Introduction

Mark E. Glickman* and Jeff Sonas

Introduction to the NCAA men's basketball prediction methods issue

DOI 10.1515/jqas-2015-0013

We are delighted to present a special collection of articles in this issue of the Journal of Quantitative Analysis in Sports (JQAS) on predicting game outcomes in the NCAA men's basketball tournament, more commonly known as "March Madness." The NCAA tournament is a single elimination knockout event that begins with 64 teams after four teams have been eliminated through a play-in round a few days before the tournament begins. The idea for this special series of papers arose from a prediction competition we hosted on the Kaggle website (along with Will Cukierski of Kaggle and Ed Feng of *The Power Rank*) in which contestants were required to predict the results of the 2014 NCAA men's basketball tournament. The competition was sponsored by Intel and came with a $15,000 cash prize for the first-place finisher. It became evident from Kaggle forum discussions that some contestants were developing innovative analytical approaches for making predictions. We felt that inviting submissions for JQAS based on these innovative methods would lead to a fascinating set of articles. The culmination of this process is on display in the current JQAS issue.

Kaggle has been hosting prediction contests since its inception in 2010. Kaggle contests involve building prediction models or algorithms for specific data questions, often posed by companies that sponsor the contests. For typical contests, the data are partitioned at random into a training set and a test set, the latter omitting the variable or variables the contestant is tasked with predicting. Our competition had a different structure. For the 2014 NCAA tournament prediction contest on Kaggle, contestants were provided with both regular season and NCAA tournament final scores from the previous 18 years, including the final scores from the 2013 to 2014 regular season upon which to develop prediction approaches. The data provided also included the location (home/away/neutral) of each game during the regular season contests. Contestants could (and did) supplement these game results with other available data, such as expert forecasts and geographic data to assess the impact of travel schedules. The competition required contestants to submit their predictions prior to the start of the 2014 tournament. In contrast to the usual Kaggle contest in which the withheld data in the test set are already recorded, the NCAA contest involved predicting the future. Thus the usual concerns with leakage in data prediction contests (Rosset et al. 2010; Kaufman et al. 2012) did not exist for the NCAA tournament prediction Kaggle competition.

Competitions for predicting the NCAA tournament, both informal and prize contests, are abundant. Most contests, including ones run by Yahoo, ESPN and the 2014 Quicken Loans $1 billion prize contest sponsored by Warren Buffet, provide the tournament bracket with the first-round match-ups and require that contestants select the winning teams during each round of the tournament. These bracket competitions are scored typically by counting the number of correct advances in the tournament, or some weighted version that rewards predicting teams that advance far into the tournament. The Kaggle contest took a different approach. Contestants were instructed to provide winning probability predictions for all $\binom{68}{2}=2278$ possible pairwise match-ups of the 68 teams announced on Selection Sunday (March 16, 2014), with the probabilities determined from the contestants' models and algorithms. Over the next four days, contestants could upload entries. Because at the end of the 4-day period four teams were already eliminated in the play-in round, the evaluation of the predictions was based on the results of the 63 games played in the knockout tournament by the 64 remaining teams. Let $i=1, \ldots, 63$ index the

***Corresponding author: Mark E. Glickman,** Center for Healthcare Organization and Implementation Research, Edith Nourse Rogers Memorial Hospital (152), Bldg 70, 200 Springs Road, Bedford, MA 01730, USA, Tel.: +(781) 687-2875, Fax: +(781) 687-3106, e-mail: mg@bu.edu; and Department of Health Policy and Management, Boston University School of Public Health, and the Center for Healthcare Organization and Implementation Research, a Veteran Administration Center of Innovation
Jeff Sonas: Sonas Consulting, 22450 Charlene Way, Castro Valley, CA 94546, USA

games in the tournament. Let y_i be 1 if the first team in a pair wins game i and 0 otherwise, and for contestant j let p_{ij} be the reported probability that the first team in the pair wins game i. The binomial deviance of game i by contestant j, a measure of predictive discrepancy, is given by

$$d_{ij} = -(y_i \log p_{ij} + (1-y_i)\log(1-p_{ij})) \quad (1)$$

Then the predictive fit criterion for contestant j was defined for the competition as

$$\text{LogLoss}_j = \frac{1}{63}\sum_{i=1}^{63} d_{ij} \quad (2)$$

Smaller values of LogLoss_j in equation (2) indicate a better overall set of predictions for the competition. The first-place finisher was the contestant with the lowest value of LogLoss_j. It is worth noting that the formula in (2) is linearly related to a predictive version of a binomial deviance function, which itself is monotonically related to the joint probability of the 63 game outcomes. Thus smaller values of LogLoss_j correspond to higher estimated joint probabilities of the tournament game results.

A total of 243 contestants/teams entered the competition submitting a combined 428 entries, with each contestant being allowed to submit up to two sets of predictions. Although the tournament spanned 19 calendar days, more than half of the 63 games were played during the first two days of play (when 32 teams were eliminated), so the contest was already halfway complete at that point. Two contestant teams, including the eventual winning team, at that point had average binomial deviance scores of 0.426 and 0.429 with an additional five contestants/teams having average scores between 0.43 and 0.44. As the tournament progressed, more games were played between similarly-seeded college teams. Typically these match-ups were associated with probability predictions close to 0.5 so that the average binomial deviance scores increased. The deviance scores may have also increased due to what was perceived as an unusual number of upsets in the early rounds. After 48 games, the leading score was 0.4753 and fewer than 20 participants had an average binomial deviance below 0.5. With three games remaining, the best score was slightly above 0.51. The eventual winner finished with a score of 0.52951, followed by four other contestant teams having scores between 0.53 and 0.55.

It is worth noting that the top contestant/teams in the Kaggle competition had impressive performances. Viewing a probability >0.5 for each match-up as predicting that team to win, and a probability equal to 0.5 as worth half a win for each team, eight different contestants/teams successfully predicted 46 or more of the 63 match-ups correctly, a 73% success rate. By comparison, a review of publicly-available rating lists, published before the tournament from more than 60 different expert organizations/individuals indicated that all of those 60+ expert sources had misclassified at least 18 of the tournament outcomes in their ratings (Massey 2014), suggesting that none of them would have successfully predicted as many as 46 of the 63 games correctly.

The five articles in the special series demonstrate various and distinct approaches to predicting NCAA tournament game probabilities. Lopez and Matthews, the ultimate winners of the 2014 Kaggle competition, demonstrated that a fairly simple model can result in strong predictive accuracy. They constructed two logistic regressions, one based on the Las Vegas point spread and the other based on Pomeroy's team performance statistics. Their final ensemble models were different weighted averages between these two logistic regressions. Given their top performance in the Kaggle contest, they examined the role of luck through a simulation study. Bornn et al. also used team performance statistics by Pomeroy in their models, but also included those of Massey, Moore, ESPN and the "Rating Percentage Index." They then fit an ensemble model consisting of three logistic regressions, a stochastic gradient boosted tree model, and a neural network model to predict NCAA game outcomes. Recognizing that their performance statistics to predict previous years' NCAA tournament incorporated data from the tournaments themselves and resulted in overfit models, they provided details on converting these variables into ones that act as though the tournament data were removed. Hoegh et al. used a linear model for score differences, but included novel game-specific adjustments to account for non-transitivity of team strengths. Specifically, they included a model term that accounts for the over- or under-performance by each team in a match-up through the outcomes against teams with the similar background covariates. Ruiz and Perez-Cruz developed a model in which the final scores of a game are modeled as Poisson-distributed as a function of both offensive and defensive parameters. Their model has strong connections to a commonly used model to estimate soccer team strengths. They fit their model using a variational inference algorithm to approximate the posterior distribution, and constructed predictions based on their model fit. Finally, Gupta, who submitted his manuscript independently of the Kaggle comptition, wrote about more traditional bracket predictions. Gupta developed a "dual-proportion" model in which the likelihood function weighted conference tournament games more heavily than regular season games, and accounted for individual games resulting in blowouts. We hope you enjoy these articles.

References

Kaufman, S., S. Rosset, C. Perlich, and O. Stitelman. 2012. "Leakage in Data Mining: Formulation, Detection, and Avoidance." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6:15.

Massey, K. 2014. "College Basketball Ranking Composite," URL http://www.masseyratings.com/cb/arch/compare2014-19.htm. Accessed on January 24, 2015.

Rosset, S., C. Perlich, G. Świrszcz, P. Melville, and Y. Liu. 2010. "Medical Data Mining: Insights from Winning Two Competitions." *Data Mining and Knowledge Discovery* 20:439–468.