Classifying PGA Tournament Placement Based On Individual Performance Data

Hayden Spinos

CSCI 7434: Data Mining

Dr. Hong Zhang

May 3, 2023

Contents

Introduction

The Professional Golfers' Association (PGA) Tour is the premier men's professional golf tour that attracts the greatest golf talents that the world has to offer. As in many sports, recent advances in information technology have become increasingly important in understanding and evaluating player performance. The PGA is no exception, as statistics are being captured in higher quantities than ever before.

Background and Motivation

In addition to information technology advances, physical golf hardware advances have greatly impacted how the game is played over the last few years. In the most recent Masters Tournament, new tee boxes had to be constructed to accommodate greater driving distances resulting from golf club and ball technology allowing players to hit farther than ever before. The increased attention to this controversy served as the inspiration for this project.

Glossary

- **Driving Distance:** A term used to reference how far a player can hit the ball off the tee.
- **Hole Par:** A numeric rating of a given hole's difficulty. This number represents the expected number of strokes a player should take to hole the ball from the tee box
- **Stroke:** A single attempt of a player swinging a golf club at their ball. Strokes can be gained through swinging and penalties.
- **Strokes Gained:** A performance metric used to analyze how a player is performing compared to their competitors. Can be tailored to specific aspects of the game: approach shots, putting, driving, etc.…
- **Round:** When a player plays every hole on a given golf course. Typically, 18 holes.

- **Tournament:** A golf competition that is hosted at a specific golf course. Typically consists of 4 rounds in which a percentage of players are cut after 2 rounds.
- **Cut:** Refers to the benchmark score that players must make to continue competing in a tournament after the first two rounds. Historically, players who did not make a cut, would not get paid for playing the first two rounds.
- **Purse:** The pool of prize money that is divided amongst players that make the cut at a tournament.

Objectives

The goal of this data mining project is to analyze PGA tour data on individual player performances from specific tournaments to predict the final finishing position. By testing multiple classification algorithms on a subset of attributes, I will predict the finishing positions of players in specific tournaments.

Data

The dataset used for this project contains 37 columns and has 36,864 instances. It contains records from 2015 through 2022. Each instance contains information pertaining to a specific player's performance at a specific tournament during a specific season. For example, one record may contain relevant information on all four rounds played by Scottie Scheffler's performance at the 2022 Masters Tournament. Many columns contained irrelevant data pertaining to the objective of this project. The decisions made in feature selection will be covered in the preprocessing section.

Data Source: "PGA Tour Golf Data – (2015-2022)" (Kaggle)

Preprocessing

There were several steps needed in processing the data into the desired format. A combination of Python, Excel, and Weka tools were used in processing the data. Below, I've sequentially outlined the steps used in processing the data.

1. Download original 'ASA All PGA Raw Data – Tourn Level.csv' file from Kaggle
2. Python (Jupyter Notebook)
    a. Import data from .csv to Pandas DataFrame
    b. Create subsets of data based on tournament and season. Follow the remaining steps on each set of data. The following data subsets were created:
        - 2022 Masters Tournament
        - 2022 PGA Championship
        - 2021 US Open
        - 2022 Waste Management Phoenix Open
    c. Drop columns that do not contain relevant information:
        - 'Player_initial_last' : Identity of golfer is not relevant to project objective
        - 'tournament id' : Tournament is not relevant to project objective
        - 'player id' : Identity of golfer is not relevant to project objective
        - 'made_cut' : Binary variable that determines if the player made cut or not. Irrelevant in project design since finishing position and strokes indicate if player made the cut or not
        - 'pos' : Final position of player upon stopping of competing (regardless of reason). Redundant information of column 'Finish'

- ‘player’ : Redundant information of column ‘Player_initial_last’. Identity of golfer is not relevant to project objective
- ‘Unnamed: 2’ : No description provided and contained only null values.
- ‘Unnamed: 3’ : No description provided and contained only null values.
- ‘Unnamed: 4’ : No description provided and contained only null values.
- ‘tournament name’ : Tournament is not relevant to project objective
- ‘course’ : Name of course played is not relevant to project objective
- ‘date’ : Date of tournament is not relevant to project objective
- ‘purse’ : Winnings are not relevant to project objective
- ‘no_cut’ : Redundant information of column ‘made_cut’.

d. Drop all rows that contained null values

e. Drop all rows where the player did not finish. Specifically, the rows where ‘Finish’ equals one of the following values:
    - ‘CUT’ : The player did not make the cut
    - ‘WD’ : The player withdrew from the tournament due to injury or illness
    - ‘W/D’ : Same as ‘WD’
    - ‘MDF’ : Player made cut but did not finish. Refers to scenario where too many players qualify at the cut line, so a new cut line is proposed. Players who do not make the second cut line are given note of ‘MDF’, allotted their portion of purse and do not continue playing third and fourth rounds.
    - ‘DQ’ : The player is disqualified from tournament

f. Update ‘Finish’ column to only contain numeric representation of finishing position. For example, if two players tied for 5th place, they each would have a value of T5 which cannot be used for mathematical operations. This step removes the ‘T’ from each instance representing a tie.

g. Cast column ‘Finish’ to be of type ‘int’

h. Move column ‘Finish’ to be the last column in the table. (Done to affect data viewing in Weka explorer)

i. Categorize finishing positions into ranges to be used for classification. The scheme for converting finish positions is described below:

   - If player finished in the top 10 **(‘Finish’ >= 1 and ‘Finish’ <= 10)**, then update value to **1**
   - If player finished in the top 20 **(‘Finish’ >= 11 and ‘Finish’ <= 20)**, then update value to **2**
   - If player finished in the top 30 **(‘Finish’ >= 21 and ‘Finish’ <= 30)**, then update value to **3**
   - If player finished in the top 40 **(‘Finish’ >= 31 and ‘Finish’ <= 40)**, then update value to **4**
   - If player finished in the top 50 **(‘Finish’ >= 41 and ‘Finish’ <= 50)**, then update value to **5**
   - All other finish positions are assigned a value of **6**

j. Output DataFrame to respective csv files

3. Microsoft Excel

a. Delete column A (autogenerated index field by pandas library) from files to avoid index out of bounds errors when importing into Weka
b. Save the updated file

4. Weka Explorer
    a. Open the files in the Weka Explorer
    b. Save the .csv file to a .arff format with the same file name
    c. Apply the unsupervised attribute filter 'NumericToNominal' on the 'Finish' attribute resulting in 6 distinct classes
    d. Select and remove all attributes relating to fantasy points and 'num rounds' (i.e. containing "DKP", "FDP", "SDP")
    e. Save the final dataset to be used in classification.
        - **'PGA_Data_Finished_Encoded.**arff' : 16,606 instances
        - **'2022_Masters_Finish_Encoded.arff'**: 49 instances
        - **'2022_PGA_Championship_Finish_Encoded.arff'** : 74 instances
        - **'2021_US_Open_Finish_Encoded.arff'** : 127 instances
        - **'2022_WM_Open_Finish_Encoded.arff'** : 67 instances

Design

There should now be five separate datasets including 1 total dataset that contains information on all tournaments from 2015 to 2022, and four subsets that contain information on the 2022 Masters Tournament, the 2022 PGA Championship, the 2021 US Open, and the 2022 Waste Management Phoenix Open. The selection of these tournaments was originally intended to be the four most recent majors however, no data was available on the most recent British Open so the Waste Management Phoenix Open was substituted in.

The rest of the project will consist of testing selected classification algorithms on each of the datasets and reporting the accuracy of the model. Each data set will be using a train/test split methodology with a 66% training set size. The goal of each classification algorithm is to classify a player's performance as either a top 10, top 20, top 30, top 40, top 50, or worse finish. The remaining attributes that each algorithm uses to classify are listed below:

- **'hole_par'** : The total par of holes played by the player during the tournament. If a course is a par 72, and the player played 4 full rounds, then 'hole_par' would equal 4*72 = 288
- **'strokes'** : The total number of strokes taken by the player during the tournament
- **'sg_putt'** : Strokes gained putting
- **'sg_arg'** : Strokes gained around the green
- **'sg_app'** : Strokes gained approaching the green
- **'sg_ott'** : Strokes gained off the tee
- **'sg_t2g'** : Strokes gained tee to green
- **'sg_total'** : Total strokes gained

Algorithms

Algorithm implementation for this project will be done in Weka. This is to abstract coding overhead and reduce time needed. In this project, there will be three different classification algorithms being used. They are listed as follows:

- Random Tree
- SVM
- J48 Decision Tree

Evaluation Metrics

Weka provides very detailed evaluation metrics for each algorithm implementation. Each result buffer will be included in the project submission. For the sake of this project however, I will be evaluating the accuracy of the models based on the percentage of correctly classified instances.

Results

	2022 Masters Tournament	2022 PGA Championship	2021 US Open	2022 WM Open	PGA Data
Random Tree	100%	84%	83.7209%	91.3043%	71.8208%
SVM	94.1176%	88%	93.0233%	82.6087%	67.995%
J48	88.2353%	100%	95.3488%	100%	83.9001%

Table 1: Percentage of correctly classified instances per classification algorithm per dataset

	Average
Random Tree	89.7563%
SVM	89.437325%
J48	95.896025%

Table 2: Average percentage of correctly classified instances per classification algorithm across selected datasets

As shown in the tables above, all three classification algorithms perform relatively well on the individual tournament datasets. We can see that J48 performed the best across all the selected tournaments. It also performed the best when running on the entire dataset. By a small margin, SVM performed the worst when comparing the selected tournaments. SVM performed much worse when run on the entire dataset. There is also a notable drop off in the accuracy of the models between the selected datasets and the total dataset.

Discussion

One reason that the classification algorithms performed better on the selected datasets is due to the smaller input size. Due to the nature of the data set, combining the data from multiple tournaments and increasing the dataset may not be helpful. Player performance varies from week to week as well as course difficulty. For this reason, model complexity must be extremely high to accommodate the changing environment of golf performance week to week. By isolating specific tournaments, you limit the variability of data by eliminating differences in course difficulty and player momentum.

Another factor that likely contributed to the accuracy of the models was the decision to classify the performance instances into percentiles rather than trying to predict a numeric finish position. If a player finished 1^{st} in a tournament and the model predicted they would finish 10^{th}, it is counted as a correctly classified instance. This forgiveness incorporated into the model can be treated essentially as a hyperparameter of the model and may be changed by adjusting the categories in the data processing code.

When attempting to classify performance on the entire dataset, the results are potentially lower due to these variables mentioned above. Additionally, by treating the entire dataset as a single tournament, players with multiple entries in the dataset may end up "competing against themselves" in essence.

In future iterations of this project, I would like to explore a dataset that incorporates driving distances with player performances rather than just various strokes gained measures. Increased driving distances in the last few years have majorly affected strategy and in turn may show favor to players who can hit the ball farther. The implications made by such a study could

be used as evidence in board decisions on equipment regulations, golf course design, player evaluations, and betting odds.

Conclusion

In this data mining project, I've shown that it is possible to, with acceptable levels of accuracy, predict which percentile a player will finish in a specific golf tournament. This method of prediction could be extended to use testing data consisting of a player's current strokes gained averages and previous performances on the specified course. The attempt to find a general classification model for predicting a player's finishing position regardless of tournament or golf course proved to be less effective than tournament specific prediction.