Stakeholder Report

Introduction

This assignment is based on a dataset from the PGA tour in Golf, which contains data from 2010-2018. In this stakeholder report, we will walk you through the most essential analyses and results from python, so you can get an understanding of our findings.

We have made a problem statement, which will guide the analyses of our assignment. The problem statement is*:*

Does the golf player's skills have influence on how much money they earn?

To answer the above problem statement, we have chosen six different skills, which we believe should have an influence on the income of the golf players.

In this assignment, we will use Machine Learning techniques to analyse and find patterns in the dataset, but also predict the players' earnings and thereby, see how precise the predictions are compared to the actual earnings during the PGA tour.

Preprocessing

The eight following variables is the foundation of our assignment.

Other variables:	Target variable:	Feature variables:
● Player Name ● Season	● Money	● AverageDistanceAfterDrive ● Total Eagles ● Driving Accuracy ● Birdie Conversion Percentage ● Ball Speed

The dataset consists of 2083 columns. We have chosen the above variables from the 2083, as we believe that these feature variables have a high influence on a golf player's abilities. This we need to know to be able to see a link between a golf players abilities and the money he earns during the PGA tour.

We have dropped the variables, which we didn't find as interesting as the ones in the table above. We have also dropped all of the non numeric values. After these droppings, we had 1667 rows left and eight variables. We renamed the variables' names, removed $ and commas in the values of the money variable and changed our feature values to floats. All of this cleaning was made to make the variables easier and more precise to work with later on.

To make it possible for us to compare the data in the different columns, we needed to scale our data. The idea behind StandardScaler is, that it will transform our data such that its distribution will have a mean value 0 and standard deviation of 1. By doing this on our feature variables, we made it possible for us to compare for example AverageDistanceAfterDrive (yards) to TotalEagles (number).

Variable	Season	Money	Average Distance After Drive	Driving Accuracy	Distance	Ball Speed	Total Eagles	Birdie Conversion
count	1667	1667	1667	1667	1667	1667	1667	1667
mean	2014	1554733	0.00	0.00	0.00	0.00	0.00	0.00
std	2.61	1482411	1.00	1.00	1.00	1.00	1.00	1.00
min	2010	24650	-3.80	-3.64	-2.53	-2.67	-1.66	-3.16
25%	2012	579297	-0.57	-0.69	-0.70	-0.69	-0.72	-0.65
50%	2014	1083787	0.02	0.00	-0.04	-0.06	-0.10	0.00
75%	2016	1955328	0.67	0.69	0.62	0.68	0.53	0.63
max	2018	13030460	2.71	3.06	3.60	3.09	4.27	3.54

The table above shows our chosen variables where the feature variables has been scaled. The table also shows the mean of the different variables. The mean of the golf players' earnings was 1,554,733$ and the golf player that earns the most money during one year of the PGA tour earnt 13,030,460$ and the one that earnt the least in our dataset earnt 24,650$.

In our problem statement, we also wanted to see if other factors than the golf players' ability played any role in their earnings. Therefore, to get a quick overview, we wanted to compare the ten people that earnt the most and the ones that had the best abilities. The different players can be seen below from some of our chosen variables.

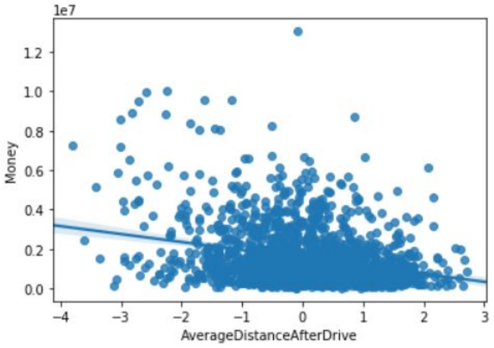
Money	AverageDistance AfterDrive	DrivingAccuracy	BallSpeed	BirdieConversion
1. Tiger Woods 2. *Jordan Spieth* 3. *Justin Thomas* 4. *Dustin Johnson* 5. *Jon Rahm* 6. Bryson DeChambe 7. *Rory McIlroy* 8. Jason Day 9. Hideki Matsuyama 10. Justin Rose	1. Brad Faxon 2. Soren Kjeldsen 3. Mike Weir 4. Jon Curran 5. Jin Park 6.Nick O'Hern 7. Omar Uresti 8. Luke Donald 9. David Lynn 10. Colt Knost	1. Omar Uresti 2. Joe Durant 3. Ryan Armour 4. Tim Clark 5. Thomas Aiken 6. Heath Slocum 7. David Toms 8. Justin Hicks 9. Jerry Kelly 10. Jim Furyk	1. Brandon Hagy 2. Ryan Brehm 3. Andrew Loupe 4. Bubba Watson 5. Tony Finau 6. Keith Mitchell 7. Peter Uihlein 8. Trey Mullinax 9. Charlie Beljan 10.Luke List	1. *Jordan Spieth* 2. *Justin Thomas* 3.Tommy Fleetwood 4. *Dustin Johnson* 5. *Jon Rahm* 6. Brooks Koepka 7. *Rory McIlroy* 8. Phil Mickelson 9. Brandon Harkins 10. Grayson Murray

From this comparison, we saw that the abilities that have the highest effect on the amount of money earnt is BirdieConversion and TotalEagles. We also saw that the other variables shown did not influence money earnt by the players that earnt the highest amount of money.

Linear regression

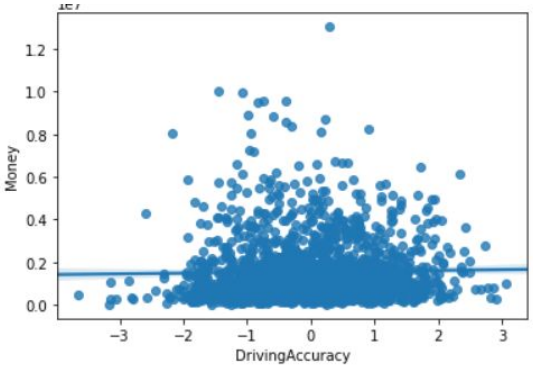
The target variable “Money” is a numerical variable and therefore, we have chosen to make some linear regressions between our target variable and features. These Linear Regressions were made to give a quick insight at the link between the variables, and thereby, get the first insight into what could be the answer to our problem statement.

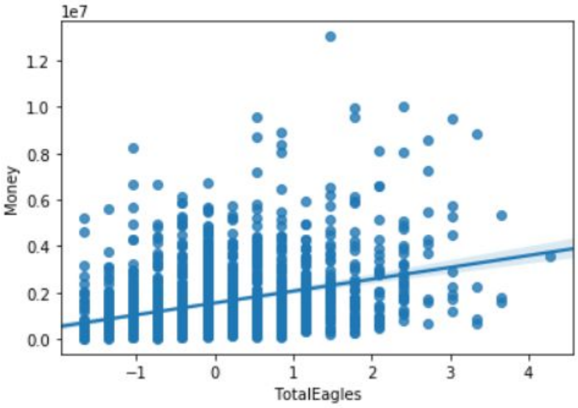
In the following, we will only visualize three of the linear regressions, due to the maximum amount of pages in this report.



Regression of Money and AverageDistance AfterDrive
Outcome: The regression outcome shows that there is a negative link between the average distance to the hole after drive and Money. That makes sense, as the longer distance the player has to the hole after the drive, the more shots he’s going to take to get close to the green.

Regression of Money and DrivingAccuracy
Outcome: There's a minor positive link between DrivingAccuracy and Money. When the golf player has a more accurate drive, he tend to earn more money. However, we expected this to have a bigger impact.





Regression of Money and TotalEagles
Outcome: The more eagles the players make, the more money they earn.
We expected this, since Total Eagles tell us something about how many tries one uses at each hole.

Unsupervised Machine Learning

Unsupervised learning is about learning from test data that has not been labeled, classified or categorized. Unsupervised learning identifies similarities in the data and reacts based on the presence or absence of such similarities in each new piece of data.

We created two clusters that each contained different values, which we calculated the mean of the features.

	Money	AverageDistance AfterDrive	Driving Accuracy	Distance	BallSpeed	Total Eagles	Birdie Conversion
Cluster 0	1,194,433	0.59	0.45	-0.63	-0.63	-0.37	-0.40
Cluster 1	2,032,117	-0.78	-0.59	0.84	0.84	0.49	0.53

By looking at the above table, we can see that Cluster 1 in average have a lower income than the players in Cluster 0.

With regards to the features, we can see that both AverageDistanceAfterDrive and DrivingAccuracy are the only features that in average have higher values in Cluster 1 than Cluster 0. This relates to our EDA analysis, since the graphs illustrated that these two features were the only ones that had either a negative link or almost no link with money.

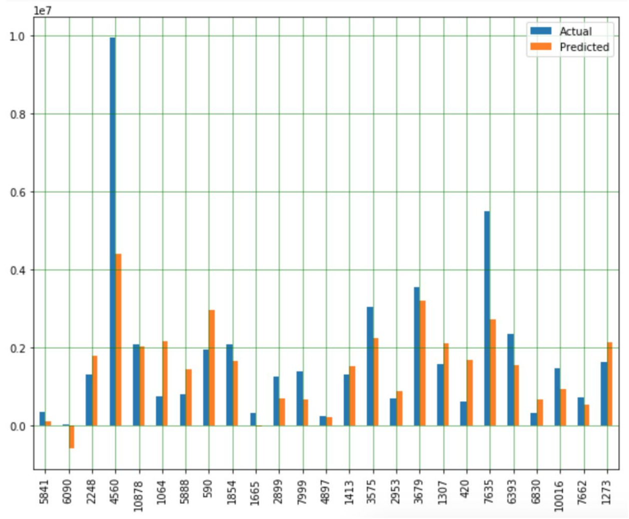
The remaining features have higher values in Cluster 0 and these features had, in the EDA analysis, a positive link with money.

Supervised Machine Learning

Supervised learning is where you have input variables and an output variable and you use an algorithm to learn the mapping function from the input to the output. The goal is to approximate the mapping function so well that when you have new input data that you can predict the output variables for that data.

To be able to answer our problem statement, we wanted to predict the golf players' earnings based on their skills. Therefore, we used different machine learning algorithms to train some of the data which makes it possible for us to predict. Then, we compared the trained data to a testset. From this, we could see that our predictions were 42% accurate. This is not very high, which makes it very difficult for us to predict the precise amount of money which the different golf players earnt on the PGA tour.

The graph below shows 25 random golf players and how much they actually earnt and how much we predicted they should have earned based on their abilities.



Conclusion

Starting this paper, we were looking at the different links between certain skill sets of golf players to see whether or not the chosen skills had any influence on the income for the different players. Furthermore, we wanted to see if other factors than the skill sets had any influence on their income.

In the EDA analysis in the beginning of this paper, we saw that there definitely were links between some of our chosen skills, but also that some of these skills, which we thought had big influence, did not by first glance seem to have much of an impact. Furthermore, we saw that Tiger Woods, who is a very famous golf player, had the largest mean income at the PGA tour, in spite of him not being in the top 10 in any of our chosen categories. To illustrate these findings, we used different illustrations and graphs, to make it easy to read as well as understandable.

In the second part of the assignment, we began to analyse our data using PCA and cluster analysis. For the cluster analysis, we created two different clusters that showed our data divided into two groups. These two groups contained the good golf players and the less good golf players, based on their income, respectively. We were thereby able to conclude that cluster 0 had a smaller income in average than cluster 1.

In the last part, we used techniques within supervised ML and used the train_test_split function to train our dataset into later on making prediction regarding the incomes. This analysis showed us that the actual incomes and the predicted incomes were far apart, and thereby indicating that the methods used for this analysis were not able to create a precise result. This can be due to that fact that a large part of the dataset consisted of NAN values, thereby making it difficult to create precise predictions. We believe this can be due to the fact that some of the golf players were not good enough and thereby not interesting enough for someone to document all data on the players.

The outcome from this paper that is relevant for further use, is that the skills chosen for this analysis do in fact influence the income of golf players as well as other factors eg. fame.

The final conclusion of this paper is that we have chosen to change our career path and start playing golf instead.

If interested the analyses and codes can be found in colab via this link:
https://colab.research.google.com/drive/1Xwu3ZVqFMwdJqjQeHr4OvbdOzgp3JxAX