Applications of Bayesian Inference and Simulation in Professional Golf

Bradley O'Bree[1]
Supervised by Associate Professor Anthony Bedford[1]
[1]RMIT University, Melbourne, Australia

This report details a summary of my AMSI scholarship research project, research findings and experience at the 2012 CSIRO Big Day In Conference in Sydney.

The Project

The project involved developing a simulation model that determines outcome probabilities in professional golf tournaments.

Project Aim

The aim of the project was to develop a simulation model that accurately and efficiently predicted outcomes in professional golf tournaments. Initially, this would involve generating random round scores for each of the competing players at each stage of the tournament.

In this work, we utilised the 2011 US Masters Tournament as a case study. This tournament is regarded as the pinnacle of not only the Professional Golfers' Association (PGA) Tour, but moreover all circuits worldwide. Only the best players in the world are invited to compete in this tournament.

Golf and Simulation Structure

Golf is a club and ball sport with worldwide popularity and origins dating back to at least the 15th century.

The objective is to complete each hole on the golf course in as few strokes (shots) as possible. Each course contains 18 holes, with each hole containing a tee off zone and a green, with the cup being located on the green. Professional players typically score an eagle, birdie, par, bogey or double bogey on each hole. Par refers to the number of strokes a professional player is expected to require to complete a hole, which is primarily judged on the length of the hole. Eagle and birdie refer to completing the hole in two or one strokes below par respectively. Bogey and double bogey refer to completing the hole in one or two strokes above par respectively.

Typically tournaments consist of four rounds, where playing 18 holes constitutes completing a round. After the second round, approximately half of the competing field of players is 'cut' from the tournament, meaning their participation is ceased. Competitors are ranked in an ascending order based on stroke counts, and those with the lower rankings (i.e. higher stroke counts, and therefore worst scores) are the ones who are cut from the tournament.

Figure 1. English golfer Luke Donald was the most prolific money winner in 2011, earning over $6.68m in the calendar year. Photo courtesy of Yahoo! Sports.

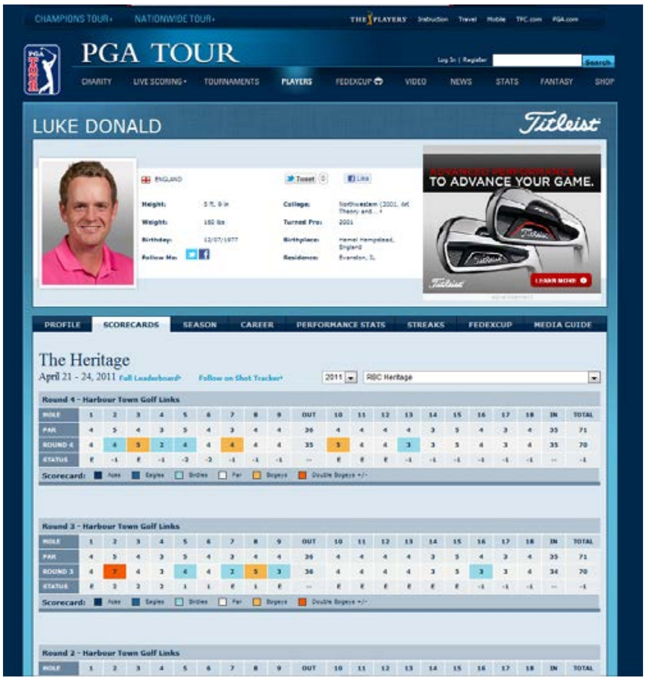
The simulation model followed the structure of the typical tournament (which is also valid for the US Masters). Scores were randomly generated for each player for each round in the tournament, with players ranked following each round (and the appropriate players cut following round two). The player with the lowest score at the end of round four was determined to be the winner.

Data

Data for the simulation model came in the form of round scores from completed professional tournaments. These were typically sourced from pgatour.com, the official website of the PGA Tour (see Figure 2).

This website provides scorecards and profiles for players competing on the PGA, Nationwide and Champions Tours; all of which feature tournaments primarily in the US. Due to time constraints, only 2011 scorecards were acquired.

In consideration of specificity of any inferences from data used, we considered only the round scores of players competing in the tournament we were simulating. This means that while we collected scores from all currently competing players, when simulating the US Masters we only actually used the scores from players competing in the tournament.



LUKE DONALD

ENGLAND

Height:	5 ft. 9 in	College:	Northwestern (2001, Art Theory and...
Weight:	160 lbs	Turned Pro:	2001
Birthday:	12/07/1977	Birthplace:	Hemel Hempstead, England
Follow Me:		Residence:	Evanston, IL

The Heritage

April 21 - 24, 2011 Full Leaderboard Follow on Shot Tracker 2011 RBC Heritage

Round 4 - Harbour Town Golf Links

HOLE	1	2	3	4	5	6	7	8	9	OUT	10	11	12	13	14	15	16	17	18	IN	TOTAL
PAR	4	5	4	3	5	4	3	4	4	36	4	4	4	4	3	5	4	3	4	35	71
ROUND 4	4	4	5	2	4	4	4	4	4	35	5	4	4	3	3	5	4	3	4	35	70
STATUS	E	-1	E	-1	-2	-2	-1	-1	-1	--	E	E	E	-1	-1	-1	-1	-1	-1	--	-1

Scorecard: ■ Aces ■ Eagles □ Birdies □ Par □ Bogeys ■ Double Bogeys +/-

Round 3 - Harbour Town Golf Links

HOLE	1	2	3	4	5	6	7	8	9	OUT	10	11	12	13	14	15	16	17	18	IN	TOTAL
PAR	4	5	4	3	5	4	3	4	4	36	4	4	4	4	3	5	4	3	4	35	71
ROUND 3	4	7	4	3	4	4	2	5	3	36	4	4	4	4	3	5	3	3	4	34	70
STATUS	E	2	2	2	1	1	E	1	E	--	E	E	E	E	E	E	-1	-1	-1	--	-1

Scorecard: ■ Aces ■ Eagles □ Birdies □ Par □ Bogeys ■ Double Bogeys +/-

Round 2 - Harbour Town Golf Links

HOLE	1	2	3	4	5	6	7	8	9	OUT	10	11	12	13	14	15	16	17	18	IN	TOTAL

Figure 2. Screenshot of Luke Donald's Scorecard from The Heritage Tournament in 2011, courtesy of pgatour.com.

Round Score Distributions and Bayesian Inference

We created a Round Score Distribution of Frequencies using the round score data acquired for each of the competing players. These frequencies were standardised to create a probability distribution of round scores. Both the frequency and probability distributions can be approximated using a binomial distribution, the parameters of which are derived using the characteristics of the Round Score Distribution.

Round scores can be randomly generated using this distribution. There is an issue however in that doing this assumes all players are of equal skill, which is obviously not the case. The objective then became to turn the score frequencies into multiple Round Score Distributions.

It was decided the best approach would be to have three different Round Score Distributions, each better suited to players of differing quality. The three distributions would be related to the probability a player will qualify for the cut (not be cut after round two), and if they qualified, the probability they will finish in the top 10.

All round scores were grouped according to the result for the player who scored them. The first grouping contained scores where the player always qualified for the cut. The second grouping contained scores where the player always failed to qualify for the cut. The third grouping contained scores where in at least one tournament they qualified for the cut, but also in at least one tournament they failed to make the cut.

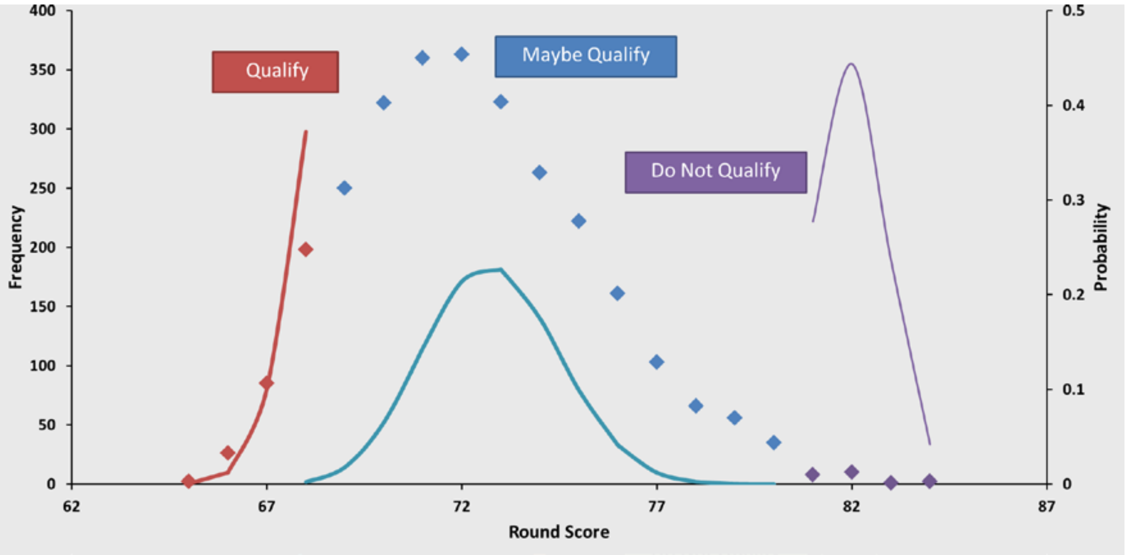


Figure 3. Plot of Round One Round Score Data and fitted Binomial Round Score Distributions

Having three Round Score Distributions allows for quality of the player to be taken into account, as well as keeping the variation in scores for players reasonable.

Players would be randomly assigned to a scoring distribution, based on their inherent ability to actually achieve a score that falls within that distribution. Using historic round score data, we can determine the probability any player will achieve a score that falls into any of the three categories. Players would be randomly assigned to one of the three groups based on these marginal probabilities.

Bayesian inference was used to update these qualification probabilities based on the player's current score.

$$P(Qualify|Score\ x) = \frac{P(x|Q)P(Q)}{P(x|Q)P(Q) + P(x|MQ)P(MQ) + P(x|DNQ)P(DNQ)}$$

At the end of each round the player's probability of falling into each qualification category is updated based on their current score. From here they were again randomly assigned to a group, and a new round score would be randomly generated.

The classification of players is based upon qualifying for the cut prior to rounds one and two while it is based on finishing in the top 10 prior to rounds three and four. The process explained above was repeated for each round in the simulation.

Project Findings

The project findings of note related to the reasonableness of generated scores and comparisons between players. The model adequately produced final scores that were considered reasonable. To demonstrate this, we will look at a sample of the simulation results from the 2011 US Masters.

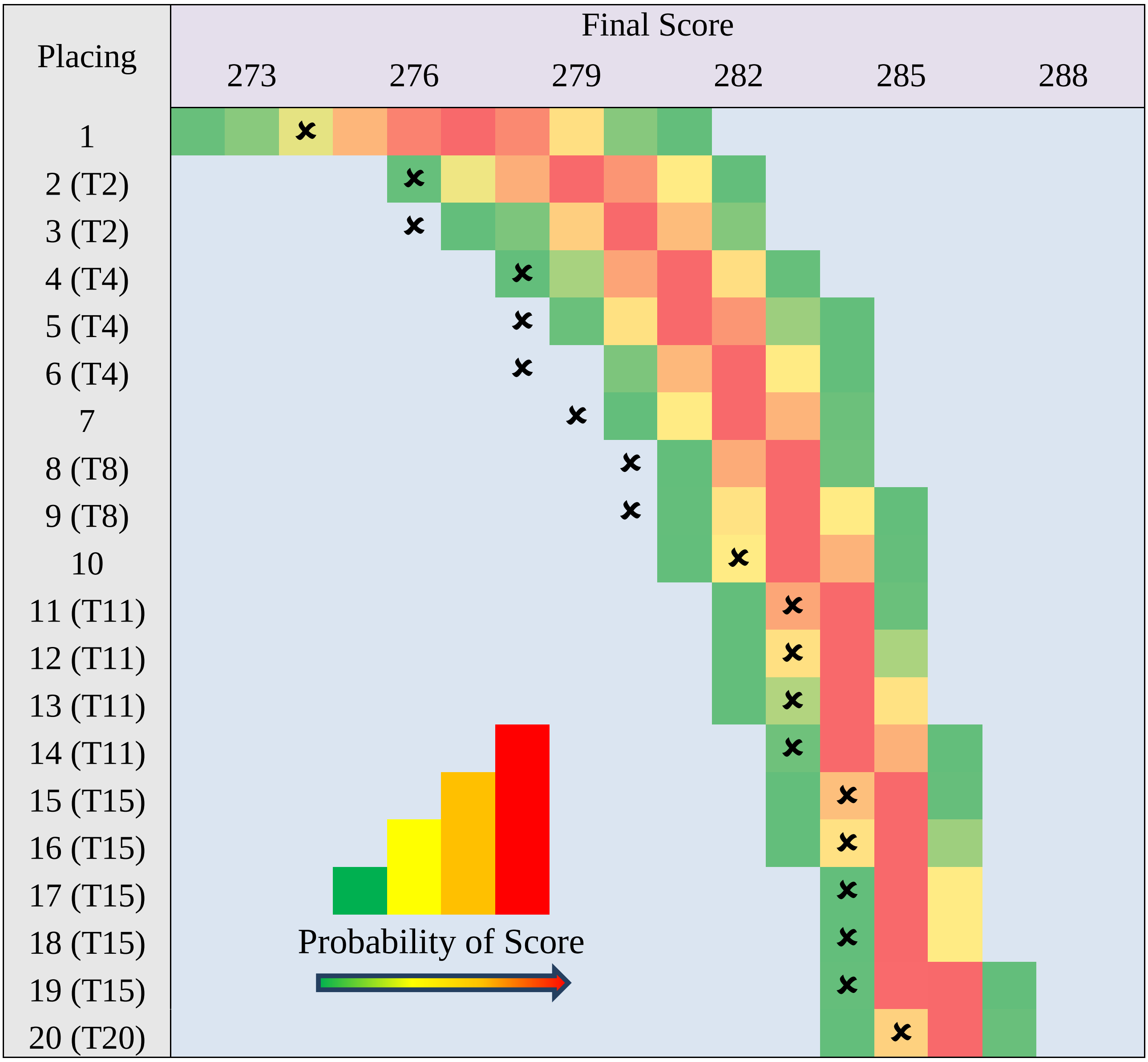


Figure 4. Heat Map of Final Score Distributions by Final Placing (Independent of Player). × indicates the actual scores of each ranked player, (T*i*) indicates a tie between players ranked *i*th on the final leaderboard.

In this simulation, actual final scores tended to be lower than predicted final scores. This difference became more prominent the higher ranked the placing, indicated by greater deviations from the expected finals scores (indicated by the red cells). Such a trend suggests that the Round Score Distribution pertaining to the Maybe Top 10 grouping is reasonably accurate (if only slightly underestimating a player's ability), while that which pertains to the Top 10 grouping is not as accurate (underestimating each player's ability by a measurement of approximately three strokes).

This result can be attributed to individual tournament conditions. Data used to create the Round Score Distributions were based on round scores from other tournaments, each with their own weather and playing conditions specific to the course they are played on. The only variable used to compare scores is course par, meaning courses in locations with notoriously poor weather conditions would influence the score distributions in a negative way. It may be in the case of the 2011 US Masters that conditions favoured the players, with the benefits reflected in lower scores across the board. Results indicate that the best players (those who finished in the top 10) were able to take advantage of the weather to a greater degree than the lesser skilled competitors.

In light of these results, development in the future would need to make the comparisons between courses and their scores a more precise undertaking, with general and time specific factors considered when tournaments are played.

Future Work

To further develop the model in the future, work will be focused on expanding the score database and simulating hole by hole instead of round by round.

The more complete the database of scores, the better we can gauge the quality of each player. Given each player's performance can vary based on current form, forecasting techniques could be used to determine which data for each player is relevant or needed to best represent their current form.

By simulating hole by hole we gain the ability to track a player's form changes throughout a round. The addition of new, extra information better informs us of the player's current abilities or form, which would result in improved accuracy of simulated outcomes, as we can use up to 17 data points when estimating a player's round score.

Experiences

Being selected as an AMSI scholar allowed me to gain experience in researching, as well as presenting research at a conference to students who were in a similar situation as myself.

Conducting research in the university setting gave me a taste of what conducting my own research would be like. At this point in time I'm preparing for my honours year, and the research I conducted as part of this program has been perfect preparation to have a solid honours year at university.

The Big Day In Conference

Participating in the Big Day In Conference was a fantastic opportunity to present my research to students from mathematical backgrounds that differed from my own. It was satisfying to see the appreciation for my work from not only my peers but also their associates and AMSI and CSIRO figures. Attending the conference allowed me to expand my contact list to include students from different universities in different states across Australia, which ultimately instills confidence in future research aspirations I may have in unfamiliar fields of mathematics others are currently developing expertise in. I see this conference as a gateway to finding students with similar passions and aspirations, and am thankful I was given the opportunity to participate.

References

Shepherd, R, Bedford, A & Schembri, A 2010, ‘A Two-Stage Simulation to Predict Medalists in Pistol Shooting’, Proceedings of the tenth Australasian Conference on Mathematics and Computers in Sport, A Bedford & M Ovens (Eds), pp.213-220.