A Generalization of the Hardy Distribution for Golf Hole Scores

Ryan Kyaw, Department of Mathematics, rkyaw@callutheran.edu

Abstract

In golf, the likelihood of success for each shot is primarily dependent on the result of the previous shot. Hence, Markov Chains may be used to model the progress that a player makes each hole. Modelling and predictions in sports have grown rapidly in recent years, and athletes and teams use predictive models to improve. In "The Hardy Distribution for Golf Hole Scores" by A.H.G.S. Van der Ven, the author uses categorical descriptors to model a golfer's progress as a random walk. To improve the detail and accuracy of Van der Ven's model, our proposed models will include the general states that may occur throughout the golf course. Primarily, we will be focusing on the distance to the hole that a player has on any given shot. We will be using the Beautiful Soup package of the Python programming language to web scrape the PGA Tour data needed to estimate state transition probabilities. Several models will be created and compared to PGA data to analyze golf hole scoring. Ultimately, we will see that our models produce fairly accurate score distributions.

Introduction

Accurate predictions can enhance a fan's experience of a certain sporting event. Predictive modelling can lead to rational insights on what could happen in a particular golf round or tournament. This plays a significant role in sports gambling, which is popular among golfers and golf fans. Furthermore, foresight can lead to advantages for the athletes. Understanding how a certain hole might play can impact a golfer, and this could help him or her create a beneficial strategy.

In my project, I will be creating a predictive model to contribute to the fan's experience along with the golfer's strategy. Using the general positions that a player could be on the course, I will be creating a Markov Chain model that will consider the possibilities of these states being present on a certain hole. My goal is for this model to remove the guesswork that often goes into predicting how difficult or easy a certain course will be for golfers.

Method

- Created multiple Markov Chains corresponding to the different lengths of golf holes
- Each state corresponds to the different positions that a golfer can be in on a golf hole
- 100-150 yard par 3, 400-450 yard par 4, 550-600 yard par 5, etc.
- Used Python and its Beautiful Soup package to web scrape ESPN and PGA Tour data
- This data was used to come up with numerical parameters for each Markov Chain

Results

- 450-500 Yard Par 4 Markov Chain displayed on the right
- Based on our model, scores appear to be better when a player drives his ball on the fairway as opposed to the rough
- Used Chi-Square Goodness-of-Fit testing to compare my Markov Chains to real-world data
- Null Hypothesis: The distribution of scores from the observed data is equivalent to the distribution of scores from the model.
- P-value = 0.093
- Hence, we fail to reject the null hypothesis which is good evidence supporting the accuracy of the models.
- Overall, many of the final models performed well against the real-world data.

- Figure 1: Generalized Markov Chain for all Golf Holes

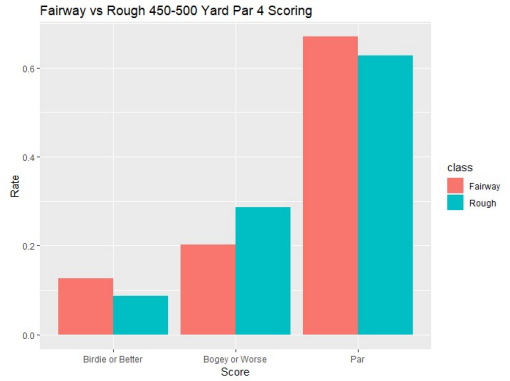
Transition Matrix: $p_{i,j}$ = the probability of reaching state j given that we start at state i

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \end{array} \begin{array}{c} \begin{array}{ccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \end{array} \\ \begin{pmatrix} 0 & p_{1,2} & p_{1,3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_{2,4} & p_{2,5} & p_{2,6} & p_{2,7} & p_{2,8} & p_{2,9} & p_{2,10} & 0 \\ 0 & 0 & 0 & p_{3,4} & p_{3,5} & p_{3,6} & p_{3,7} & p_{3,8} & p_{3,9} & p_{3,10} & 0 \\ 0 & 0 & 0 & 0 & p_{4,5} & p_{4,6} & p_{4,7} & p_{4,8} & p_{4,9} & p_{4,10} & 0 \\ 0 & 0 & 0 & 0 & p_{5,5} & p_{5,6} & p_{5,7} & p_{5,8} & p_{5,9} & p_{5,10} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{6,9} & p_{6,10} & p_{6,11} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1-p_{7,11} & p_{7,11} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1-p_{8,11} & p_{8,11} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1-p_{9,11} & p_{9,11} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1-p_{10,11} & p_{10,11} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

- Figure 2: 450-500 Yard Par 4 Markov Chain

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \end{array} \begin{array}{c} \begin{array}{ccccccccccc} 1 & 2 & 3 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \end{array} \\ \begin{pmatrix} 0 & 0.61 & 0.39 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.36 & 0.3 & 0.15 & 0.1 & 0.07 & 0.02 & 0 \\ 0 & 0 & 0 & 0.56 & 0.2 & 0.12 & 0.07 & 0.03 & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 0.05 & 0.05 & 0.15 & 0.6 & 0.15 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.15 & 0.8 & 0.05 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.84 & 0.16 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.70 & 0.30 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.45 & 0.55 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.03 & 0.97 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

- Figure 3: 450-500 Yard Par 4 Scoring



References

- Ge, Karen. *Expected Value and Markov Chains*. 16 Sept. 2016, www.aquatutoring.org/ExpectedValueMarkovChains.pdf.
- "Golf Stat and Records: PGA TOUR." *PGATour*, www.pgatour.com/stats.html
- "PGA Hole Statistics." *ESPN*, ESPN Internet Ventures, www.espn.com/golf/stats/hole.
- Ross, Sheldon M. *Introduction to Probability Models*. Academic Press, 2010.
- Shasky, Wade. (2015). *Markov Chains and Its Applications to Golf*. Lakehead University.
- Van der Ven, A.H.G.S. "The Hardy Distribution for Golf Hole Scores." The Mathematical Gazette, vol. 96, no. 537, 2012, pp. 428–438., doi:10.1017/s0025557200005052.

Acknowledgements

- A Special Thanks to:
- Dr. Christopher Brown
- Dr. Allan Knox
- OURCS

California Lutheran UNIVERSITY