Crucial Statistics on the PGA TOUR

Tyler Rozwadowski
Project URL:
github.com/tylerroz/cse482project

ABSTRACT

Golf can generate a lot of statistics. More specifically, professional golf on the PGA TOUR produces large amounts of data every year. Many distinct statistical categories are measured, but there is no use to all the data if it is not interpreted properly. In this project, I sought to find relationships between different statistics on the PGA TOUR. I was curious to find which categories had the most influence on success in golf at the highest level. By setting up the problem as a market basket analysis problem and preforming association mining, I was able to determine which part of a golfer's game is most important. Through this data analysis, I was able to determine that approach shots (iron shots) are the most significant factor in influencing a player's average score, top 10 finishes, and money earned. Other stats such as scrambling percentage and strokes gained on tee shots also have significant influence on players' success on the PGA TOUR.

1. INTRODUCTION

Every year, the best golfers in the world compete in tournaments hosted by the PGA TOUR. In 2008, the PGA TOUR implemented ShotLink, a system that was able to capture more unique golf statistics than ever before [1]. In golf, every shot is important. One stroke can be the difference between winning a tournament or not getting a paycheck. Therefore, it is important to leverage this ShotLink data to determine what categories are most important to the success of a professional golfer. Below are different statistical categories and their definitions. These are some of the key areas that were analyzed:

- Scoring Average: The average score for a player during a given round of golf (lower is better)
- Top 10 Finishes: Number of times a player has placed in the top-10 in a tournament
- Points Earned: Number of FedEx Cup points a player has accumulated throughout a season
- Money Earned: Cumulative amount of prize money a player has won in a given year
- Strokes Gained off the Tee: Average number of shots gained as compared to the rest of the field by driving off the tee
- Strokes Gained on Approach: Average number of shots gained as compared to the rest of the field by hitting approach shots to greens (typically with irons)
- Strokes Gained Scrambling: Average number of shots gained as compared to the rest of the field on shots near the green
- Scrambling Percentage: percentage of time a player makes par after missing the green with his approach shot (higher is better)
- Strokes Gained Putting: Average number of shots gained as compared to the rest of the field through putting
- Driving Distance: Average number of yards per measured drive

Please note that for all "Strokes Gained" statistics, a higher number is better. It can be interpreted as "a player was on average 1.2 shots better than the other players (the field) in this category." These definitions come from the official PGA TOUR website [2].

The goal of this project was to analyze player statistics and determine which ones were most likely to lead to success on the PGA TOUR. For this report, we will define success in the following ways:

- Being a top money earner
- Having a low scoring average
- Finishing in the Top 10/Winning tournaments
- Being a top FedEx Cup points earner

To perform any analysis, professional golf data first had to be collected. After it was determined that there were no suitable "ready-made" data sets available on the internet for this topic, I turned to the PGA TOUR's official website. While they did have a vast amount of data available, it was only able to be downloaded at a cost. Therefore, I decided to collect this data through a web scraper. After much work and tweaking, this crawler was able to download all the necessary data from the TOUR's official web page.

The plan for this report is to perform association mining on the data using the Apriori algorithm. In order to do this, all the collected ratio data must be converted into ordinal item sets.

After all of the data had been collected from the website, it needed to be preprocessed before being analyzed. The first step was to remove attributes that were not necessary to the project. Null or empty values also had to be imputed with the proper values. Once the data had been cleaned, it was converted from rows of ratio attributes to rows of ordinal attributes. Each data point was replaced with a label in comparison to the lower, middle, and upper quartiles. Those item sets were then passed to the Apriori algorithm. After certain parameter and argument tweaks, I was able to obtain the desired result.

After association mining, we see that the "Strokes gained on approach shots" is the statistic that most strongly correlates with success in all measures we defined. Certain other statistical categories, such as "strokes gained off the tee" and higher scrambling percentages also led to greater success on the PGA TOUR.

2. DATA

2.1 Data Collection: Web Crawling

The datasets in this project were collected through a Python web crawler written utilizing the library BeautifulSoup. This crawler allowed me to specifically scrape the attributes I was most interested in for data analysis. All of the collected data in this project comes from the PGA TOUR's official statistics website [2]. In fact, scraping this data may have been the most challenging part of the project because of the way data is formatted and stored on the PGA TOUR's website, where each statistical category has its own page.

While crawling, I created a separate DataFrame for each statistic, which were then all merged using the player's name as a key. This process was repeated for multiple years' worth of data. For this report, I gathered data from 2009 to 2018, giving me 10 years of golf data. This resulted in 1862 records (each for a player in a different year), with 17 attributes (statistical categories) each.

2.2 Data Cleaning

Once all of the data had been collected, I then had to perform preprocessing before I could conduct analysis. Because the data had been crawled as text, Pandas treated them as 'object' type. However, I could not perform statistical analysis on this type. Before converting my data to a numerical form, I first had to do cleaning. Any commas or dollar signs had to be removed. In the cases of 'Top 10', '1st', 'Points', and 'Money', I imputed null or missing values with 0. This was done after reviewing the other ways the data may have been imputed (i.e. using averages) and deciding they would not be suitable for these attributes. After the data had been cleaned, appropriate attributes were converted to the proper numerical type. For example, percentages became *float64*, whereas round number attributes (i.e. number of rounds played) became type *int16*.

2.3 Preprocessing for Association Mining

I then began to preprocess the data for association mining. Because the Apriori algorithm takes item sets as input, I needed to devise a discretization method to convert my records of ratio attributes to ordinal data. I decided that I would take each attribute and categorize it in relation to that statistic's first, second, and third quartiles. So, for each attribute, I would bin it into one of these four categories: *below_q1, between_q1_q2, between_q2_q3, above_q3*.

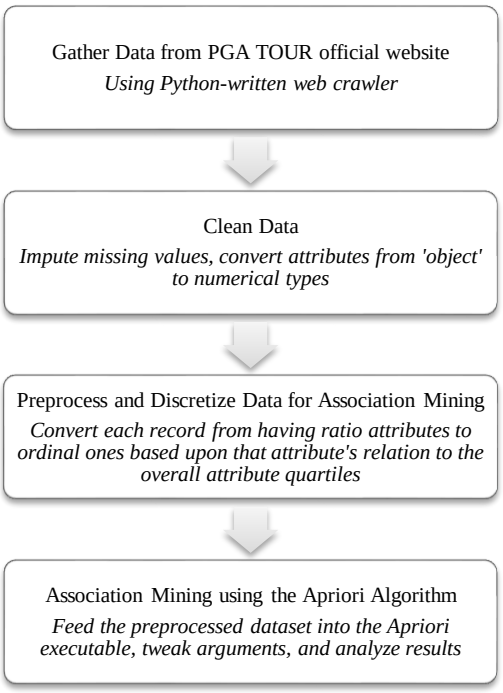
I then iterated through each record in my data set. Each ratio attribute in that record (i.e. a player's 2014 driving distance statistic) was replaced by the appropriate label in relation to that year's quartiles. For example, if the player's value was 272 and the Q1 value for that attribute in 2014 was 274.5, it would be given the label *"below_q1_driving_distance."*

This process was repeated for each numerical record, keeping the size of the data at 1862 records. However, each record now only had 15 attributes, because the *Player Name* and *Year* were no longer important attributes.

Each record no longer contained numerical data, but instead labels for each attribute corresponding to the appropriate quartile bin. After all records were written to a text file (856 KB), this data was now ready to be passed to the Apriori algorithm.

3. METHODOLOGY

Below is a flowchart that shows the way in which data was collected, preprocessed, then analyzed.



The Apriori algorithm/executable used was obtained from Christian Borgelt's webpage [3]. This allowed me to perform association rule mining and determine which PGA TOUR statistics were most closely related (had high support and/or confidence values).

The code written for this project is all contained in the file *golf_data_analysis.ipynb*. The sections for web scraping, data cleaning, preprocessing, and association mining are all clearly distinguished.

4. EXPERIMENTAL EVALUATION

Here we will see how the experiment was conducted and analyze the results.

4.1 Experimental Setup

Running the experiment was simple enough. The Apriori executable can be run on nearly any computer, and simply takes an input and output file name. However, its execution can also be modified with additional arguments. These parameters are able to tweak how the data is output, making the results closer in line with the desired result [4]. After much tweaking and experimentation, I decided on the following parameters:

- -s20: minimum support of 20%
- -c35: minimum confidence of 35%
- -m2n2: restrict maximum and minimum number of supporting items
- -tr: set the target type to association rules

One last argument helped me narrow the results to be what I wanted. The Apriori executable can also be modified to only show certain desired values in the consequent. Because had defined which statistics represented success, I only wanted associations that led to those results. These were my "rules" that I passed to the Apriori algorithm using the "-R" parameter:

```
ante
below_q1_scoring_average cons
above_q3_top10_finished cons
above_q3_points_earned cons
above_q3_1st_place_finishes cons
above_q3_money_earned head
```

The algorithm ran quickly and outputted the results in a file on the local filesystem.

4.2 Experimental Results

The Apriori algorithm outputted 30 rules that had only one item in the head and body of each rule, with the specified support and confidence thresholds. Because it had been filtered, it was easy to visually determine which statistics had the most influence on a golfer being "successful." For example, a rule looked like this:

above_q3_money_earned <- above_q3_strokesgained_approach (25.0269, 52.5751)

This says that if a player is in the 75th percentile for approach (iron) shots, there is approximately 25% support and 52% confidence that they will also be in the 75th percentile of money earned as well. The table below shows the rules with the most significant support and confidence values:

Consequent	Antecedent	Support	Confidence
below_q1_scoring_average	above_q3_strokesgained_approach	25.03	57.73
below_q1_scoring_average	above_q3_scrambling_percentage	25.02	48.71
above_q3_top10_finished	above_q3_strokesgained_approach	25.03	44.42
above_q3_top10_finished	above_q3_strokesgained_tee	24.97	37.63
above_q3_points_earned	above_q3_strokesgained_approach	25.03	44.42
above_q3_points_earned	above_q3_strokesgained_tee	24.97	41.08
above_q3_money_earned	above_q3_strokesgained_approach	25.03	52.58
above_q3_money_earned	above_q3_scrambling_percentage	25.03	41.2
above_q3_money_earned	above_q3_driving_distance	25.03	39.27

Through this, we see that being in the upper quartile of "strokes gained on approach" is the statistic that is most strongly correlated with our measures of success, as defined in the Introduction. Keep in mind that for only scoring average, success will be falling below the first quartile (a lower score is better in golf).

However, we also see that being in the 75th percentile of scrambling can also lead to being in the 75th percentile of scoring average and money earned.

Overall, this project was successful. The results all seem to make logical sense. If you are successful in other aspects of the game, you will most likely be successful overall. However, the goal of this project was to determine which statistics were most likely to lead to success on the PGA TOUR, and that goal was accomplished.

5. CONCLUSIONS

There's a common saying in golf: "Drive for show, putt for dough." Essentially stating that it's more beneficial to be better at putting than at driving the golf ball. However, our analysis proves that to be wrong. In fact, it's neither; but instead iron (approach) shots that are most strongly connected to making money and other successful measures on the PGA TOUR. Although hitting the golf ball may be appealing, it's actually shorter shots that make the biggest difference.

Further research or analysis of these data may find other correlations between different statistics. Other association methods may be attempted to see if they yield the same results as presented in this report.

6. REFERENCES

[1] http://www.shotlink.com/about/background
[2] https://www.pgatour.com/stats.html
[3] http://www.borgelt.net/apriori.html
[4] http://www.borgelt.net/doc/apriori/apriori.html