2016 NUS Statistics Society
Data Analysis Competition

Introduction

This competition is kindly sponsored by **Nielsen**, a leading global information and measurement company.

nielsen

In this year's data analysis competition, we focus on the application of data mining in sports event prediction. More specifically, you are required to make predictions for the results of 2016 U.S. Open Championship (golf), the second of the four major golf championships. You can work in a team with no more than 4 members.

This competition will give you a chance to put your statistical knowledge into practice. Through this competition, you can expect to learn/hone the essential skills that a good data analyst should acquire, which include but are not limited to,

1) Data collection
2) Statistical model fitting & selection
3) Result presentation

A Glimpse of Golf

*Golf is a club and ball sport in which players use various clubs to hit balls into a series of holes on a course in **as few strokes as possible**.*

*A **hole** is classified by its **par score**, meaning the number of strokes a skilled golfer should require to complete play of the hole.*

*A golfer's **score** is expressed as the difference between the player's number of strokes and the par score.*

There are at least twenty professional golf tours, each run by a PGA (Professional Golfers' Association) or an independent tour organization, which is responsible for arranging events, finding sponsors, and regulating the tour.

The men's major golf championships are the four most prestigious annual tournaments in professional golf. In order of their playing date, the current majors are:

April – Masters Tournament

June – U.S. Open

July – The Open Championship

August – PGA Championship

Goal

You are required to predict the **final scores** of a list of professional golf players in **2016 U.S. Open Championship** (Jun 16-19, 2016). You can use the following software in your analysis: R, Python, Excel/VBA, MATLAB, C++, JAVA.
You should submit your prediction results along with one team report before June 15, 2016.

The list of players are as follows:

1. Spieth, Jordan
2. Mcilroy, Rory
3. Day, Jason
4. Fowler, Rickie
5. Scott, Adam
6. Johnson, Dustin
7. Rose, Justin
8. Watson, Bubba
9. Stenson, Henrik
10. Matsuyama, Hideki
11. Woods, Tiger
12. Oosthuizen, Louis
13. Mickelson, Phil
14. Grace, Branden
15. Kuchar, Matt
16. Garcia, Sergio
17. Walker, Jimmy
18. Reed, Patrick
19. Koepka, Brooks
20. Snedeker, Brandt
21. Furyk, Jim
22. Casey, Paul

Data

You are supposed to find your own datasets.
Basically, you can include in your model whatever predictors you find relevant to the final prediction. There is **NO** requirement on the size of the dataset you may use to train your model.

Note: Most data you find online may not be directly analyzable through standard statistical software, so you should be comfortable with working with different formats (e.g. data cleaning and format conversion)

A sample dataset would contain but not limited to the following 3 parts

1) Player information
    a) Nationality
    b) Swings
    c) Turned Pro (year)

    d) PGA Debut (year)
    e) Birthday
    f) ...
2) Performance of the season (Up to the tournament)
    a) Ranking
    b) Events Played
    c) AVG
    d) Earnings
    e) …
3) Tournament information
    a) Par
    b) Tour it belongs to
    c) Location
    d) Length
    e) Prize Fund
    f) Month Played
    g) Aggregate Record
    h) To Par Record
    i) …

For example, one row of your dataset may look like this

Player	Nationality	Swings	Turned Pro	PGA Debut	Brithday
Scott, Adam	Australia	R	2000	2000	16/7/80

Events Played	AVG	Eearnings	Ranking
9	68.7	$4,442,698	1

Tournament	Par	Tour	Location	Length(yards)
Arnold Palmer Invitational	72	PGA Tour	United States	7381

Prize Fund	Month played	Aggregate Record	To Par Record
$6,300,000	March	264	-23

And the last column, which is <u>the response (scores)</u> we are interested in predicting,

Score
-9

<u>Prediction Results</u>

You are required to sumbit a text file containing your prediction results.
On each line, give the predicted score for that player.
The prediction results should be given in the same order as the list of players.

Report

You are required to submit one report as a team. The report should be within 10 pages (**NOT** including appendix) and explain the steps you took throughout the data mining process, which should at least include the following 4 parts

1) **Data Collection** - Briefly mention the sources of the data you used
2) **Selection** - Explain which attributes you used to build the model; Explain whether you did any transformation of your data and whether you created new attributes from the existing attributes
3) **Model Fitting** - Explain how your built your prediction models
4) **Model Selection** - Explain what model you used in the final predition; Explain why you favored this model against other models

Judging Criteria

Your work will be judged based on both the test MSE of your prediction results and your data mining process as described in your team report.

Note: The BEST PREDICTION PRIZE will be presented to the team with the lowest test MSE, regardless of the quality of the team report. However, if there is a tie in the MSE, the quality of the team report will be considered.

Prize Presentation

The prizes of the competition are as follows:
BEST PREDICTION PRIZE (1 team): $100 / team
THIRD PRIZE (2 teams): $100 / team
SECOND PRIZE (1 team) : $300 / team
GRAND PRIZE (1 team): Surprising Prizes

The results of the competition will be announced around the end of July, 2016
There will be a prize presentation ceremony in the second of August, 2016.
The best 3 teams will have the option to present their works to students, professors and professional data analysts in the prize presentation ceremony.