Capstone Project

Description of the problem and a discussion of the background:

Golf is one of the world's most popular sports, with over 450 million fans all around the globe. Professionally, golf is played all around the world with the most popular league being the Professional Golfer's Association, also known as the PGA. According to our dataset, professional golfers on the PGA tour brought in a total of nearly $300 million in 2018, with top earner Justin Thomas bringing in a healthy $8,694,821. In our entire dataset of golfers from 2010 to 2018, the single biggest earner in a calendar year was Jordan Spieth, who in 2015 brought in a hefty $12,030,465. This came as a result of playing 91 rounds, winning four tournaments, and finishing in the top-10 fourteen times.

While there are large sums of money up for grabs in the form of tournament winnings, the majority of professional golfers' salary is made up of endorsement deals. Looking at Spieth again for example, according to the Forbes list of highest-paid athletes, Spieth made $41 million in 2018. However, our dataset, as well as Golf Digest, indicate Spieth made only approximately $3 million in prize money in 2018. This means Spieth made $38 million from endorsements alone!

It is a lucrative amount of money, but total sum is made up of multiple sponsors. Spieth's biggest sponsor, Under Armour, often sees direct returns on their sponsorship. According to CNN business, Under Armour reported a 25% increase of sales in the business quarter following Spieth's win at The Masters, one of golf's most prestigious tournaments, in 2015. Another study published in Research Gate in 2013 looked at Tiger Wood's impact on the sales of Nike Golf Balls. Tiger Woods signed contracts with Nike worth a lucrative $181 million through 2000-2010. However, this study found that Nike recouped nearly 57% of their investment through the sales of golf balls alone, showing an additional profit of $103 million on Nike's part attributed to golf balls.

Clearly golf is a wealthy sport, which leads into the description of our problem. The aim of this analysis is to figure out what separates top players, those who win tournaments and consistently finish in the top 10, from your average tour professional. This information is relevant to fans of the sport, golfers looking to improve their game, but most of all, large sport and golf brands who are looking to sponsor pro players. This final subset of golf fanatics is who we will be emulating in this analysis. We want to know, as a top golf brand such as TaylorMade, Callaway, Ping or Nike, which golfers we should sponsor for the upcoming season.

Because Golf has such a large global audience, the top golfers on the PGA have a massive fan following. This makes them excellent brand ambassadors for sports brands such as Nike, as fans will want to emulate their style. As seen with Under Armour products and Nike golf balls, sponsoring the right player can have a profound effect on sales. Players can also be sponsored by multiple brands, for example using clubs from a golf brand like TaylorMade, but then wearing Nike shirts, hats and pants. Because of this, we will not be excluding players who already have brand deals in this analysis.

A Description of the data, and how it will be used to solve the problem:

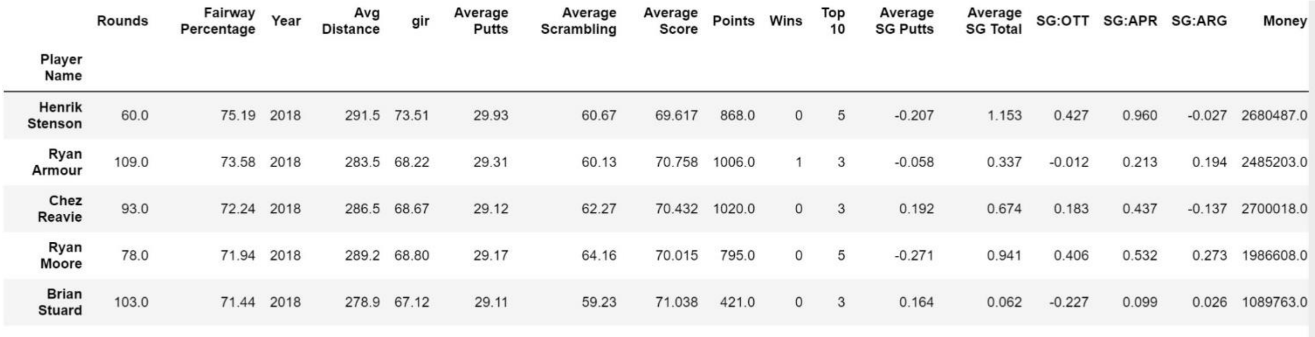
Golf is a game of numbers, which leads to a significant number of metrics being tracked. The data we will be using was obtained from a Kaggle dataset labeled 'PGA Tour data'. This dataset contains information of professional golfers from the PGA Tour between 2010 and 2018. The creator of the dataset built it by scraping the PGA Tour site for key metrics. The dataset contains 18 columns, from which we can identify key features that represent success. The dataset is clearly labeled, with metrics such as 'Wins, Top-10, and Money' being indicators of success.

Between 2010 and 2018, we can see that there were 150 unique players with at least one win, out of a total of 438 players. This indicates around 1 in every 3 golfers on the PGA tour will eventually get a tournament win. But if we dive deeper by examining the number of winners per year, we see that we have between 28 and 33 unique winners per year out of a range of 180 – 195 golfers. That's an average of 83% of players who will go the entire season without a win. While we will compare a dataset of winners vs non-winners, the fact that wins are so rare makes it a difficult target variable in our analysis.

For this reason, the main target variable we will be looking at is Top-10 finishes. In contrast to wins, finishing in the top 10 can be common for the average PGA tour pro. On average there are only 36 golfers a year who will go the whole year without finishing in the top 10, which in a field of around 190 golfers is only about 19%. With about 80% of golfers finishing in the top 10 at least once this year, we see slightly stronger correlation between features if the target variable is finishing in the top-10. This will help us build our machine learning model to predict if a golfer will finish in the top 10.

Data Preparation

Some of the data needed to be cleaned initially. Columns containing 'NaN' (Not a Number) values had to be changed to 0s, because we also want to track when a player gets no wins in a season. The remaining 'NaN' values were dropped, as most pertain to players having played in no tournaments, which were data entries we do not want to use. Data types also had to be changed – 'Top 10' and 'Wins' categories need to be of type 'int'. We also changed the index to Player Name so that individual players could easily be called in the analysis. Once the data had been properly cleaned, we were left with 1674 entries. A snapshot of the data is shown:



	Rounds	Fairway Percentage	Year	Avg Distance	gir	Average Putts	Average Scrambling	Average Score	Points	Wins	Top 10	Average SG Putts	Average SG Total	SG:OTT	SG:APR	SG:ARG	Money
Player Name																	
Henrik Stenson	60.0	75.19	2018	291.5	73.51	29.93	60.67	69.617	868.0	0	5	-0.207	1.153	0.427	0.960	-0.027	2680487.0
Ryan Armour	109.0	73.58	2018	283.5	68.22	29.31	60.13	70.758	1006.0	1	3	-0.058	0.337	-0.012	0.213	0.194	2485203.0
Chez Reavie	93.0	72.24	2018	286.5	68.67	29.12	62.27	70.432	1020.0	0	3	0.192	0.674	0.183	0.437	-0.137	2700018.0
Ryan Moore	78.0	71.94	2018	289.2	68.80	29.17	64.16	70.015	795.0	0	5	-0.271	0.941	0.406	0.532	0.273	1986608.0
Brian Stuard	103.0	71.44	2018	278.9	67.12	29.11	59.23	71.038	421.0	0	3	0.164	0.062	-0.227	0.099	0.026	1089763.0

Feature description

While features such as ‘wins’, and ‘top-10’ are fairly self-explanatory, the number of wins and top-10 finishes a player has in the given year, others, such as ‘Average SG Total’, aren’t quite as self-explanatory. A key metric that pops up frequently is ‘SG’, which refers to ‘strokes gained’. Strokes is the number of shots a player takes to get their golf ball in the hole. ‘Average SG total’ simply put refers to the number of strokes a player was ahead of ‘the field’ – the average of the players playing the same event. As the lowest score wins in golf, being ahead refers to the less strokes a player has taken relative to opponents. Other features which will be examined include average number of putts, average distance of tee-shots, and others.

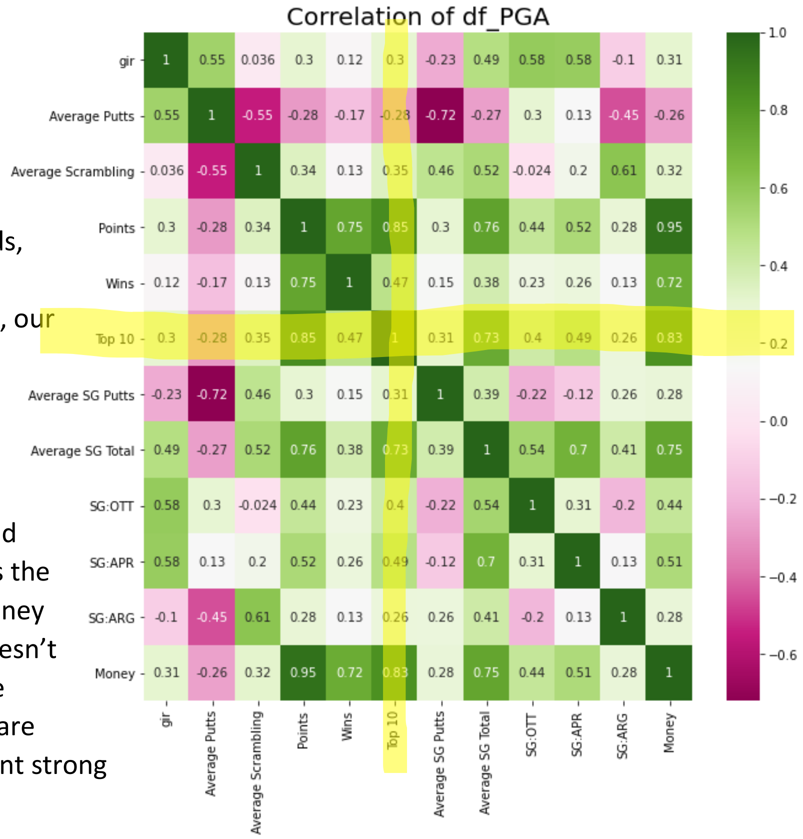
- Index – Player Name. Each row is referring to a specific player in a particular year
- Rounds – The number of golf rounds played in a year
- Fairway Percentage – The percentage of time golfers hit the fairway with their tee shot
- Year – Each row is assigned to a player for a particular year
- Avg Distance – The average distance a player hits the ball off the tee
- Gir – Greens in Regulation %. When a player lands the ball on the green with two strokes under par, this is considered a Green in Regulation. This gives players an opportunity to score under par. Gir here is expressed as a percentage, so the percentage of holes in which a player achieves a green in regulation
- Average Putts – The average number of putts per round
- Average Scrambling – Scrambling is when a player does not hit a green in regulation. A successful scramble is achieved when a player makes par. Expressed as a percentage
- Average Score – Average Score of all the rounds in a year. Ex – a player shoots 68, 69, 70 and 71 in 4 rounds in a year, he has an average score of 69.5.
- Points – Players receive points for placement in tournaments. Total points in a year. These points count towards an end of year cup, the Fed-Ex Cup
- Wins – The number of wins a player achieves in a year
- Top 10 – The number of Top 10 finishes a player achieves in a year
- Average SG Putts – The number of strokes gained by a player through putting alone. Shots gained was previously explained

- Average SG Total – The average of all strokes gained in the year. Includes SG Putts, SG:OTT, SG:APR and SG:ARG.
- SG:OTT – Shots gained off the tee.
- SG:APR – Shots gained from approach shots. Approach shots are typically players second or third shots.
- SG:ARG – Shots gained around the green. Doesn't include putting – mostly revolves around chipping
- Money – The total amount of money earned by a player in that year.

Data Exploration

We started by exploring the correlation between features. For the correlation matrix, we removed features Rounds, Fairway Percentage and Year, as they had no correlation with target variables. As mentioned, our target variable will be Top 10 as a predictor of success.

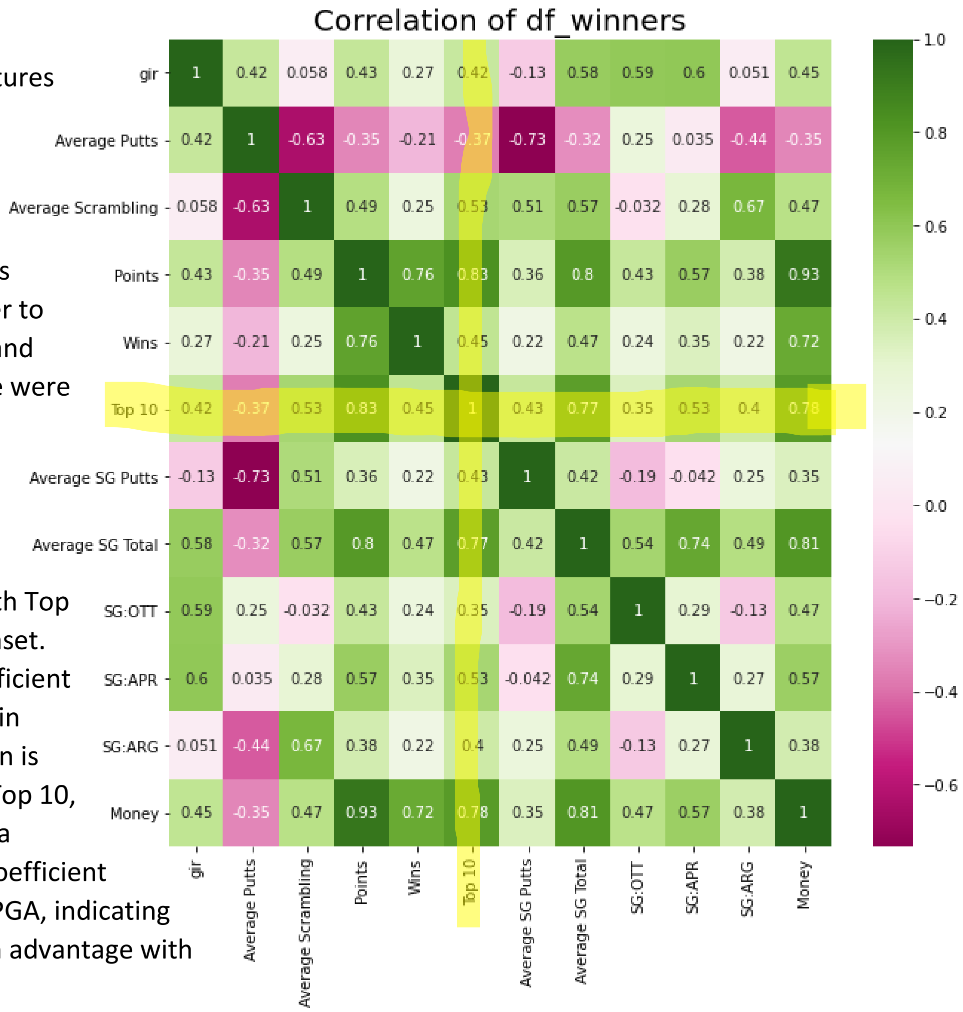
We see moderate correlation with Top 10. The strongest correlations are seen with Points, Money, and Average SG Total. Points and Money have strong correlations with Top 10, as the more Top 10 finishes you receive, the more Money and Points you finish with. This is logical but doesn't help us too much, as none of these features are actually influenced by the players. Rather they are external results that are achieved with consistent strong play.



	gir	Average Putts	Average Scrambling	Points	Wins	Top 10	Average SG Putts	Average SG Total	SG:OTT	SG:APR	SG:ARG	Money
gir	1	0.55	0.036	0.3	0.12	0.3	-0.23	0.49	0.58	0.58	-0.1	0.31
Average Putts	0.55	1	-0.55	-0.28	-0.17	-0.28	-0.72	-0.27	0.3	0.13	-0.45	-0.26
Average Scrambling	0.036	-0.55	1	0.34	0.13	0.35	0.46	0.52	-0.024	0.2	0.61	0.32
Points	0.3	-0.28	0.34	1	0.75	0.85	0.3	0.76	0.44	0.52	0.28	0.95
Wins	0.12	-0.17	0.13	0.75	1	0.47	0.15	0.38	0.23	0.26	0.13	0.72
Top 10	0.3	-0.28	0.35	0.85	0.47	1	0.31	0.73	0.4	0.49	0.26	0.83
Average SG Putts	-0.23	-0.72	0.46	0.3	0.15	0.31	1	0.39	-0.22	-0.12	0.26	0.28
Average SG Total	0.49	-0.27	0.52	0.76	0.38	0.73	0.39	1	0.54	0.7	0.41	0.75
SG:OTT	0.58	0.3	-0.024	0.44	0.23	0.4	-0.22	0.54	1	0.31	-0.2	0.44
SG:APR	0.58	0.13	0.2	0.52	0.26	0.49	-0.12	0.7	0.31	1	0.13	0.51
SG:ARG	-0.1	-0.45	0.61	0.28	0.13	0.26	0.26	0.41	-0.2	0.13	1	0.28
Money	0.31	-0.26	0.32	0.95	0.72	0.83	0.28	0.75	0.44	0.51	0.28	1

It was interesting to see that Top 10 didn't have a strong correlation with Greens in Regulation (0.3), or with Average Putts (-0.28). Logically you would assume that the more opportunities you give yourself to score under par, aka greens in regulation, the better results you would achieve. The same could be said about Average Putts, where generally you could assume that the fewer putts you take a round, the lower your score will be. Neither of these ended up being the case, with weak correlations shown.

As we zero in on potential features to use for predicting, Average SG Total is a clear choice due to its strong correlation with Top 10. The problem with Average SG Total is that it should inherently be correlated with Wins and Top 10s, but that is because it is relative to the rest of the field in each round. So, if a player is consistently gaining strokes on their competitors, then they will be winning tournaments and finishing in the top 10.



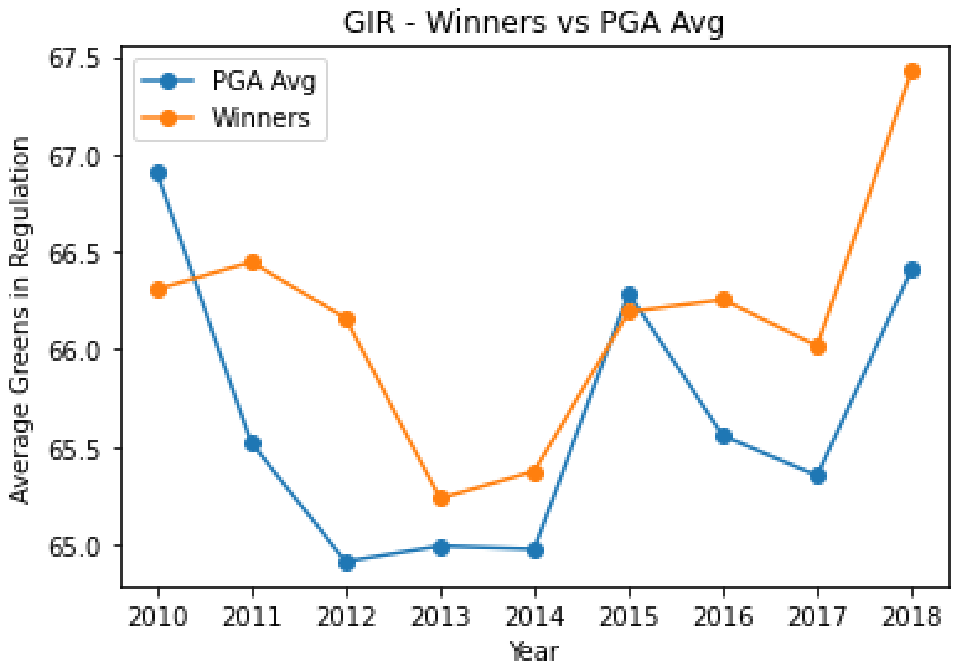
In search of more potential features to be used, as well as stronger correlations, we built a new dataset called df_winners. This dataset only contains data points where players won a tournament in that year. This was done for two main reasons, first in order to look at what the top players are doing and what is bringing them success. Next, we were wary of entries in the full dataset that contained players maybe playing only 5 rounds in an entire year, which would be data entries we don't want to use.

We see stronger correlation with Top 10 when looking at the df_winners dataset. Average SG Total has a correlation coefficient of 0.77 with Top 10, in contrast to 0.73 in df_PGA. Another interesting observation is Average Scrambling's correlation with Top 10, a value of 0.53. SG:APR is also showing a moderate correlation with Top 10 – a coefficient of 0.53. The coefficient was 0.49 in df_PGA, indicating that top players gain slightly more of an advantage with their approach shots.

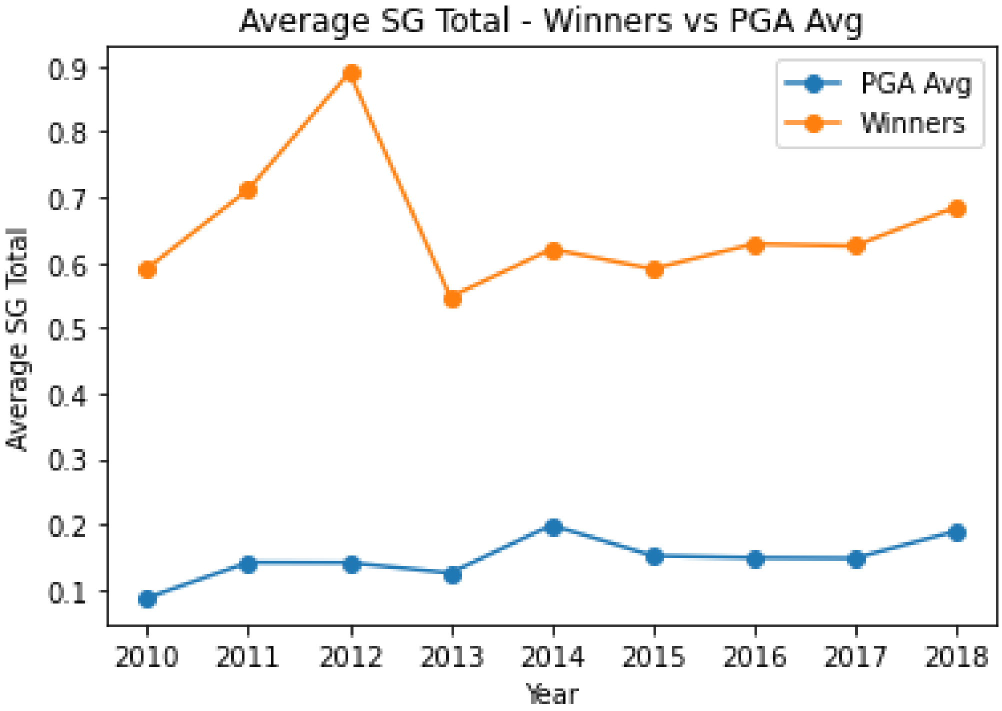
Since we will be using Average SG Total as our predicting feature, it is worth delving into the correlations of Average SG Total itself. Average SG Total shows a strong correlation with SG:APR (0.74), and much stronger than the correlation with SG:OTT (0.54), SG:ARG (0.49) or Average SG Putts (0.42). This tells us that of the shots gained statistic, the shots gained with the approach shots have the most impact on SG Total, and ultimately, Wins and Top 10 finishes. This is also consistent with the correlation seen between Top 10 and shots gained statistic, with SG:APR showing the strongest correlation.

Looking at Average SG Total correlation, we also see moderate correlation with average scrambling, similar to Top 10, but we also see a moderate correlation with greens in regulation. Average Scrambling percentage and Greens in Regulation could be looked at in future models looking to predict success based on controllable factors by the players.

Another idea behind creating a dataset containing only players with a win, was to visualize certain areas where the top PGA pros are much more successful than the average player. Below, we can see a line graph depicting the differences in the number of Greens in Regulation hit.



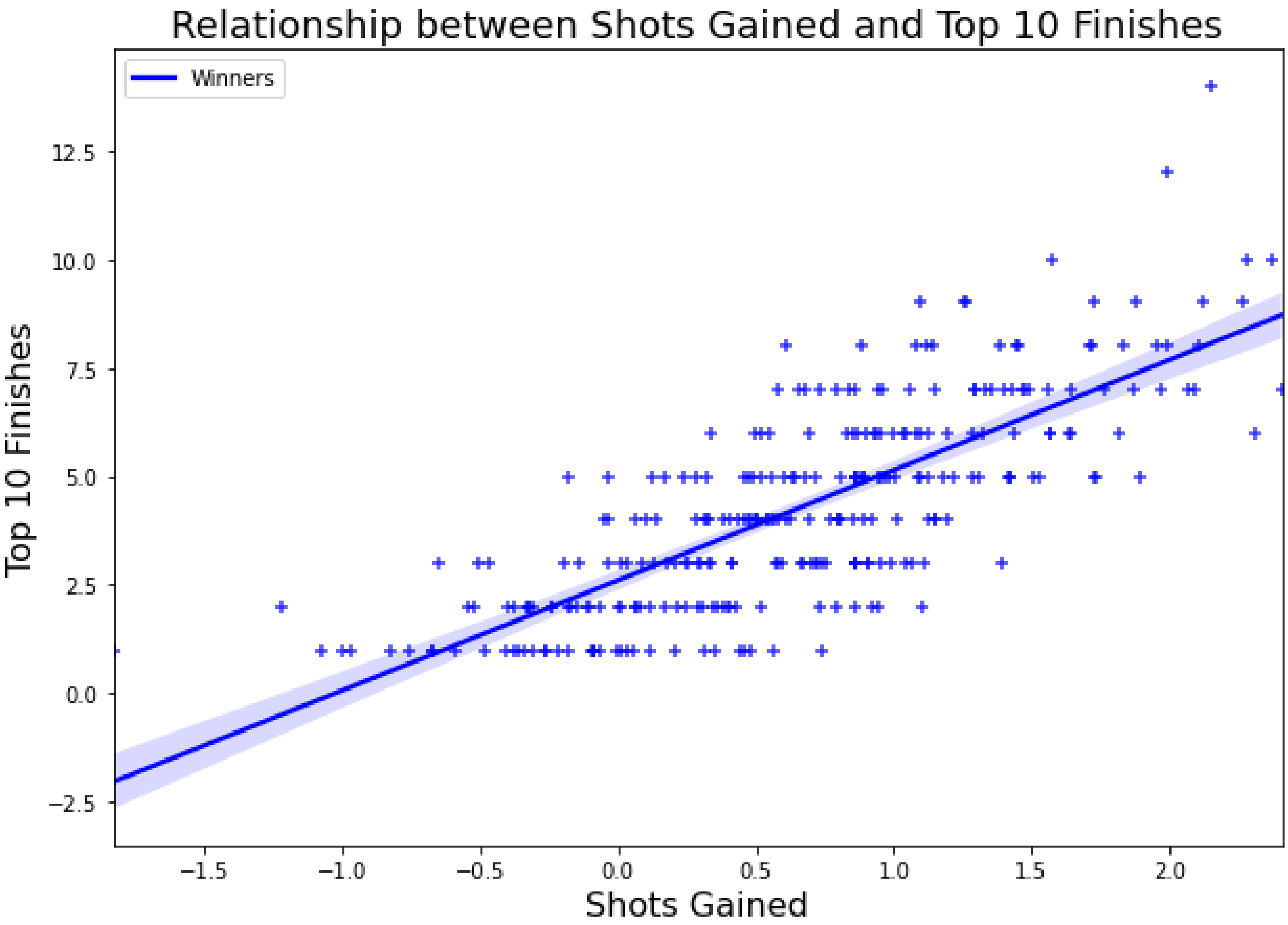
As we can see, there is not much of a difference between your average PGA tour pro and winners in terms of the number of greens hit in regulation, around half a percent. Even, in 2010 and 2015 we can see that the average tour pro actually hits more greens in regulation than the top PGA pros. There were other examples where the average PGA player was on par with the winners, such as avg distance and fairway percentage. For the sake of clarity, we only included gir.



Now in contrast, if we look at a line chart depicting the differences in Average SG Total between your average tour pro and a winner over time, we see a huge disparity between the groups. Average tour pros are averaging anywhere between 0.1 to 0.2 shots gained over a round, while the winners are averaging anywhere between 0.6 to 0.9 shots gained per round. In 2012 this distinction was the starkest, with players winning in this year averaging 0.9 shots gained, versus around 0.15 for the average PGA pro.

There were other features that displayed similar disparity between the winners and the average player. Unsurprisingly, following the trend of Average SG Total, all the shots gained statistics indicated large differences between the winners. We say unsurprising because the shots gained statistic is relative to other golfers in a given round. Thus, it makes sense that the average pro is barely gaining shots on each other, versus when compared with winners. Other features that displayed large disparities included Money, Points, and Average Score. This is also extremely logical as those who win tournaments will post lower average scores, net more total points, and haul in more money.

Model



We decided to run a basic Linear Regression model in order to find a relationship between one of our features, Average SG Total, and our target variable, Top 10 finishes. Average SG Total was chosen because it had the strongest correlation with Top 10 finishes.

We see a moderate to strong positive correlation with these variables, clearly indicating that as golfers gain more shots, they finish in more Top 10s. There were a few outliers, such as Justin Thomas netting 14 Top 10 finishes in 2015. Spieth finished the year with 2.15 shots gained on average.

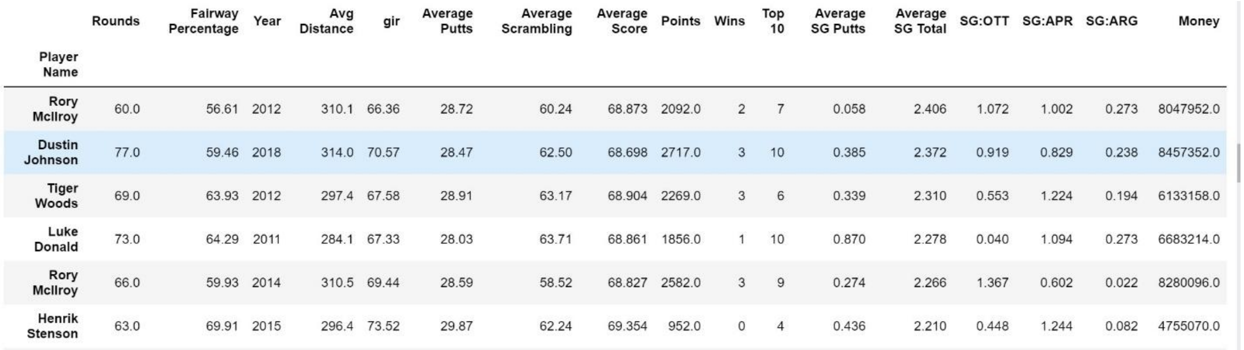
The model has its limitations, however. It's predicting how many times a golfer should finish in the Top 10 based on the number of shots gained but doesn't technically tell us anything about the relationship between features and actually winning a tournament. But as previously mentioned, Top 10 finishes are strongly correlated with wins so it's a good place to start.

Another option which was considered was a decision tree. If we wanted to build a model that would predict if a given golfer would win a tournament based on certain statistics, a decision tree would be a great resource. However, for this analysis we went with the approach of predicting a continuous variable, the number of Top 10 finishes in a year, rather than predicting a categorical variable (Yes or No that a player will win a tournament). If one were to go with this approach, then average scrambling and greens in regulation could be good features to build with.

Conclusion

We analyzed a dataset containing data from PGA pros between 2010 and 2018 and looked at which factors most influence success on the PGA Tour. This information will be relevant to amateur golfers, fans of the sport, but most of all, large sports and golf brands such as Nike or TaylorMade.

We learned that Average SG Total is the best predictor of success, which in our case was the target variable of Top 10 finishes. Below is a snapshot of the golfers who posted the best Average SG Total statistic. As we are in 2020, we are looking for golfers who are recently posting great SG numbers, and we have a clear winner in Dustin Johnson. Johnson posted his best year in 2018, with 2.372 shots gained on average, and ended the year with 10 top 10 finishes. He would be our pick to be the next sponsored athlete by top golf brand.



	Rounds	Fairway Percentage	Year	Avg Distance	gir	Average Putts	Average Scrambling	Average Score	Points	Wins	Top 10	Average SG Putts	Average SG Total	SG:OTT	SG:APR	SG:ARG	Money
Player Name																	
Rory McIlroy	60.0	56.61	2012	310.1	66.36	28.72	60.24	68.873	2092.0	2	7	0.058	2.406	1.072	1.002	0.273	8047952.0
Dustin Johnson	77.0	59.46	2018	314.0	70.57	28.47	62.50	68.698	2717.0	3	10	0.385	2.372	0.919	0.829	0.238	8457352.0
Tiger Woods	69.0	63.93	2012	297.4	67.58	28.91	63.17	68.904	2269.0	3	6	0.339	2.310	0.553	1.224	0.194	6133158.0
Luke Donald	73.0	64.29	2011	284.1	67.33	28.03	63.71	68.861	1856.0	1	10	0.870	2.278	0.040	1.094	0.273	6683214.0
Rory McIlroy	66.0	59.93	2014	310.5	69.44	28.59	58.52	68.827	2582.0	3	9	0.274	2.266	1.367	0.602	0.022	8280096.0
Henrik Stenson	63.0	69.91	2015	296.4	73.52	29.87	62.24	69.354	952.0	0	4	0.436	2.210	0.448	1.244	0.082	4755070.0

Overall the analysis was successful, and we saw a correlation coefficient of 0.77 in our model. These results could be improved, and an alternate model could prove to be a better predictor of success. We hope you enjoyed the analysis.