making a rain gauge ineffective.

Additionally, the grass was cut at first light daily, the time of day when these sessions took place was not consistent and as the temperatures of each day fluctuated. This may have had a minimal effect on the growth of the grass impacting the speed of the green[49].

This research would have benefitted from a measure green speed using a stimp meter, this provides a distance that the ball rolls[49]. A consistent measurement can be carried out prior to each data collection session.

To ensure a like for like comparison between the models, a consistent methodological approach was used. For example, the use of a GLM function helped reduced the number of variables, aiming to minimise overfitting and strengthen the relationship between variables and outcomes.

It could be argued that each model, despite these similarities requires a tailored approach to reach optimal predictive performance. Consistency was prioritized due to the time constraints involved in identifying the optimal parameters of each model. Also, using an optimal approach might lead to different configurations and potentially bringing the integrity of the analysis into question.

This limitation must be acknowledged due to the findings the SVM results. As this model may have achieved a better ranking if the results of false positives were improved by using an additional lay of analysis. Similarly, the logistic regression model may have benefited from a feature engineering approach, which could have reduced the risk of overfitting to a higher degree and producing better results. In addition, this project was carried out by an individual researcher and due to time constraints, this level of analysis was not feasible.

Future Work

As previously discussed, this research focused on the putting performance of a single participant. A key limitation was identified that contamination of the data due to a compensation in the golfer’s technique. Future research of a larger sample size could reduce the impact of these compensation, improving the model’s predictions.

Additional measurements that were not recorded during the data collection process could have benefitted this study. Metrics such as green speed, slope angle and surface moisture would add value able context; by adding this additional layer it could improve the performance and reliability of these predictions[2, 49].

Although putt outcomes were captured on camcorder, two additional details such as direction and distance could have been added on missed putts, highlighting where and how these putts were missed. By introducing labels such as “left” or “right” (direction variable) and “long” or “short” (distance variable), this could provide more areas for the models to predict.

If this approach was used on this research project, the new predictions would be based on the 506 missed putts. The benefit of this addition would be more areas to predict, which could enhance the performance of these models. These supplementary results could benefit a performance analyst as they could tailor lessons on the insights provided by the new predictions.

There are many opportunities for further analysis using the current dataset, for example reviewing the segments or sessions for trends could reveal if a model worked best in a specific area. Did the model make better predictions on putts that did not have a slope? Or did the model perform better at predicting putts that sloped right to left as opposed to left to right?

The results showed the SVM model performed better in predicting successful putts, future research could establish if there is a correlation between the golfer’s performance and the model’s prediction accuracy.

Further studies could use an optimal approach rather than a consistent methodology used in this research. Such as adjusting the training/test dataset split and altering the probability threshold, as these may need fine tuning independently for each model.

Had this methodology used two or more putters it would have been possible to determine with an accuracy of 92.89% which putter the golfer performed best with. This could benefit the club manufacturer during its design and fitting process, also the golfer could potentially hole more putts as they will be using a putter tailored to their needs.

The full analytical process has been uploaded to the Github repository, including the dataset and R code, with a step-by-step instruction making it easy to use. This open-source platform allows golfers, coaches and individuals who are new to coding access to perform their own data analysis, this challenges the elitist view on the use of data in sport[34].

The methodology is adaptable for further use in golf or other unrelated research, such as aiming to predict the outcome of penalties in football over a season. Requirements include a new dataset to be uploaded and minimal alterations to the code to provide a set of new predictions. Future studies would also contribute to how well these models performed over time using methods such as time series analysis.

Chapter 6 – Conclusion

Background & Aim

Due to there being a limited amount of previous research in putting it was identified that predicting the outcomes of putts appears to be a novel idea. The aim of this research was to establish which machine learning classification model best predicts the outcome of 5-foot putts.

Pursuing this study addressed a gap in research bringing both machine learning and putting prediction together. This research has the potential to act as a benchmark for future research, the methods used are transferrable to wider areas in golf and could be applied to different sports and other areas predicting classification studies.

Methods

The data gathering process was time-consuming, the benefit of using primary data was that full control was maintained regarding the validity and reliability of the data collected. The data collected was fit for purpose avoiding use of secondary data, secondary data is often collected for a different purpose. Therefore, this was a key strength of this study, which in reflection was the correct approach.

Using a machine learning classification approach, this enabled models to make predictions to be made on a binary outcome. As the dependent variable was recorded, the use of supervised learning was applied using a dataset that was split into training and testing data.

There was a level of judgement on choosing the training and testing proportional split, on reflection this area could be considered for further research due to the marginal differences between the results.

The results were shown in a confusion matrix, additional steps were taken, and subsequent measurements facilitated a review of the results to a higher degree. This approach has shown to be a strength of this study, without these additional steps, determining the outright best performer would have been difficult to quantify.

Key findings

The results show Random Forest as the best performing model. However, it was noted that the difference was marginal, a little alteration in the methods could return XGBoost being the best performing model.

Support Vector Machines produced strong results, the weakness within this model was the number of false positive results. SVM handles multiple dimensions very well, further study would be beneficial in determining if this model perform better with additional variables.

Contribution to the field

The contribution to data analytic within golf that this research provides should not be understated, as this approach appears under-explored. It also challenged the view that data analysis is an elitist approach[34].

This was done by using affordable equipment such as a “Blast motion sensor”, along with other analytical tools which are open source and free to use. This research provides golfers and coaches the method and tools required to carry out a training session using machine learning to determine the outcome of putts. This contribution is a strength to this study, due to the potential influence of a real-world application.

Future research

Additional binary metrics could be recorded in the event the putts are missed, providing the golfer and coach a more detailed analysis. As stated in previous chapters, by including additional variables there is potential to improve the performance of these models. Due to the missed opportunity of gathering additional data, this could be seen as a weakness in this study. However, further research could explore what model performs best with the additional dependant variables previously mentioned.

Updates are easily made as the R Studio code is available for public use on GitHub. It could be argued that this project has the potential to be future proof due to the ease in which GitHub updates work. This could be considered as the most impactful strength of this study, its potential to enable future researchers to use this dataset and code to take this study to another direction.

Final remarks

This study has made a valid contribution to the use of data analytics in putting, an area where research is limited. This study has provided a benchmark which data analysts, golfers and performance analysts can build on and use for their own work. Looking back on this project the author ends this study with a sense of curiosity, what shape future research will take and how this contributes to the application of data analysis in putting.

# References

[1] H. Gillmeister, "Golf on the Rhine: On the origins of golf, with sidelights on polo," *The International Journal of the History of Sport,* vol. 19, no. 1, pp. 2-30, 2002.

[2] B. W. Holmes, "Putting: How a golf ball and hole interact," *American Journal of Physics,* vol. 59, no. 2, pp. 129-136, 1991.

[3] D. L. Alexander and W. Kern, "Drive for show and putt for dough? An analysis of the earnings of PGA Tour golfers," *Journal of sports Economics,* vol. 6, no. 1, pp. 46-60, 2005.

[4] Y. Hasegawa, S. Koyama, and K. Inomata, "Perceived distance during golf putting," *Human Movement Science,* vol. 32, no. 6, pp. 1226-1238, 2013.

[5] I. Muhammad and Z. Yan, "SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY," *ICTACT Journal on Soft Computing,* vol. 5, no. 3, 2015.

[6] E. Beauxis-Aussalet and L. Hardman, "Visualization of confusion matrix for non-expert users," in *IEEE Conference on Visual Analytics Science and Technology (VAST)-Poster Proceedings*, 2014: sn, pp. 1-2.

[7] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-label confusion matrix," *Ieee Access,* vol. 10, pp. 19083-19095, 2022.

[8] P. Wheeler, "A gaze at the galleries: spectators at championship golf events in Kent 1892–1938," *Sport in History,* vol. 43, no. 4, pp. 410-441, 2023.

[9] M. S. Couceiro, A. G. Araújo, and S. C. Pereira, "InPutter: an engineered putter for on-the-fly golf putting analysis," *Sports Technology,* vol. 8, no. 1-2, pp. 12-29, 2015.

[10] A. Ríos Fuentes, S. Calero Morales, R. E. Eizméndiz Domínguez, Y. Aguiar Alpízar, and B. Sánchez Córdova, "Variables in the Cuban baseball batter's performance using the Blast Motion Baseball sensor," *PODIUM-Revista de Ciencia y Tecnologia en la Cultura Física,* vol. 15, no. 3, 2020.

[11] R. Gilgen-Ammann, T. Schweizer, and T. Wyss, "Accuracy of distance recordings in eight positioning-enabled sport watches: instrument validation study," *JMIR mHealth and uHealth,* vol. 8, no. 6, p. e17118, 2020.

[12] J.-M. Huang, T.-K. Ho, Y.-C. Liu, and Y.-H. Lin, "A discussion on the user intention of golfers toward golf GPS navigation," *Journal of Hospitality and Tourism Technology,* vol. 6, no. 1, pp. 26-39, 2015.

[13] M. D. Hughes and R. M. Bartlett, "The use of performance indicators in performance analysis," *Journal of sports sciences,* vol. 20, no. 10, pp. 739-754, 2002.

[14] "Top Tracer Website." <https://www.toptracer.com/about/> (accessed.

[15] "Indoor Golf Bar." <https://thegolfbar.co.uk/> (accessed.

[16] B.-H. Seong and C.-Y. Hong, "Corroborating the effect of positive technology readiness on the intention to use the virtual reality sports game “Screen Golf”: Focusing on the technology readiness and acceptance model," *Information Processing & Management,* vol. 59, no. 4, p. 102994, 2022.

[17] T. Oh and J.-H. Kang, "Analysis of the sports broadcasting market in the television broadcasting industry," *European Sport Management Quarterly,* vol. 22, no. 6, pp. 856-875, 2022.

[18] S. J. Robertson, A. F. Burnett, and R. U. Newton, "Development and validation of the Approach-Iron Skill Test for use in golf," *European journal of sport science,* vol. 13, no. 6, pp. 615-621, 2013.

[19] F. Alam *et al.*, "A study of golf ball aerodynamic drag," *Procedia Engineering,* vol. 13, pp. 226-231, 2011.

[20] "Inside the golf-ball rollback debate: Here’s what both sides are saying." <https://golf.com/news/distance-rollback-pro-anti-both-sides/> (accessed.

[21] B. Dyer, "The controversy of sports technology: a systematic review," *SpringerPlus,* vol. 4, pp. 1-12, 2015.

[22] P. O'Donoghue, *An introduction to performance analysis of sport*. Routledge, 2014.

[23] B. Taheri, C. Porter, N. Valantasis-Kanellos, and C. König, "Quantitative data gathering techniques," *Research methods for business and management,* vol. 4, no. 14, pp. 913-930, 2015.

[24] C. L. Winchester, M. J. Salji, and V. Kasivisvanathan, "Gathering preliminary data," *Journal of Clinical Urology,* vol. 10, no. 6, pp. 568-572, 2017.

[25] C. Marquardt, "The SAM puttLab: concept and PGA tour data," *International Journal of Sports Science & Coaching,* vol. 2, no. 1\_suppl, pp. 101-120, 2007.

[26] J. Karlsen, G. Smith, and J. Nilsson, "The stroke has only a minor influence on direction consistency in golf putting among elite players," *Journal of Sports Sciences,* vol. 26, no. 3, pp. 243-250, 2008.

[27] C. Fernández-Echeverría, I. Mesquita, J. González-Silva, and M. P. Moreno, "Towards a more efficient training process in high-level female volleyball from a match analysis intervention program based on the constraint-led approach: the voice of the players," *Frontiers in Psychology,* vol. 12, p. 645536, 2021.

[28] M. Stein *et al.*, "Bring it to the pitch: Combining video and movement data to enhance team sport analysis," *IEEE transactions on visualization and computer graphics,* vol. 24, no. 1, pp. 13-22, 2017.

[29] H. Elzbieta, U. Piotr, and O. D. Peter, "Reliability of a computerised system for recording location on a tennis court," *International Journal of Performance Analysis in Sport,* vol. 8, no. 3, pp. 138-144, 2008.

[30] S.-M. Cooper, M. Hughes, P. O’Donoghue, and M. A. Nevill, "A simple statistical method for assessing the reliability of data entered into sport performance analysis systems," *International Journal of Performance Analysis in Sport,* vol. 7, no. 1, pp. 87-109, 2007.

[31] H. Liu, W. Hopkins, A. M. Gómez, and S. J. Molinuevo, "Inter-operator reliability of live football match statistics from OPTA Sportsdata," *International Journal of Performance Analysis in Sport,* vol. 13, no. 3, pp. 803-821, 2013.

[32] E. Nordmann, P. McAleer, W. Toivo, H. Paterson, and L. M. DeBruine, "Data visualization using R for researchers who do not use R," *Advances in Methods and Practices in Psychological Science,* vol. 5, no. 2, p. 25152459221074654, 2022.

[33] S. Simsek, A. Albizri, M. Johnson, T. Custis, and S. Weikert, "Predictive data analytics for contract renewals: a decision support tool for managerial decision-making," *Journal of Enterprise Information Management,* vol. 34, no. 2, pp. 718-732, 2021, doi: 10.1108/JEIM-12-2019-0375.

[34] D. Greenbaum, "Wuz you robbed? Concerns with using big data analytics in sports," *American Journal of Bioethics,* vol. 18, no. 6, pp. 32-33, 2018.

[35] L. Dabbish, C. Stuart, J. Tsay, and J. Herbsleb, "Social coding in GitHub: transparency and collaboration in an open software repository," in *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 2012, pp. 1277-1286.

[36] A. T. Yigit, B. Samak, and T. Kaya, "An XGBoost-lasso ensemble modeling approach to football player value assessment," *Journal of Intelligent & Fuzzy Systems,* vol. 39, no. 5, pp. 6303-6314, 2020.

[37] Y. Eratlı Şirin and M. Şahin, "Investigation of factors affecting the achievement of university students with logistic regression analysis: School of physical education and sport example," *Sage Open,* vol. 10, no. 1, p. 2158244020902082, 2020.

[38] B. Bačić, "Predicting golf ball trajectories from swing plane: An artificial neural networks approach," *Expert Systems with Applications,* vol. 65, pp. 423-438, 2016.

[39] O. Wiseman, "Using machine learning to predict the winning score of professional golf events on the PGA tour," Dublin, National College of Ireland, 2016.

[40] Y. Li and Y. Mu, "Research and performance analysis of random forest-based feature selection algorithm in sports effectiveness evaluation," *Scientific Reports,* vol. 14, no. 1, p. 26275, 2024.

[41] B. S. Baumer, G. J. Matthews, and Q. Nguyen, "Big ideas in sports analytics and statistical tools for their investigation," *Wiley Interdisciplinary Reviews: Computational Statistics,* p. e1612, 2023.

[42] E. Beauxis-Aussalet and L. Hardman, "Simplifying the visualization of confusion matrix," in *26th Benelux conference on artificial intelligence (BNAIC)*, 2014.

[43] S. Wilkens, "Sports prediction and betting models in the machine learning age: The case of tennis," *Journal of Sports Analytics,* vol. 7, no. 2, pp. 99-117, 2021.

[44] "Blast Motion Golf Sensor." <https://blastmotion.com/products/golf/> (accessed.

[45] N. R. Higdon, W. H. Finch, D. Leib, and E. L. Dugan, "Effects of fatigue on golf performance," *Sports Biomechanics,* vol. 11, no. 2, pp. 190-196, 2012.

[46] D. J. Thompsett, K. A. Vento, C. Der Ananian, D. Hondula, and F. C. Wardenaar, "The effects of three different types of macronutrient feedings on golf performance and levels of fatigue and alertness," *Nutrition and Health,* vol. 28, no. 4, pp. 509-514, 2022.

[47] N. Perkins-Ceccato, S. R. Passmore, and T. D. Lee, "Effects of focus of attention depend on golfers' skill," *Journal of sports sciences,* vol. 21, no. 8, pp. 593-600, 2003.

[48] R. J. Leach, S. E. Forrester, A. Mears, and J. R. Roberts, "How valid and accurate are measurements of golf impact parameters obtained using commercially available radar and stereoscopic optical launch monitors?," *Measurement,* vol. 112, pp. 125-136, 2017.

[49] J. Richards, D. Karcher, T. Nikolai, J. Henderson, and J. Sorochan, "A comparison of two devices used to measure golf course putting green speed. Online. Applied Turfgrass Science doi: 10.1094," ATS-2009-0724-02-RS, 2009.

# Appendix

## Appendix A

[SGD80/MScModelPerformance](https://github.com/SGD80/MScModelPerformance) – Click here if accessed via a PC

<https://github.com/SGD80/MScModelPerformance> - URL address

## Appendix B

Backstroke Time

The time it takes to complete your Backstroke Address to top of backstroke, measured in seconds.

Forward Stroke Time

The time it takes to complete your Forward Stroke, measured in seconds.

Tempo

Tempo is the ratio between Backstroke Time and Forward Stroke Time.

Total Stroke Time

The time it takes to make a complete putting stroke. From start of the stroke to impact, measured in seconds.

Backstroke Length

The length the putter face travels. From address to the top of the backstroke.

Loft Change

The increase or decrease in Loft between the start of the stroke and impact, measured in degrees.

Impact Stroke Speed

The speed of the sweet spot of the putter face at the moment of impact, measured in miles per hour.

Face Angle at Impact

The difference in rotation between Backstroke Rotation and Forward Stroke Rotation, measured in degrees.

Lie Change

The increase or decrease in the Lie Angle between the start of the stroke and impact, measured in degrees.

Backstroke Rotation

The rotation of the putter face during the backstroke, measured in degrees.

Forward Stroke Rotation

The rotation of the putter face during the forward stroke, measured in degrees.

## Appendix C

BackS - Backstroke Time

ForwardS – Forward Stroke Time

TST – Total Stroke Time

Tem - Tempo

ISS – Impact Stroke Speed

BSL – Backstroke Lenght

LoftCh - Loft Change

BROT – Backstroke Rotation

ForRot - Forward Rotation

FAI – Face Angle at Impact

LieCh – Lie Change