Capstone Project – Week 1

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.

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'. 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.

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.