Using artificial intelligence to predict golfer's final scores in the US Masters.

Mark Kennedy

15312186

Abstract

The sport of golf has evolved dramatically in recent years with the emergence of new performance metrics changing how players approach competitions. It is hard to dispute that anyone has had a bigger impact on the sport recently than Byrson Dechambeau, who has been dubbed the 'mad scientist' for how he tailors his game around improving certain performance metrics. The abundance of metrics gives us great insights into player's strengths and weaknesses, but can they be used to accurately predict player's final scores in competitions? The aim of this project is to apply artificial intelligence techniques (such as linear regresson and neural networks) to historical data to try and predict the final scores of golfers in this years US Masters (April 2021). Feature importance will then be applied to the best performing model to gain insights into what performance metrics are most influential in determining a player's final score in the competition.

1 Introduction

Golf like most sports has become increasingly data driven in recent years. ShotLink is a system developed by the PGA to keep track of the players scores. ShotLink is always evolving as it integrates the latest technology that enable the capture of new metrics, which give us even greater insights into the sport. The abilty to accuractely record the distance of shots led to the creation of 225 new metrics within the first few years of ShotLink's introduction. The latest developments include the capturing of club swing speed through the use of lasers at tee boxes. With the abundance of captured data and metrics now widely available, it is no surprise to see players such as Byrson DeChambeau fixate over how they can leverage the insights gained from these metrics. Recently we have seen a trend developing in golf where golfers such as Byrson DeChambeau have prioritised achieving more distance in their drives, even if it does mean sacrificing their driving accuracy. Through the application of artificial intelligence (AI) techniques, the main task of this research project is to see if some of the main performance metrics available today can be used to accurately predict the players actual final score in the competition. This could lead to some interesting insights into what metrics have the greatest influence on a players score. However, the underlying goal of this project is to explore the possibilities and limitations that face applications of AI techniques today. The modern concept now commonly known as AI was first coined by it's founding fathers in 1956. Their proposition was "that every aspect of learning or any other feature of intelligence can in principal be so precisely described that a machine can

be made to simulate it" [1]. Over half a century later and the goals set out by the founding fathers have still to be realised. Although the premise of Artifical Intelligence remains the same, the term itself has an almost ambigous element to it today. In order to be able to define Artifical Intelligence one must first be able to define what exactly intelligence is first. Naturally when people hear the term they immediately begin to compare the performance of today's AI with human intelligence, something which Luc Steels believes "can only lead to disappointment"[2]. Some of the biggest advancements in the field of AI have been spurred on by challenges set to the field, as opposed to disputing the true definition of the term. Alan Turing who was one of the most renouned mathematicians of his generation decided to "side-step" defining intelligence and instead decided to develop a test to help identify it. The Turing test is a "classical approach to determining whether a machine is intelligent"[3] or not. According to the Turing test a machine is considered inteligent if human judges cannot effectively discrimate between a computer and a human through text conversation. Although these tests and challenges spurred on great advancements in the field of AI in the past, they have now become inadiquate tests of intelligence in today's environment due to the advancements in technology. The development of cloud computing now means that huge banks of data can be stored and accessed extremely quickly, enabling machines the ability to appear intelligent when in fact they are just implementing basic search algorithms and heuristics. The only task of intelligence in these instances is to "avert the ever-present threat of exponential explosion of search"[4]. The field of AI now finds itself at a crossroads where it must define what exactly intelligence is as it is not enough anymore to merely resemble inteligence. Many believe that inorder for machines to have human-like intelligence they would require "them to have human-like being in the world, which would require tham to have bodies" of their own [5]. Other common beliefs are that modern-day AI lacks 'common sense' and 'understanding' of its own. This has led to the establishment of the field of computational neuroscience which studys "what information processing is actually carried out by natural brains as well as offering new ideas to AI about what mechanisms might be needed for intelligence"[2]. This is no easy task as Jordan Pollack states that "most of what our minds our doing involves mindless chemical activity" [5]. It's hard to predict what the filed of AI will be able to achieve in coming years as "computers will continue to get faster by a factor of 1000per decade"[6] in accordance to Moore's law. The task for practitioners now is to ensure that the progress with software remains up to speed. This will of course become easier with the advancements in computational power as "many tasks that are hard for today's software on present machines will become easy without even fundamentally changing the algorithms" [6]. A safer prediction for the future of AI is that it will continue in its cycle of "periods of optimistic predictions and massive investment and periods of dissapointment, loss of confidence and reduced funding" [1], which Melanie Mitchell refers to as AI springs and AI winters in her paper 'Why AI is Harder Than We Think".

The background research will explore the different machine learning techniques that are capable of outputting continuous values in the form of golf scores. The experimental design section will then describe in detail the research questions being asked, the data processing steps which were taken, and the parameters that were used when training the models, along with any other details needed to replicate this work. The results and

discussion sections will then explore the various findings from this research project and how they relate to the initial research questions set out in the experimental design section.

2 Background

2.1 Neural Networks

In recent years neural networks have become a hot topic in the machine learning branch of AI. Neural networks "are what power all of the major AI advances we've seen in the past decade, including speech recognition, machine translation, chat bots, image recognition, game playing, and protein folding among others" [1]. The brain-inspired multi-layered approach is modelled on the interconnected neurons in the human brain, and is capable of learning complex patterns and relationships within the data, which reduces the need for feature engineering and human understanding of the input data. The simplest component of a neural network is a perceptron. A perceptron (figure 1) is an artificial neuron which is capable of taking in multiple inputs, and producing a single output. Each input is assigned a weight in accordance with their importance to the output. The weighted sum of these inputs is then passed to an activation function which determines if the perceptron activates or not.

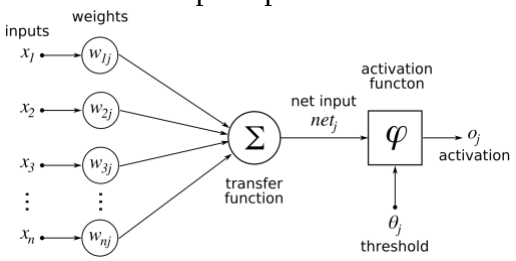


Figure 1: *Perceptron*

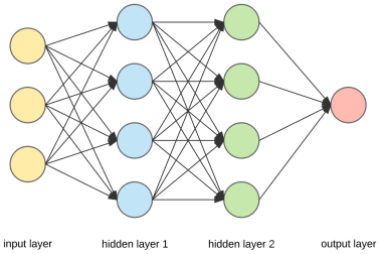


Figure 2: *Neural Network*

A bias neuron can be added to the perceptron to adjust the threshold at which the perceptron's activate. Activation functions capable of outputting continuous values include the sigmoid, tanh, and relu functions which can all be seen below in figure 3.

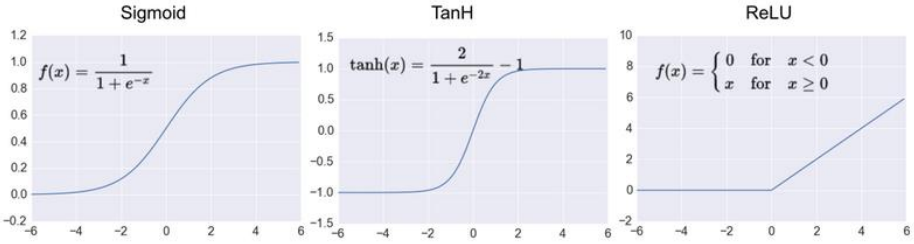


Figure 3: Activation Functions

A multi-layer neural network is made up of many layers of these perceptron's, the inner layers of the network are hidden which is why many people refer to this approach as a 'black box' approach as we cannot see the inner workings of the model (figure 2). The multi-layered approach enables the model to understand non-linear relationships unlike the linear regression approach which will be discussed below. Neural networks generally take a long time to train due to the large quantity of data that they require, making them

computationally expensive to use. Neural networks are one of the most advanced techniques currently used in the field artificial intelligence today and have led to some ground-breaking applications such as speech recognition. To achieve true artificial intelligence AI must be able to think for itself "and be able to use what it has learned in different contexts" [5]. In my opinion in other to achieve this it would almost certainly entail a 'black box' element to it, like what we see with the neural network approach described above. However, with new regulations being proposed by governing bodies such as the EU [8], any AI implementations will be required to have a certain level of transparency to ensure that ethical guidelines are met. While "adding code to provide insight into weights in different layers of a neural network, is not an impossibility – it's a latency issue"[8] which slows down an already computationally expensive process.

2.2 Linear Regression

One of the most common approaches is to apply a linear regression model to the training dataset. Linear regression was widely used in the field of statistics before it was adopted by machine learning. The model is extremely simple to understand unlike many other machine learning techniques. Linear regression involves fitting a line of best fit to the data so that is as close to all of the data points as possible (see figure 4). Once the regression line is determined we can use it to predict the output values (y) for new input values (x). Linear regression models are generally quick to implement, which means that the regression line can easily adjusted if needed in the event of the training set growing in volume. Multiple linear regression is an adaptation of linear regression, which enables the use of multi-dimensional input data. The main difference between the two is that instead of using a regression line, multiple linear regression uses a regression plane as seen in figure 5. The multiple linear regression approach will be suitable for this research problem as we will be using multiple features in the form of performance metrics to predict the players' final scores.

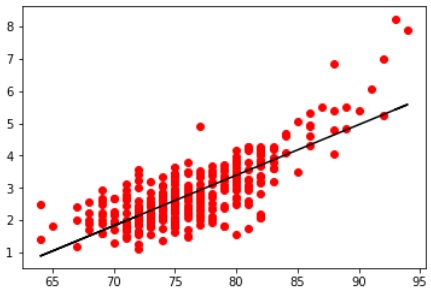


Figure 4: *Linear Regression*

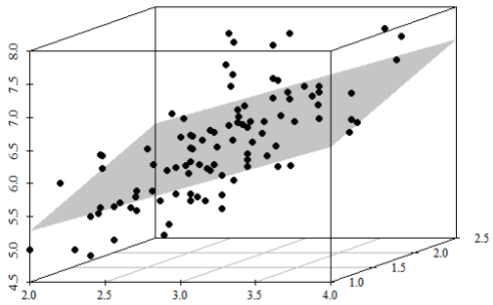


Figure 5: *Multiple Linear Regression*

Linear regression is a simple model which has been borrowed from the field of statistics. Its linear approach means that it does not perform well when non-linear relationships exist within the data. Therefore, it is often seen as a tool for machine learning as opposed to actual artificial intelligence. I will adopt this approach in this research project as it is highly suitable for the problem being addressed and will serve as a good benchmark to compare more intelligent approaches such as neural networks to.

Other noteworthy approaches include: K-Nearest Neighbour (KNN), Regression Tree

3 Experimental Design

This section will detail all of the programming steps carried out during this project, so that it can be easily replicated in the future. It is also worth noting that through the background research I discovered that it was common practice to normalise data before feeding it to machine learning techniques, to reduce training times, and to also ensure that no feature is prioritised during the training of models. As the standard of scoring varied from year-to-year due to variables such as the weather and course difficulty (i.e. pin positions) I decided that it was not appropriate to apply normalisation in this case. For example the driving distances in one year might be longer or have a skewed distribution than previous years due to variables such as higher winds, and so it would be wrong to normalize the driving distances in this case. I have however included an appendix section at the end of this paper where I have included the results obtained from models trained on normalized data.

GitHub Repository: https://github.com/markkennedy3/Predicting_Golf_Scores

3.1 Research Questions

Core Research Question 1: What trends or insights can be found within the performance metrics of players who competed in past US Masters competitions?
Core Research Question 2: Can artificial intelligence techniques, such as neural networks and regression models, be used to accurately predict golfers final scores in the US Masters, from using just the players performance metrics over the four rounds of the competition? Compare different the performance of the different techniques/models used.
Stretch Question: Which performance metric(s) have the biggest influence on a player's final score in the competition?

3.2 The Data Preperation

The data for this research project was sourced from two websites. The performance metrics were scraped from the official PGA website [10], while the official scores were scraped from the ESPN website [11]. The performance metrics gathered were as follows:

1. **Average driving distance:** "The average number of yards per measured drive."
2. **Driving accuracy:** "The percentage of time a tee shot comes to rest in the fairway (regardless of club)."
3. **Greens in regulation:** "The percent of time a player was able to hit the green in regulation (greens hit in regulation/holes played)."
4. **Scrambling:** The percent of time a player misses the green in regulation, but still makes par or better.
5. **Sand saves:** "The percent of time a player was able to get 'up and down' once in a greenside sand bunker (regardless of score)."
6. **Putting average:** "The average number of putts per green in regulation"

I chose these metrics because I believe they give the greatest insight into how players are performaning and cover all aspects of a player's game (i.e. driving, approach to green and putting). The performance metrics from the past twenty years (2001-2021) were scraped form the PGA website. Once the data was collected I then began to prepare and cleanse the data. The following years were then excluded from the data set.

2020: This was the first US masters to be played outside of the month of April due to the Covid-19 pandemic. The weather conditions in November led to the record lowest ever winning score being recorded by Dustin Johnson (-20)
2007: The 2007 Masters was effected by extreme winds and cold weather which led to the joint highest ever winning score being recorded by Zach Johnson (+1). This is an outlier when you consider that the average winning score of the masters tournament is -11.
2002 & 2001: The Augusta National Golf club was reconstructed in 2002 doubling its size. Therefore any competitions played before 2003 were practically played on a different course.

Now that the data was cleansed and outliers were removed, the players final scores were added to the dataset. The two datasets were joined on player names, and any players who failed to make the cut were removed from the dataset as there was no performance metrics available for them on the PGA website. After these pre-processing steps were complete there was seventeen years of data left. I then conducted exploratory analysis on the data to examine trends and gain insights. The next step then was to split the data into a training and test set. For the training set I took the sixteen years of data prior to the 2021 competition, this data was then used to train the models. The test set then was the 2021 US masters competition data for which I would try to predict the final scores for.

3.3 The Algorithms

The implementations of the neural network and linear regression models used in this research project were sourced from the scikit-learn machine learning library in python [12]. The parameters used for both models can be seen below in figures 6 and 7 respectively. For the neural network model, I created three different instances of the model using the sigmoid (logistic), tanh and relu activation functions to see which one performed best.

```
{'activation': 'logistic',
 'alpha': 0.0001,
 'batch_size': 'auto',
 'beta_1': 0.9,
 'beta_2': 0.999,
 'early_stopping': False,
 'epsilon': 1e-08,
 'hidden_layer_sizes': (100,),
 'learning_rate': 'constant',
 'learning_rate_init': 0.001,
 'max_fun': 15000,
 'max_iter': 200,
 'momentum': 0.9,
 'n_iter_no_change': 10,
 'nesterovs_momentum': True,
 'power_t': 0.5,
 'random_state': None,
 'shuffle': True,
 'solver': 'adam',
 'tol': 0.0001,
 'validation_fraction': 0.1,
 'verbose': False,
 'warm_start': False}
```

Figure 6: NN Model Parameters

```
{'copy_X': True,
 'fit_intercept': True,
 'n_jobs': None,
 'normalize': False}
```

Figure 7: LR Model Paramaters

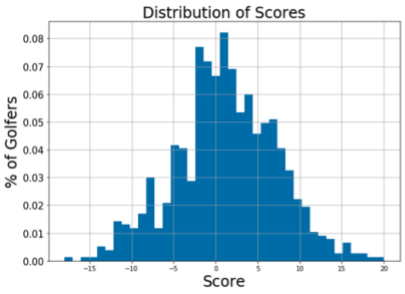
3.4 Advanced: Feature Importance

The coefficients property (coeff_) for each of the input features can be taken as indications of their importance in the linear regression model. However, as mentioned above it was not deemed appropriate to normalise the data in this instance. This might

mean that the results of the feature importance might be flawed. Calculating the feature importance for a neural network is more complex however due to its 'black box' nature. One option is to remove each feature form the model one-by-one and to replace them with random noise, and then check how the performance of the model changes. However, due to the time constraint for this project this was unfeasible.

4 Results

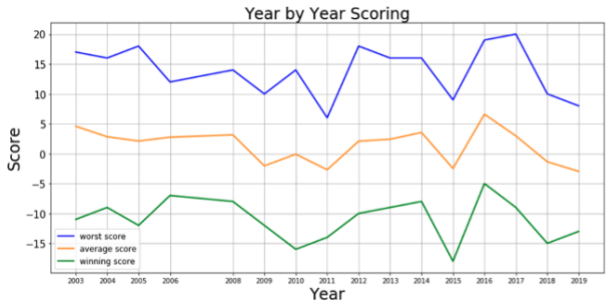
4.1 Exploratory Analysis



	Score
Lowest (Best) Score	-18
Average Score	+1.26
Highest (Worst) Score	+20

Figure 8: Scoring Distribution.

The first analysis I carried out was to plot the distribution of all the final scores in the dataset (figure 8). I then found the lowest, average, and highest scores recorded over the 16 years, which can be seen in the table above. There was a range of 38 scores between the worst score in the dataset and the best score. The average score of all players that have played in the US Masters is +1.26.



	Std (σ)
Winning Scores	3.54
Average Scores	2.89
Worst Scores	4.28

Figure 9: Standard Deviation of Scoring.

Figure 9 shows a timeseries of the of the winning, average, and worst scores of each year. From looking at the graph in figure 9, it is clear to see that the level of scoring tends to fluctuate year-on-year. Different variables can attribute to this, such as weather conditions and course layout (i.e., pin positions). I then calculated the standard deviation for the winning, average, and worsts scores of each year which can also be seen in figure 9. The standard deviation of the average scores each year is 2.89, which demonstrates just

how much the level of scoring can fluctuate each year. Not shown in this figure but worth noting is that the average winning score in the US Masters is -11.

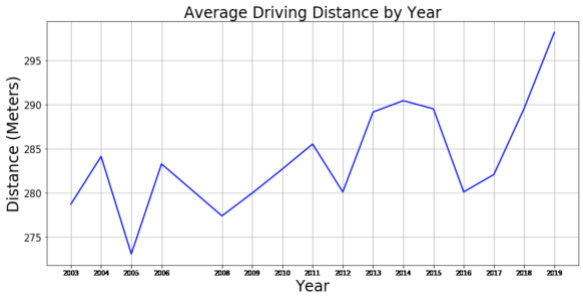


Figure 10: Driving Distance

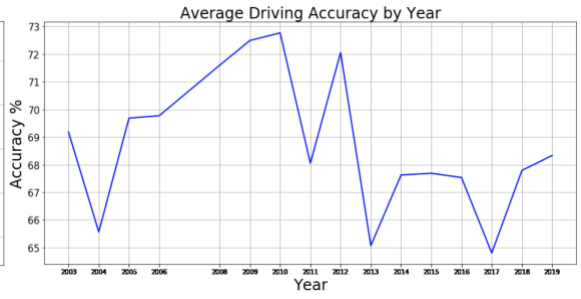


Figure 11: Driving Accuracy

Figure 10 shows that the average driving distance for each year is on an upward trend. The average golfer in the 2019 US Masters was driving the ball twenty meters further than their counterparts in the 2003 competition. While the average driving accuracy of golfers in the 2019 US masters has fallen to 68.33% from its high of 72.77% in 2010 (a fall of 4.44%).

4.2 AI Model Predictions and Performances

	Linear Regression	Neural Network (Sigmoid)	Neural Network (Tanh)	Neural Network (Relu)
Mean Absolute Error	1.62	3.56	3.83	3.40
Mean Squared Error	4.35	19.06	21.33	17.79
Root Mean Squared Error	2.08	4.37	4.62	4.22

Figure 12: Model Performances

The best performing model by far was the linear regression model which had a mean absolute error of 1.62. This means that for every prediction made by the model it was on average 1.62 shots away from the actual score. The best performing neural network model was the model that used the relu activation function, which had a mean absolute error of 3.4.

Actual Top 10:

	Player	Actual	Predicted
45	Hideki Matsuyama	-10	-9.0
23	Will Zalatoris	-9	-9.0
39	Jordan Spieth	-7	-9.0
8	Xander Schauffele	-7	-7.0
30	Marc Leishman	-6	-5.0
3	Jon Rahm	-6	-6.0
42	Justin Rose	-5	-3.0
32	Corey Conners	-4	-2.0
18	Patrick Reed	-4	-3.0
29	Cameron Smith	-3	-4.0

Predicted Top 10:

	Player	Actual	Predicted
39	Jordan Spieth	-7	-9.0
23	Will Zalatoris	-9	-9.0
45	Hideki Matsuyama	-10	-9.0
8	Xander Schauffele	-7	-7.0
3	Jon Rahm	-6	-6.0
30	Marc Leishman	-6	-5.0
5	Justin Thomas	0	-4.0
40	Kevin Na	-2	-4.0
29	Cameron Smith	-3	-4.0
41	Brian Harman	-2	-3.0

Figure 13: Model Performances

Feature	Importance
Average Driving Distance	-0.06
Driving Accuracy	-0.001
Greens in Regulation	-0.54
Scrambling	-0.28
Sandsaves	-0.0005
Putting average	41.03

Figure 14: Feature importance

Figure 13 shows the top ten golfers predicted by the model, when we compare this to the actual top ten we can see that the model correctly predicted seven of them. Figure 14

shows the feature importance from the linear regression model. We can see that the only really significant feature was the putting average, however this is likely affected by the unscaled data that was fed to the model for training.

5 Discussion

The main insight gained from the exploratory analysis on the performance metrics is that players on average are driving the ball further than they have ever done before. The players in the 2019 US Masters drove the ball on average 20 meters further than the players who played the same competition in 2003. This has however come at the cost of accuracy. Although the accuracy has fluctuated year-on-year it is on a downward trend since its height of 72.77% recorded in 2010. This was an observation that was mentioned at the start of the paper and is commonly discussed by experts in the media. Byrson "the mad scientist" Dechambeau is renowned for tailoring his game to prioritise distance over accuracy and is currently the longest driver on the professional tour. To see this trend evident in the historical data gathered on the US Masters tournament suggests that more players are following suit as they have begun to realise the benefits in accepting the trade-off between achieving more distance in exchange for less accuracy.

The best performing model implemented was the linear regression model which achieved a mean absolute error of 1.62. This means that for every prediction made by the model it was on average 1.62 shots away from the actual score. This is impressive when we consider the fluctuation in scoring year-on-year due to variables such as weather and course difficulty (i.e., pin positions), which is reflected in the standard deviation of average scores each year, which is 2.89. This model was also able to predict seven of the top ten finishers which shows that it is good at ranking the performances of players based on their performance metrics. As this was the best performing model, it was that model that the feature importance was calculated for. According to the results the putting average was the only significant feature in the dataset. Although it is hard to dispute that putting is one of the most important skills in the game, I would not read too much into the dominance of its importance shown in figure 14 for two reasons. The first reason is that the data was not normalised prior to training the model (for reasons previously mentioned), and as result the putting average feature was likely prioritised in the model. But also, because the putting average is highly related to how well a player performs in a competition (i.e., the less puts a player has, the lower their putting average is, and generally the lower their final score is).

The neural network's performances were very disappointing in my opinion. Having researched ground breaking applications of neural networks, I had high hopes for its performance when applied to this research question. For the neural network approach to have been beaten so heavily by a simple linear regression model was disappointing for me. The best performing instance of the neural network (relu activation function) only managed an mean absolute error of 3.4. Reflecting on the results of the neural networks I have concluded that perhaps there was not enough data passed to the neural networks for them to achieve accurate scores. Also the inability to examine the inner workings of the neural networks made it hard for me to understand where it was going wrong.

6 Conclusions & Future Work

The main objective of this research project was to predict the final scores of players in this year's US Masters based on their performance metrics alone. This was achieved best by the linear regression model which achieved a mean absolute error of 1.6. The relatively strong performance of this model is encouraging, and suggests that there could be great insights gained from future work in this area. For a possible future project I would like to include features based on the weather and course difficulty (i.e. pin positions) to see if the accuracy of the linear regression model could be improved upon. Incorporating such features would also make it more appropriate for normalising the performance metrics, as for example we could normalise the data such as driving distance in accordance with average wind and temperature recorded over the competition. Through normalising the data we would gain more reliable insights into each feature's importance. In a future project I would also like to include more of the performance metrics that are available on the PGA website, such as the club swing speed metrics which are captured through the use of lasers. There are over 100 features currently available on the website of which I only used 6 (due to time constraints). I believe we would then begin to see better performances from the neural network models as they are capable of identifying complex relationships within large datasets (something which linear regression models would struggle with). Finally the original idea for this project was to take the performance metrics from the players' first two rounds in the competition, and to predict their performances over the last two rounds. However, such data was not available on the PGA website, and instead the only data available was aggerate data for the players who completed the full four rounds. This is would be a really interesting project to undertake in the future should the data be made available.

References

[1] Lungarella, M., Iida, F., Bongard, J. and Pfeifer, R., n.d. AI in the 21st Century – With Historical Reflections. 50 Years of Artificial Intelligence, pp.1-8.

[2] Steels, L., 2007. Fifty Years of AI: From Symbols to Embodiment - and Back. 50 Years of Artificial Intelligence, pp.18-28.

[3] Legg, S. and Hutter, M., 2007. Tests of Machine Intelligence. 50 Years of Artificial Intelligence, pp.232-242.

[4] Vernon, D. and Furlong, D., n.d. Philosophical Foundations of AI. 50 Years of Artificial Intelligence, pp.53-62.

[5] Nilsson, N., n.d. The Physical Symbol System Hypothesis: Status and Prospects. 50 Years of Artificial Intelligence, pp.9-17.

[6] Schmidhuber, J., 2007. 2006: Celebrating 75 Years of AI - History and Outlook: The Next 25 Years. 50 Years of Artificial Intelligence, pp.29-41.

[7] Mitchell, M., 2021. Why AI is Harder Than We Think. [online] Youtube.com. Available at: <https://www.youtube.com/watch?v=WF_nm0axBzo&ab_channel=CentrodeCienciasdelaComplejidad>.

[8] Teich, D., 2021. The European Union Is Proposing Regulations For Artificial Intelligence. [online] Forbes. Available at: <https://www.forbes.com/sites/davidteich/2021/04/21/the-european-union-is-proposing-regulations-for-artificial-intelligence/?sh=5985b50e2b46>.

[10] PGATour. 2021. Golf Stat and Records | PGA TOUR. [online] Available at <https://www.pgatour.com/stats.html>

[11] ESPN. 2021. 2021 Masters Tournament - Golf Leaderboard and Results - ESPN. [online] Available at: <https://www.espn.com/golf/leaderboard/_/tournamentId/401243010>

[12] Scikit-learn.org. 2021. scikit-learn: machine learning in Python — scikit-learn 0.24.2 documentation. [online] Available at: <https://scikit-learn.org/stable/index.html>

7 Appendix

7.1 What is Golf?

Golf is an individual sport which is usually played over 18 holes. The aim of the game is to get the golf ball from the tee into the hole in as few shots as possible. Professional competitions (such as the Masters) are played over 4 rounds, and the player with the lowest score at the end of the competition is declared the winner.

7.2 Normalisation

I decided to apply some dimension reduction techniques to the data to see what effects normalisation had on the data. The normalisation technique used was min-max normalisation.

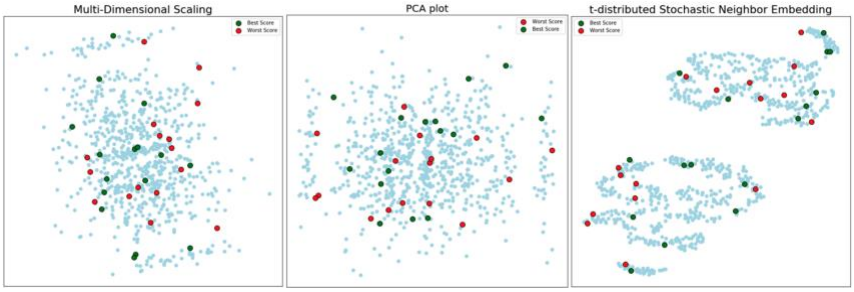


Figure 15: Dimension Reduction without Normalisation.

In figure 15 we can see that the dimension reduction techniques found it hard to distinguish the winners from the worst performers in the non-normalised data.

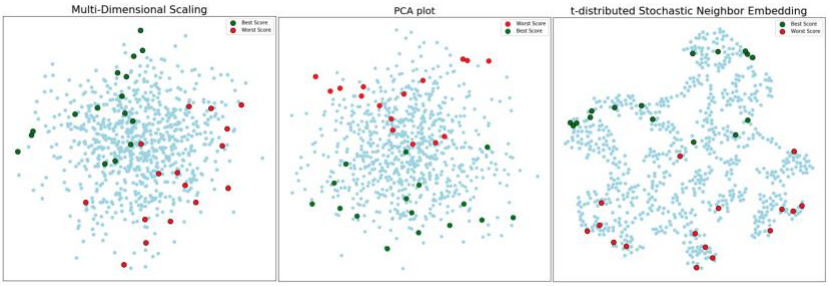


Figure 16: Dimension Reduction after Normalisation.

In figure 16 we can see that the dimension reduction techniques were better able to distinguish the winners from the worst performers in the normalised data.

	Linear Regression	Neural Network (Sigmoid)	Neural Network (Tanh)	Neural Network (Relu)
Mean Absolute Error	2.423	2.52	2.48	2.5
Mean Squared Error	8.46	8.90	8.87	8.96
Root Mean Squared Error	2.91	2.98	2.98	2.99

Figure 17: Model Performances (Normalised Data).

In figure 17 we can see the performances of the models which were trained on normalised data. We can see that the neural network models performed better than their counterparts that were trained on the non-normalised data. However, we can still see that the linear regression models performed slightly better still.

	Player	Actual	Predicted
45	Hideki Matsuyama	-10	-7.0
23	Will Zalatoris	-9	-5.0
39	Jordan Spieth	-7	-6.0
8	Xander Schauffele	-7	-3.0
30	Marc Leishman	-6	-5.0
3	Jon Rahm	-6	-0.0
42	Justin Rose	-5	-2.0
32	Corey Conners	-4	0.0
18	Patrick Reed	-4	-3.0
29	Cameron Smith	-3	-2.0

	Player	Actual	Predicted
45	Hideki Matsuyama	-10	-7.0
39	Jordan Spieth	-7	-6.0
23	Will Zalatoris	-9	-5.0
30	Marc Leishman	-6	-5.0
8	Xander Schauffele	-7	-3.0
18	Patrick Reed	-4	-3.0
40	Kevin Na	-2	-2.0
48	Webb Simpson	-2	-2.0
29	Cameron Smith	-3	-2.0
42	Justin Rose	-5	-2.0

Figure 18: Predicted Top 10 vs Actual Top 10 (Normalised Data).

In figure 18 we can see that the best performing model predicted 8 of the actual top 10 players, which suggest that it is a slightly better model to use for ranking players based on their performance metrics. (Which is to be expected due to the normalisation)

Feature	**Importance**
Driving Distance	2.19
Driving Accuracy	1.92
Greens in Regulation	-12.63
Scrambling	-8.36
Sand Saves	-1.62
Putting Average	12.51

Figure 18: Feature Importance.

Finally in figure 18 we can see feature importance for the linear regression model. Here we can see that the putting average feature is less dominant than it was in the model trained on the non-normalised data. Also the driving distance and accuracy have a greater influence in this model.