Michael J. Lopez* and Gregory J. Matthews

Building an NCAA men's basketball predictive model and quantifying its success

Abstract: Computing and machine learning advancements have led to the creation of many cutting-edge predictive algorithms, some of which have been demonstrated to provide more accurate forecasts than traditional statistical tools. In this manuscript, we provide evidence that the combination of modest statistical methods with informative data can meet or exceed the accuracy of more complex models when it comes to predicting the NCAA men's basketball tournament. First, we describe a prediction model that merges the point spreads set by Las Vegas sportsbooks with possession based team efficiency metrics by using logistic regressions. The set of probabilities generated from this model most accurately predicted the 2014 tournament, relative to approximately 400 competing submissions, as judged by the log loss function. Next, we attempt to quantify the degree to which luck played a role in the success of this model by simulating tournament outcomes under different sets of true underlying game probabilities. We estimate that under the most optimistic of game probability scenarios, our entry had roughly a 12% chance of outscoring all competing submissions and just less than a 50% chance of finishing with one of the ten best scores.

Keywords: basketball; NCAA; predictive modeling; simulations; tournament.

DOI 10.1515/jqas-2014-0058

1 Introduction

Each March, more than an estimated 50 million Americans fill out a bracket for the National Collegiate Athletic Association (NCAA) men's Division 1 basketball tournament (Barra 2014). While paid entry into tournament pools is technically outlawed, prosecution has proved rare and ineffective; an estimated $2.5 billion was illegally wagered on the tournament in 2012 (Boudway 2014; Tsu 2014).

Free tournament pools are legal, however, and Kaggle, a website that organizes free analytics and modeling contests, hosted its first college basketball competition in the early months of 2014. Dubbed the "March Machine Learning Mania" contest, and henceforth simply referred to as the "Kaggle contest," the competition drew more than 400 submissions. We submitted two entries, detailed in Section 3, one of which earned first place in the 2014 contest.

This manuscript describes how we combined standard statistical methods with relevant data to build a men's college basketball prediction model. Formally, building on themes first suggested by Carlin (1996), we blend information from the Las Vegas point spread with team-based possession metrics by using a weighted average of the predictions generated from logistic regression models. The success of this algorithm reinforces longstanding themes of predictive modeling, including the benefits of combining multiple predictive tools and the importance of using the best possible data.

Next, we use simulations to estimate the fraction of luck that our entry required to outperform competitors, using different underlying sets of probabilities for each conceivable 2014 tournament game. If one of our two submissions contained the exact win probabilities, we estimate that submission increased our chances of winning by about a factor of 50, relative to if the contest winner were to have been randomly chosen. Despite this advantage, due to the large number of contest entries, that submission would have had no more than about a 50–50 chance of finishing in the top 10, even under the most optimal of conditions.

This paper is laid out as follows. Section 2 describes the data, methods, and scoring systems pertinent to predicting college basketball outcomes. Section 3 details our submission, and in Section 4, we present simulations with the hope of quantifying what proportion of luck was needed for our entry to succeed. Section 5 summarizes and concludes.

***Corresponding author: Michael J. Lopez,** Skidmore College – Mathematics and Computer Science, 815 N. Broadway Harder Hall, Saratoga Springs, New York 12866, USA, Tel.: +9784072221, e-mail: mlopez1@skidmore.edu
Gregory J. Matthews: Loyola University Chicago – Mathematics and Statistics, Chicago, Illinois, USA

2 NCAA tournament modeling

2.1 Data selection

Two easily accessible sets of predictors for NCAA basketball tournament outcomes are information from prior tournaments and results from regular season competition. Regular season data would generally include information like each game's home team, away team, location, and the final score. For tournament games, additional information would include each team's seed (No. 1 through No. 16), region, and the distance from each school's campus to the game location.

The specific viability of using team seed to predict tournament success has been examined extensively; see, for example, Schwertman, Schenk, and Holbrook (1996) and Boulier and Stekler (1999). In place of team seeds, which are approximate categorizations of team strengths based mostly on perceived talent, we supplemented regular season data with two types of information that we thought would be more relevant towards predicting tournament outcomes: (1) the Las Vegas point spread and (2) team efficiency metrics.

2.1.1 The Las Vegas point spread

One pre-game measurement available for the majority of Division 1 men's college basketball games over the last several seasons is the Las Vegas point spread. This number provides the predicted difference in total points scored between the visiting and the home team; a spread of –5.5, for example, implies that the home team is expected to win by 5.5 points. To win a wager placed on a 5.5 point favorite, one would need that squad to win by six points or more. Meanwhile, a bet on the underdog at that same point spread would win either if the underdog loses by 5 points or fewer, thereby covering the spread, or if the underdog wins. In principal, the point spread accounts for all pre-game factors which might determine the game's outcome, including relative team strength, injuries, and location.

Rules of efficient gambling markets imply that, over the long run, it is nearly impossible to outperform the point spreads set by sportsbooks in Las Vegas. A few landmark studies, including Harville (1980) and Stern (1991), used data from National Football League (NFL) games to argue that, in general, point spreads should act as the standards on which to judge any pre-game predictions. While recent work has looked at gambling markets within, for example, European soccer (Constantinou, Fenton, and Neil 2013), the Women's National Basketball Association (Paul and Weinbach 2014), the NFL (Nichols 2014), and NCAA men's football (Linna et al. 2014), most research into the efficiency of men's college basketball markets was produced more than a decade ago. Colquitt, Godwin, and Caudill (2001), for example, argued that, overall, evidence of market inefficiencies in men's college basketball was limited. These authors also found higher degrees of efficiency in betting markets among contests in which a higher amount of pre-game information was available. Paul and Weinbach (2005) highlighted inefficiencies with respect to larger point spreads using men's college basketball games played between 1996–1997 and 2003–2004 and found that placing wagers on heavy underdogs could be profitable. Lastly, Carlin (1996) modeled tournament outcomes from the 1994 NCAA season, finding that the point spread was among the easiest and most useful predictors.

As a result of our relative confidence in the efficiency of men's basketball markets, we extracted the point spread from every Division 1 men's basketball contest since the 2002–2003 season using www.covers.com and linked these results to a spreadsheet with game results.

2.1.2 Efficiency metrics

One aspect lost in the final scores of basketball games is the concept of a possession. Given that NCAA men's teams have 35 seconds on each possession with which to attempt a shot that hits the rim, the number of possessions for each team in a 40-minute game can vary wildly depending on how quickly each squad shoots within each 35-second window. In the 2013–2014 season, for example, Northwestern State led all of Division 1 with 79.3 possessions per game, while Miami (Florida) ranked last of the 351 teams with 60.6 per game (TeamRankings 2014). As a result, it is not surprising that Northwestern State scored 20.1 more points per game than Miami on average, given the large discrepancy in each team's number of opportunities (TeamRankings 2014). As score differentials will also be impacted by the number of possessions in a game, offensive and defensive per-possession scoring rates may provide a greater insight into team strength relative to the game's final score.

Several examples of possession-based metrics can be found on a popular blog developed by Ken Pomeroy (www.kenpom.com). Pomeroy provides daily updated rankings of all Division 1 teams using offensive and defensive efficiency metrics that he adjusts for game location and opponent caliber. The larger umbrella of possession-based statistics, of which Pomeroy's metrics fall under, are summarized by Kubatko et al. (2007).

Pomeroy's website provides team-specific data for all seasons since 2001–2002. We extracted several different variables that we thought would plausibly be associated with a game's results, including a team's overall rating and its possession-based offensive and defensive efficiencies. These metrics provide a unique summary of team strength at each season; one downside, however, is that the numbers that we extracted were calculated *after* tournament games, meaning that postseason outcomes were included. As a result, fitting postseason outcomes using Pomeroy's end-of-postseason numbers may provide too optimistic a view of how well his numbers produced at the end of the regular season would do. Given that there are many more regular season games than postseason games, however, we anticipated that changes between a team's possession-based efficiency metrics, as judged at the end of the regular season and at the end of the postseason, would be minimal. Relatedly, Kvam and Sokol (2006) found that most of the variability in a team's success during the tournament could be explained by games leading up to mid-February, implying that games at the end of the regular season do not have a dramatic impact on evaluation metrics.

2.2 Tournament scoring systems

Standard systems for scoring NCAA basketball tournament pools, including those used in contests hosted online by ESPN (11 million participants in 2014) and Yahoo (15 million), award points based on picking each tournament game winner correctly, where picks are made prior to start of the tournament (ESPN 2014; Yahoo 2014). In these pools, there are no lost points for incorrect picks, but it is impossible to pick a game correctly if you had previously eliminated both participating teams in earlier rounds. The standard point allocation ranges from 1 point per game to 32 points for picking the tournament winner, or some function thereof, with successive rounds doubling in value. With the final game worth 32 times each first round game, picking the eventual tournament champion is more or less a prerequisite for a top finish. For example, among roughly one million entries in one 2014 ESPN pool, the top 106 finishers each correctly pegged the University of Connecticut as the champion (Pagels 2014). In terms of measuring the best prognosticator of all tournament games, the classic scoring system is inadequate, leading some to call for an updated structure among the websites hosting these contests (Pagels 2014); for more on optimal strategies in standard pools, see Metrick (1996) and Breiter and Carlin (1997).

Systems that classify games as "win" or "lose" fail to provide probability predictions, and without probabilities there is no information provided regarding the strength of victory predictions. For example, a team predicted to win with probability 0.99 by one system and 0.51 by another would both yield a "win" prediction, even though these are substantially different evaluations.

An alternative structure, used in the Kaggle competition, involved submitting a probability of victory for every potential matchup among the 68 teams prior to the play-in round. This corresponded to 2278 probability predictions. A submission would then be evaluated by a measure of predictive discrepancy on the 63 games that were played in the bracket of 64 teams. For simplicity, we refer generically to participants in a matchup as Team 1 and Team 2.

Let $\hat{y}_{ij}$ be the predicted probability of Team 1 beating Team 2 in game i on submission j, where i=1,…, 2278 and j=1,…, S, where S is the total number of submissions. Let y_i equal 1 if Team 1 beats Team 2 in game i and 0 otherwise. Letting $\mathrm{I}(Z_i=1)$ be an indicator for whether game i was actually played, define $\mathrm{LogLoss}_{ij}$ for entry j from game i as

$$\mathrm{LogLoss}_{ij}=-(y_i\log(\hat{y}_{ij})+(1-y_i)\log(1-\hat{y}_{ij})*\mathrm{I}(Z_i=1)).$$

Only the 63 games which were eventually played counted towards the participants' standing; i.e.,

$$\mathrm{LogLoss}_j=\frac{1}{\sum_{i=1}^{2278}\mathrm{I}(Z_i=1)}\sum_{i=1}^{2278}[\mathrm{LogLoss}_{ij}]=\frac{1}{63}\sum_{i=1}^{2278}[\mathrm{LogLoss}_{ij}].$$

Smaller log-loss scores are better, and the minimum log-loss score (i.e., picking all games correctly with probability 1) is 0.

There are several unique aspects of this scoring system. Most importantly, the probabilities that minimize the log-loss function for a submission are the same as the probabilities that maximize the likelihood of a logistic regression function. As a result, we begin our prediction modeling by focusing on logistic regression.

Further, all games are weighted equally, meaning that the tournament's first game counts as much towards the final standings as the championship game. Finally, as each entry picks a probability associated with every possible contest, prior incorrect picks do not prevent a submission from scoring well in future games.

2.2.1 Predicting NCAA games under probability based scoring function

Despite the intuitiveness of a probability based scoring system, little research has explored NCAA men's basketball predictions based on the log-loss or related functions. In one example that we build on, Carlin (1996) supplemented team-based computer ratings with pre-game spread information to improve model performance on the log-loss function in predicting the 1994 college basketball tournament. This algorithm showed a better log-loss score when compared to, among other methods, seed-based regression models and models using computer ratings only. However, Carlin's model was limited to computer ratings based only on the final scores of regular season games, and not on possession-based metrics, which are preferred for basketball analysis (Kubatko et al. 2007). Further, the application was restricted to the tournament's first four rounds in one season, and may not extrapolate to the final two rounds or to other seasons.

Kvam and Sokol (2006) applied similar principles in using logistic regression as the first step in developing a team ranking system prior to each tournament and found that their rankings outperformed seed-based evaluation systems. However, this proposal was focused on picking game winners rather than improving scoring under the log-loss function.

3 Model selection

Our submission was based on two unique sets of probabilities, $\hat{\boldsymbol{y}}_{\boldsymbol{m}_1}=[\hat{y}_{1,m_1},\ldots\hat{y}_{2278,m_1}]$ and $\hat{\boldsymbol{y}}_{\boldsymbol{m}_2}=[\hat{y}_{1,m_2},\ldots\hat{y}_{2278,m_2}]$, generated using a point spread-based model (M_1) and an efficiency-based model (M_2), respectively.

For M_1, we used a logistic regression model with all Division 1 NCAA men's basketball games from the prior 12 seasons for which we had point spread information. For game g, g=1,…, 65043, let y_g be our outcome variable, a binary indicator for whether or not the first team (Team 1) was victorious. Our only covariate in M_1 is the game's point spread, spread$_g$, as shown in Equation (1),

$$\text{logit}(Pr(y_g=1))=\beta_0+\beta_1{}^*\text{spread}_g. \quad (1)$$

We used the maximum likelihood estimates of β_0 and β_1 from Equation (1), $\hat{\beta}_{0,m_1}$ and $\hat{\beta}_{1,m_1}$, to calculate $\hat{y}_{i,m_1}$ for any i using spread$_i$, the point spread for 2014 tournament game i such that

$$\hat{y}_{i,m_1}=\hat{\pi}_i=\frac{\exp^{\hat{\beta}_{0,m_1}+\hat{\beta}_{1,m_1}{}^*\text{spread}_i}}{1+\exp^{\hat{\beta}_{0,m_1}+\hat{\beta}_{1,m_1}{}^*\text{spread}_i}}.$$

The actual point spread was available for the tournament's first 32 games; for the remaining 2246 contests, we predicted the game's point spread using a linear regression model with 2013–2014 game results.[1] Of course, just 31 of these 2246 predicted point spreads would eventually be needed, given that there are only 31 contests played in each tournament after the first round.

An efficiency model (M_2) was built using logistic regression on game outcomes, using seven team-based metrics for each of the game's home and away teams as covariates. These covariates are shown in Table 1. Each team's rating represents its expected winning percentage against a league average team (Pomeroy 2012). Offensive efficiency is defined as points scored per 100 possessions, defensive efficiency as points allowed per 100 possessions, and tempo as possessions per minute. Adjusted versions of offensive efficiency, defensive efficiency, and tempo are also shown; these standardize efficiency metrics to account for opposition quality, site of each game, and when each game was played (Pomeroy 2012). We also include an indicator variable for whether or not the game was played at a neutral site.

We considered several different logistic regression models using different combinations and functions of the 15 variables in Table 1. Our training data set, on which models were fit and initial parameters were estimated, consisted of every regular season game held before March 1 using each of the 2002–2003 through 2012–2013 seasons. For our test data, on which we averaged the log-loss function and selected our variables for M_2, we used all contests, both regular season and postseason, played after March 1 in each of these respective regular seasons. We avoided

Table 1: Team-based efficiency metrics.

Variable	Description	Team
X_1	Rating	Home
X_2	Rating	Away
X_3	Offensive efficiency	Home
X_4	Offensive efficiency	Away
X_5	Defensive efficiency	Home
X_6	Defensive Efficiency	Away
X_7	Offensive efficiency, adjusted	Home
X_8	Offensive efficiency, adjusted	Away
X_9	Defensive efficiency, adjusted	Home
X_{10}	Defensive efficiency, adjusted	Away
X_{11}	Tempo	Home
X_{12}	Tempo	Away
X_{13}	Tempo, adjusted	Home
X_{14}	Tempo, adjusted	Away
X_{15}	Neutral	N/A

1 This specific aspects of the model have been previously used for proprietary reasons, and we are unfortunately not at liberty to share it.

only using the Division 1 tournament outcomes as test data because only about 1% of a season's contests are played during these postseason games. Given that March includes conference tournament games, which are perhaps similar to those in the Division 1 tournament, and our desire to increase the pool of test data, we chose the earlier cutoff. Table 2 shows examples of the models we considered and their LogLoss score averaged on the test data.

While not the complete set of the fits that we considered, Table 2 gives an accurate portrayal of how we determined which variables to include. First, given the improvement in the loss score from fits (a) to (b) and (c) to (d), inclusion of X_{15}, an indicator for if the game was played on a neutral court, seemed automatic. Next, fit (f), which included the overall team metrics that had been adjusted for opponent quality, provided a marked improvement over the unadjusted team metrics in fit (e). Meanwhile, inclusion of overall team rating (fit (h)), and linear functions of team efficiency metrics (fit (g)), failed to improve upon the log-loss score from fit (f). Higher order terms, as featured in models (i), (j), and (k), resulted in worse log-loss performances on the test data, and an ad-hoc approach using trial and error determined that there were no interaction terms worth including.

The final model for M_2 contained the parameter estimates from a logistic regression fit of game outcomes on X_7, X_8, X_9, X_{10}, and X_{15} (Adjusted offensive efficiency for home and away teams, adjusted defensive efficiency for home and away teams, and a neutral site indicator, respectively). We estimated $\hat{y}_{m_2}$ using the corresponding team specific metrics from kenpom.com, taken immediately prior to the start of the 2014 tournament.

Table 2: Model building results.

Fit	Variables	*LogLoss*^
(a)	(X_1-X_2)	0.509
(b)	$(X_1-X_2), X_{15}$	0.496
(c)	X_1, X_2	0.510
(d)	X_1, X_2, X_{15}	0.496
(e)	$X_3, X_4, X_5, X_6, X_{15}$	0.538
(f)	$X_7, X_8, X_9, X_{10}, X_{15}$	**0.487**
(g)	$(X_7-X_8), (X_9-X_{10}), X_{15}$	0.487
(h)	$X_1, X_2, X_7, X_8, X_9, X_{10}, X_{15}$	0.487
(i)**	$(X_7, X_8, X_9, X_{10}, X_{15})^2$	0.488
(j)**	$(X_1, X_2, X_7, X_8, X_9, X_{10}, X_{13}, X_{14}, X_{15})^2$	0.488
(k)***	$(X_1, X_2, X_7, X_8, X_9, X_{10}, X_{13}, X_{14}, X_{15})^3$	0.493

^Games after March 1, in each of the 2002–2003 to 2012–2013 seasons.
**All two-way interactions of these variables.
***All three-way interactions of these variables.
Chosen model is bold.

Our final step used ensembling, in which individually produced classifiers are merged via a weighted average (Opitz and Maclin 1999). Previous work has shown that ensemble methods work best using accurate classifiers which make errors in different input regions, because areas where one classifier struggles would be offset by other classifiers (Hansen and Salamon 1990). While our two college basketball classifiers, M_1 and M_2, likely favor some of the same teams, each one is produced using unique information, and it seems plausible that each model would offset areas in which the other one struggles.

A preferred ensemble method takes the additional step of calculating the optimal weights (Dietterich 2000). For weight w, $w\in\{0, 0.01, \ldots, 1\}$, we used a weighted average of the predicted probabilities from M_1 and M_2, $w*\hat{y}_{m_1}+(1-w)*\hat{y}_{m_2}$, and calculated a LogLoss score averaged over each of the Division 1 tournaments between 2008 and 2013. The balance yielding the best LogLoss score used w=0.31, giving a weight of 0.31 to M_1 and 0.69 to M_2, and implying that efficiency metrics were slightly more predictive of tournament outcomes than the model based on spreads.

For the Kaggle contest, we did not submit an entry that was a mixture between the spread-based model and efficiency-based model specifically with weights 0.31 and 0.69. Given that we were allowed two submissions, we derived a range of weights that provided reasonable bounds for minimizing the LogLoss score using the tournament outcomes from previous years. The LogLoss score using $w\in\{0.10, 0.65\}$ was within ±0.01 of the minimum score that we found using w=0.31. As stated previously, however, we were concerned that because the efficiency metrics we used were calculated based on tournament results in addition to games over the regular season, their use for predicting the 2014 tournament may not have been as reliable. We therefore estimated that a more robust upper bound for w would be 0.75. Thus, one of our submissions to the Kaggle contest used this upper bound of w=0.75, and we used w=0.25 for the other submission. The latter submission resulted in nearly an identical LogLoss score as the mixture with w=0.31, and the former checked the sensitivity to placing greater weight on the spread-based model. This yielded two submissions, $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, where

$$\boldsymbol{S}_1=0.75*\hat{\boldsymbol{y}}_{m_1}+0.25*\hat{\boldsymbol{y}}_{m_2}$$
$$\boldsymbol{S}_2=0.25*\hat{\boldsymbol{y}}_{m_1}+0.75*\hat{\boldsymbol{y}}_{m_2}$$

The correlation between $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ was 0.94, and 78% of game predictions on the two entries were within 0.10 of one another.

4 Simulation study

In order to evaluate the luck involved in winning a tournament pool with probability entries, we performed a simulation study, assigning each entry a LogLoss score at many different realizations of the 2014 NCAA basketball tournament. The contest organizer provided each of the 433 submissions to the 2014 Kaggle contest for this evaluation. These entries were submitted by 248 unique teams; each team was allowed up to 2 entries, with only the team's best score used in the overall standings.

To simulate the tournament, "true" win probabilities must be specified for each game. We evaluate tournament outcomes over five sets of true underlying game probabilities: $\boldsymbol{S}_1$, $\boldsymbol{S}_2$, M($\boldsymbol{S}_{All}$), M($\mathbf{S}_{Top10}$), and **0.5**, listed as follows.

- Our first entry ($\boldsymbol{S}_1$)
- Our second entry ($\boldsymbol{S}_2$)
- Median of all Kaggle entries (M($\boldsymbol{S}_{All}$))
- Median of the top 10 Kaggle entries (M($\boldsymbol{S}_{Top10}$))
- All games were a coin flip (i.e. p=0.5 for all games) (**0.5**)

Let rank($\boldsymbol{S}_1$) and rank($\boldsymbol{S}_2$) be vectors containing the ranks of each of our submissions across the 10,000 simulations at a given set of game probabilities. We are interested in the median rank and percentiles (2.5, 97.5) for each submission (abbreviated as M (2.5–97.5)), across all simulations. We are also interested in how often each submission finishes first and in the top 10.

Lastly, we extract the number of unique winners across the simulations, which can give us a sense of how many entries had a reasonable chance of winning at each set of underlying game probabilities.

The results of the simulations appear in Table 3. Each column represents a different "true" probability scenario and each row records the results of a statistic of interest. The first and second rows show results of our first and second entry, respectively. We can see that if the "true" probabilities were $\boldsymbol{S}_1$, our entry finished at a median of 11th place, whereas if the true probabilities were $\boldsymbol{S}_2$, our median finish was 14th place. If the true probabilities were $\boldsymbol{S}_1$, our entry containing those probabilities would finish in first place around 15% of the time. Likewise, with $\boldsymbol{S}_2$ as "true" probabilities, that entry would win around 12% of the time. Relative to a contest based entirely on luck, where each entry would have a 1 in 433 chance of finishing first, our chances of winning were roughly 50 to 60 times higher using $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ as the truth. This conceivably represents the upper bound of our submission's "skill."

On the whole, our simulations indicated that the amount of luck required to win the Kaggle contest was enormous; even when our predictions were the true underlying game probabilities, those submissions only won roughly 1 in 8 times, respectively! Further, if our submissions were correct, we only finished in the top 10 about 49% and 45% of the time, respectively, for $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$.

If the median of all entries (M($\boldsymbol{S}_{All}$)) or the median of the top 10 entries (M($\boldsymbol{S}_{Top10}$)) is used as the true probabilities, our chances of winning diminish. For M($\boldsymbol{S}_{Top10}$), our chances of winning on entries $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ were both about 2%. For M($\boldsymbol{S}_{All}$), our chances of winning on entries $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ were both <1%. Lastly, if each game was truly a coin flip, neither of our entries finished first in any of the simulations.

Of the 433 total entries, fewer than 350 finished in first place at least once in each of the simulations with our entries as the truth. This suggests that if either of our submitted probabilities were close to the "true" probabilities, about 20% of entries had little to no chance of winning. Further, there were roughly 80 submissions that never won when using the median of all contest entries as the true probabilities. As a result, it would be reasonable to argue that the true pool of submissions that we were competing against was closer to 350.

Lastly, Figure 1 shows the smoothed density estimates (dark line) of all winning LogLoss scores from 10,000

Table 3: Simulation results.

Outcome	Type	True Game Probabilities				
		$\boldsymbol{S}_1$	$\boldsymbol{S}_2$	M($\boldsymbol{S}_{Top10}$)	M($\boldsymbol{S}_{All}$)	0.5
Rank ($\boldsymbol{S}_1$)	M (2.5–97.5)	11 (1–168)	59 (1–202)	99 (2–236)	145 (4–261)	264 (186–299)
Rank ($\boldsymbol{S}_2$)	M (2.5–97.5)	53 (2–205)	14 (1–164)	92 (2–245)	146 (5–266)	226 (135–285)
Rank ($\boldsymbol{S}_1$)=1	%	15.57	3.90	2.02	0.88	0
Rank ($\boldsymbol{S}_2$)=1	%	2.22	11.65	1.89	0.63	0
Rank ($\boldsymbol{S}_1$)≤10	%	48.79	17.69	8.85	5.04	0
Rank ($\boldsymbol{S}_2$)≤10	%	20.72	44.47	11.96	4.77	<0.01
Unique winners	Total	332	336	337	348	217

M, Median, 2.5: 2.5th percentile, 97.5: 97.5th percentile.

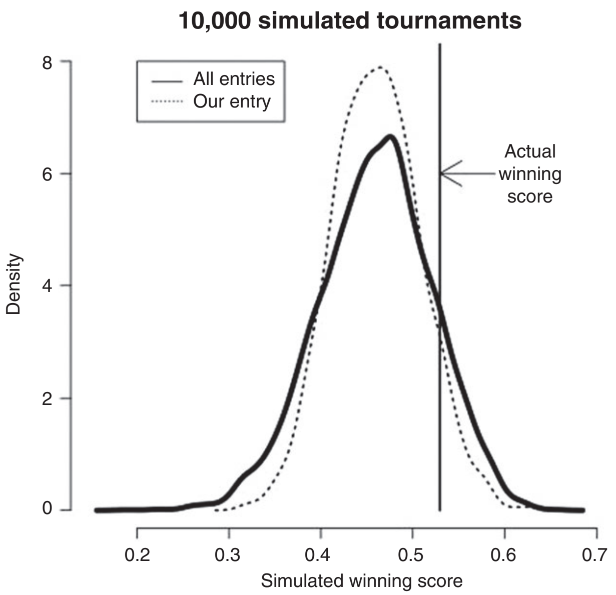

Figure 1: Smoothed density estimates of winning scores across 10,000 tournament simulations using underlying game probabilities $\boldsymbol{S}_2$. The dotted line refers to the winning scores on simulations won by the $\boldsymbol{S}_2$ entry.

simulated tournaments under the game probabilities in $\boldsymbol{S}_2$, along with the density estimates for $\boldsymbol{S}_2$ on simulations in which that entry was the winner. The winning score of $\boldsymbol{S}_2$ in 2014 (0.529) is shown by a vertical line. Relative to the simulated winning scores, $\boldsymbol{S}_2$ winning scores have a lower density in the tails. The 2014 winning score was relatively higher than most of the scores that won the simulated tournaments, perhaps because the University of Connecticut, a seven seed, won all six of its games en route to becoming Division 1 champions. Prior to Connecticut's win, only four previous champions since 1979 were seeded lower than three [four seed Arizona (1997), six seeds Kansas (1988) and North Carolina (1983), and eight seed Villanova (1985)]. As a result, most predictive models would have a seven seed as a substantial underdog in games against higher seeded opponents in the later rounds, leading to comparatively larger values of the loss function than when favorites prevail. Another explanation for the comparatively higher winning score is that our simulated winning scores were computed after assuming that our entry probabilities were the true probabilities, when in reality, this was unlikely to be the case.

5 Conclusion

In this manuscript, we describe a model for predicting NCAA tournament outcomes using logistic regression models with efficiency metrics and the Las Vegas point spread as predictors. The probabilities generated from this model outperformed more than 400 other sets of predictions in terms of the log-loss score for the 2014 NCAA basketball tournament, in spite of the modest statistical methods employed. We further looked at how much luck was involved in producing predictions that minimized log-loss relative to other contest submissions.

In place of more complex modeling strategies, we placed our focus and efforts into finding the best possible data. This required the domain-specific knowledge of both where to look for and find the point-spread and possession data. It is extremely difficult to generate predictive models that outperform the Las Vegas point spread, particularly in high profile games like the ones in the NCAA tournament, and both the point spread and team ratings have previously been shown to work well in predicting college basketball outcomes (Carlin 1996). Conceptually, one could argue that the Las Vegas point spread is a subjective prior based on expert knowledge, whereas Pomeroy's ratings are based entirely on data. In this way, our ensembling of these two sources of data follows the same principles as a Bayesian analysis.

While we are unable to ascertain the models or algorithms used by other entrants, several participants were willing to share their ideas post-hoc (Kaggle 2014). These included, among other strategies, random forests and generalized boosted models. Incidentally, both of these algorithms have been shown to provide more accurate predictions than the logistic regression models that we employed (Caruana and Niculescu-Mizil 2006). Several participants also indicated that they used additional data, including each team's travel distance, seed information, conference, and past tournament performances (Kaggle 2014). In principal, however, much of this information is built into the point-spread and/or team efficiency metrics. As a result, future work appears warranted to combine the point-spread and efficiency data with more advanced predictive algorithms.

There are many different types of NCAA tournament pools, and to win any pool takes some combination of skill and luck. We attempted to identify how much luck was involved in winning in one specific contest format. To assess luck, we simulated outcomes from the 2014 NCAA tournament under several assumptions to compare the predictions generated from our model to the other sets of predictions. It is reasonable to estimate that our models increased our chances of winning anywhere from between five-fold to fifty-fold, relative to a contest that just randomly picked a winner. However, even under the most optimistic scenario, we found that our prediction set had less than a 50% chance of being one of the ten best and less than a 20% chance of producing the minimum log-loss of all sets of prediction. Under different, but fairly realistic true probability scenarios, the chances of our predictions producing the minimum log-loss decreased to be <2%.

The findings of our simulations in part reflect the quality of submissions to the 2014 contest, and as a result, it may be difficult to generalize to other pools with different sets of prediction entries. Given these results, however, it seems fair to conclude that the use of informative data with traditional statistical tools can still result in predictions whose accuracy rivals that of more complex models. Moreover, no matter how good of a predictive model that one builds, an immense amount of luck is also needed to win an NCAA tournament pool.

References

Barra, A. 2014. *Is March Madness a Sporting Event – or a Gambling Event?*. URL http://www.theatlantic.com/entertainment/archive/2014/03/is-march-madness-a-sporting-event-or-a-gambling-event/284545/ (accessed June 1, 2014).

Boudway, I. 2014. *The Legal Madness Around NCAA Bracket Pools*. URL http://www.businessweek.com/articles/2012-03-15/the-legal-madness-around-ncaa-bracket-pools (accessed June 1, 2014).

Boulier, B. L. and H. O. Stekler. 1999. "Are Sports Seedings Good Predictors?: An Evaluation." *International Journal of Forecasting* 15:83–91.

Breiter, D. J. and B. P. Carlin. 1997. "How to Play Office Pools if You Must." *Chance* 10:5–11.

Carlin, B. P. 1996. "Improved NCAA Basketball Tournament Modeling Via Point Spread and Team Strength Information." *The American Statistician* 50:39–43.

Caruana, R. and A. Niculescu-Mizil. 2006. "An Empirical Comparison of Supervised Learning Algorithms." In *Proceedings of the 23rd International Conference on Machine Learning*, ACM. pp. 161–168.

Colquitt, L. L., N. H. Godwin, and S. B. Caudill. 2001. "Testing Efficiency Across Markets: Evidence from the NCAA Basketball Betting Market." *Journal of Business Finance & Accounting* 28:231–248.

Constantinou, A. C., N. E. Fenton, and M. Neil. 2013. "Profiting from an Inefficient Association Football Gambling Market: Prediction, Risk and Uncertainty using Bayesian Networks." *Knowledge-Based Systems* 50:60–86.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Multiple classifier systems* (pp. 1–15). Berlin, Heidelberg: Springer.

ESPN. 2014. *Official Rules*. URL http://games.espn.go.com/tournament-challenge-bracket/2014 (accessed June 1, 2014).

Hansen, L. K. and P. Salamon. 1990. "Neural Network Ensembles." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12:993–1001.

Harville, D. 1980. "Predictions for National Football League Games Via Linear-Model Methodology." *Journal of the American Statistical Association* 75:516–524.

Kaggle. 2014. *Competition Forum*. URL https://www.kaggle.com/c/march-machine-learning-mania/forums (accessed June 1, 2014).

Kubatko, J., D. Oliver, K. Pelton, and D. T. Rosenbaum. 2007. "A Starting Point for Analyzing Basketball Statistics." *Journal of Quantitative Analysis in Sports* 3:1–22.

Kvam, P. and J. S. Sokol. 2006. "A Logistic Regression/Markov Chain Model for NCAA Basketball." *Naval Research Logistics (NrL)* 53:788–803.

Linna, K., E. Moore, R. Paul, and A. Weinbach. 2014. "The Effects of the Clock and Kickoff Rule Changes on Actual and Market-Based Expected Scoring in NCAA Football." *International Journal of Financial Studies* 2:179–192.

Metrick, A. 1996. "March Madness? Strategic Behavior in NCAA Basketball Tournament Betting Pools." *Journal of Economic Behavior & Organization* 30:159–172.

Nichols, M. W. 2014. "The Impact of Visiting Team Travel on Game Outcome and Biases in NFL Betting Markets." *Journal of Sports Economics* 15:78–96.

Opitz, D. and R. Maclin. 1999. "Popular Ensemble Methods: An Empirical Study." *Journal of Artificial Intelligence Research* 11:169–198.

Pagels, J. 2014. *Challenging the Tournament Challenge: Devising a More Equitable Bracket Scoring System*. URL https://www.bsports.com/statsinsights/ncaa/march-madness-scoring.

Paul, R. J. and A. P. Weinbach. 2014. "Market Efficiency and Behavioral Biases in the WNBA Betting Market." *International Journal of Financial Studies* 2:193–202.

Paul, R. and A. Weinbach. 2005. "Market Efficiency and NCAA College Basketball Gambling." *Journal of Economics and Finance* 29:403–408.

Pomeroy, K. 2012. *Ratings Glossary*. URL http://kenpom.com/blog/index.php/weblog/entry/ratings_glossary (accessed June 1, 2014).

Schwertman, N. C., K. L. Schenk, and B. C. Holbrook. 1996. "More Probability Models for the NCAA Regional Basketball Tournaments." *The American Statistician* 50:34–38.

Stern, H. 1991. "On the Probability of Winning a Football Game." *The American Statistician* 45:179–183.

TeamRankings. 2014. *NCAA BB Team Possessions per Game*. URL http://www.teamrankings.com/ncb/ (accessed June 1, 2014).

Tsu, T. 2014. *March Madness: Distracted Workers, Illegal Gambling, Loss of Sleep?* URL http://articles.latimes.com/2012/mar/12/business/la-fi-mo-march-madness-20120312 (accessed June 1, 2014).

Yahoo 2014. *Official Rules*. URL https://www.quickenloansbracket.com/rules/rules.html (accessed June 1, 2014).