Journal of Sports Research
2016 Vol.3, No.1, pp.10-17
ISSN(e): 2410-6534
ISSN(p): 2413-8436
DOI: 10.18488/journal.90/2016.3.1/90.1.10.17

PREDICTING RESULTS OF MARCH MADNESS USING THREE DIFFERENT METHODS

Gang Shen[1] --- Di Gao[2] --- Qian Wen[3] --- Rhonda Magel[4†]
[1,2,3,4]*Department of Statistics North Dakota State University Fargo, ND*

ABSTRACT

Three methods are used to predict the results for two years of the Men's NCAA Division1 March Madness Basketball Tournament. These methods include using the machine-learning method of the support vector machine, the data mining method of the random forest, and a newly developed Bayesian model using the property of probability self-consistency as an extension of Shen et al. (2015). The random forest method and the support vector machine method are found to possibly do slightly better than the Bayes model, although the results vary. Possible ideas as to how to extend the Bayes model are given.

Keywords: Random forest, Support vector machine, Bayes model, Single, Double scoring system.

Received: 20 April 2016/ **Revised:** 11 May 2016/ **Accepted:** 17 May 2016/ **Published:** 21 May 2016

Contribution/ Originality

This study develops a Bayesian model using the property of self-consistency and compares this model with two other modeling techniques in predicting the results of March Madness.

1. INTRODUCTION

The National Collegiate Athletic Association (NCAA) Division I Men's Basketball tournament, also known as March Madness is played annually. There are currently 68 teams involved in the tournament each year with 8 of the teams playing in the first round, known as First 4 Round. The tournament is played in March and early April (March Madness, 2014). After the first round, 64 teams are left, and each of the teams are seeded with a single seed number ranging from 1 (strongest) to 16 (weakest). The tournament is divided up into the following rounds after the First 4 Round: Round 64; Round 32; Sweet 16; Elite 8; Final 4; and Championship Round. The tournament is single elimination, and thus, if a team loses a game, they are out of the tournament. The 16 seeds are assigned to each team in each of four regions (East, South, West, and Midwest) by the NCAA selection committee. In Round 64, teams with seeds in the same region adding up to 17 play each other. Teams that win, go to the next round. The four teams winning each of their regions, go to the Final 4 (East vs South, West vs Midwest), and winners from this go to the Championship game. Brackets are explained by Breiter and Carlin (1997).

Each year, before Round 64, several million people complete brackets as to which team they think will win each game and then the championship. It was estimated that in 2014, the money that was bet on March Madness exceeded the amount of money that was bet on the Super Bowl by $2 billion (Barra, 2014). There are several types of scoring systems that can be used on March Madness, but we will mention just two of the more popular scoring systems. These systems include the single scoring system and the double scoring system. In the single scoring system, an individual is awarded one point for each game that they predict correctly. Since there are 63 games starting with Round 64, the maximum number of points that an individual can attain is 63. In the double scoring system, points for each game predicted correctly double after each round and the number of possible points for each

† Corresponding author

round is 32. Each game predicted correctly in Round 64 is worth 1 point, and in Round 32 is worth 2 points, and in Sweet 16 is worth 4 points, and Elite 8 is worth 8 points, and Final 4 is work 16 points, and predicting the champion is worth 32 points. In the double scoring system, an individual can earn up to 192 points.

In this research, we will discuss three ways which could be used to predict the results of March Madness and thus, help an individual fill out a bracket before the games begin. The first of these methods uses an extension of the probability self-consistent method proposed by Shen *et al.* (2015). In this case, a Bayesian approach is used with this probability self-consistent method. The second method involves using the machine learning method with the support vector machine. The third method uses the data mining method of the random forest. Each of these methods will be used on both the 2015 and 2016 March Madness tournaments, assuming results are unknown at the beginning of Round 64 and results obtained from the single scoring system and double scoring system will be compared for each of the methods.

2. SOME PAST RESEARCH ON MODELING NCAA DIVISION 1 BASKETBALL GAMES

There has been much research conducted on modeling for NCAA men's basketball games because of March Madness. We will discuss some of this research. Magel and Unruh (2013) studied various factors that influenced the point spread of an NCAA men's basketball game. The factors they found to be significant in the final model were differences between the two teams of the following: free throw attempts; defensive rebounds; assists; and turnovers. In order to determine if this information could be used to predict the outcome of a future basketball game, a sample of 100 games from the 2011-2012 season was selected. One team in each of the games was randomly selected to be the "team of interest" and the other team as the "opposing team". Data was collected on the previous 4 games that each of the teams played on the four factors. Medians of the in-game statistics were found for each team for the four factors discussed and the differences of these medians were taken for each team and put into the developed model to predict the point spread in the order of "team of interest" minus "opposing team". The models had an accuracy of 64% to 68% when the medians were used.

Shi *et al.* (2013) used both a machine learning method and a random forest method, in addition to two other methods to predict results of individual NCAA Division 1 basketball games over the seasons 2009 to 2013. The machine learning method they used appeared to do slightly better than the random forest method. Their conclusions were that the models used did not make so much difference as the factors considered in the models. They also concluded that more training data did not necessarily make better models. Their methods were used on individual games only and not completing a bracket for March Madness.

A logistic linear model with probability self-consistency was developed by Shen *et al.* (2015) for bracketing March Madness. The property of probability self-consistency means that the probabilities of each team making it to the next round given the team won all the rounds up to that point, in a set of teams that could potentially compete in that round, must add up to 1. Shen *et al.* (2015) used the classical maximum likelihood estimation technique of estimating coefficients in their model. When their method was compared to the restricted OLRE method proposed by West (2006); West (2008); Pomeroy (2015, 2016) and RPI (2015) for the 2014 March Madness tournament results, their method did better than the other methods, particularly in regard to the double scoring.

3. VARIABLES CONSIDERED IN THE MODELS

Overall, before beginning any of the modeling procedures, data was collected on fourteen variables beginning with the 2007-2008 season and ending with the 2013-2014 season based on the teams in March Madness and the results from the tournaments. The data collected involved variables that the previous other studies mention have found to be important in predicting March Madness results or results from individual NCAA men's basketball games.. The variables in which data were collected on include the following regular seasonal averages for each team playing in March Madness (NCAA, 2015, 2016): FGM (average number of field goals made per game); 3PM

(average number of 3 point field goals made per game); FTA (average number of free throws attempted per game); ORPG (average number of offensive rebounds per game); DRPG (average number of defensive rebounds per game) ; APG (average number of assists per game); PFPG (average number of personal fouls per game); ASM (average scoring margin); Seed number; SAGSOS (Sagarin strength of schedule) (Sagarin, 2015, 2016) ATRATIO (average assists to turnover ratio); AdjO (adjusted offensive efficiency – estimates average points scored by team per 100 possessions playing against an average D1 defense) (Pomeroy, 2015, 2016) AdjD (adjusted defensive efficiency – estimates average points allowed based on 100 possessions against an average D1 offense) (Pomeroy, 2015, 2016) and Pyth (a team's expected winning percentage against an average D1 team) (Pomeroy, 2015, 2016). The margins of predictors (of two teams, i and j) playing in a game are normalized through the transformation:

$$(x_i - x_j)/(|x_i| + |x_j|).$$

Data was also collected between 2001-2002 and 2006-2007 seasons on the number of times each combination of seeds played each other and then which seed won in order to estimate relative probabilities. This was done for the Bayesian model.

4. BAYESIAN LOGISTIC LINEAR MODEL WITH PROBABILITY SELF-CONSISTENCY

A model was developed which extends the work of Shen *et al.* (2015) by using a Bayesian approach of estimating coefficients in the probability self-consistency model. In this case, the historical winning rate between different seeds was used based on all March Madness games played between the 2001-2002 and 2006-2007 season. For example, there were a total of 12 games played between a number 1 and a number 8 seed during this time. In 10 of these games, the number 1 seed won. Thus, the estimated historical probability of the number 1 seed winning when it played a number 8 seed is 0.833. As another example, a number 1 seed played a number 4 seed 7 times during this period, winning 5 of these times, for an estimated probability of .714 of a number 1 seed beating a number 4 seed. During this time, a number 1 seed beat a number 16 in 20 games out of 20 for an estimated probability of 1 of a 1 seed beating a 16 seed. The historical winning rates between 2001-01 and 2006-07 are given in Table 1.

To illicit the prior for the model coefficients from the historical winning rate between different seeds, we first adjust the historical winning rate between seeds as follows, $p_{ij} = 0.9r_{ij} + 0.05$, where r_{ij} is the observed winning rate between seed i and j. This transformation guarantees the winning probability between any two seeds to be in the interval $[0.05, 0.95]$. We consider a multivariate normal prior for p_{ij}, i.e.,

$$p_{ij} \sim N(\hat{p}_{ij}, \hat{p}_{ij}(1-\hat{p}_{ij})/n_{ij})$$

where $\hat{p}_{ij}$is the observed historical winning rate of seed i over seed j and n_{ij} is the number of games played between the two seeds in March Madness seasons 2001-2006. Then, by delta method, $\log\frac{p_{ij}}{1-p_{ij}} \sim N(\log\frac{\hat{p}_{ij}}{1-\hat{p}_{ij}}, \frac{1}{n_{ij}\hat{p}_{ij}(1-\hat{p}_{ij})})$ approximately. Let p, n be the collection of all the p_{ij} and n_{ij} observed in the seasons, respectively, and so is $\hat{p}$.

Note the logistic linear model implies $\log\frac{p}{1-p} = X\beta$ for all the games, where X is the matrix with rows x_i'. In this model, the X matrix contains the data for the 13 variables listed in Section 2, not including the seed since this was used in the prior. The data for the X matrix in this model included the data from the 2007-2008 through 2012-2013 March Madness results and team statistics. No data on the March Madness results was used except in the calculation of the probabilities of seed i beating seed j. Hence, by abusing the notation a little bit, we may derive a multivariate normal prior of β: $\beta \sim N(\beta_0, \Sigma_0)$, where $\beta_0 = (XX')^{-1}X\log\frac{\hat{p}}{1-\hat{p}}$, and

$$\Sigma_0 = (XX')^{-1}X[n\hat{p}(1-\hat{p})]^{-1}X'(XX')^{-1}.$$

Note the logistic linear model implies $\log\frac{p}{1-p} = X\beta$ for all the games, where X is the matrix with rows x_i'. In this model, the X matrix contains the data for the 13 variables listed in Section 2, not including the seed since this was used in the prior. The data for the X matrix in this model included the data from the 2007-2008 through 2012-2013 March Madness results and team statistics. No data on the March Madness results was used except in the calculation of the probabilities of seed i beating seed j. Hence, by abusing the notation a little bit, we may derive a multivariate normal prior of β: $\beta \sim N(\beta_0, \Sigma_0)$, where $\beta_0 = (XX')^{-1}X\log\frac{\hat{p}}{1-\hat{p}}$, and $\Sigma_0 = (XX')^{-1}X[n\hat{p}(1-\hat{p})]^{-1}X'(XX')^{-1}$.

Table-1. Historical Winning Rates 2001-02 through 2006-07

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.333 6	0.200 5	0.714 7	0.800 10	-	-	0.833 12	0.875 8	-	-	-	-	-	-	1.000 20
2	-	0.545 11	-	-	-	0.615 13	-	-	0.429 7	-	-	-	-	1.000 20	-
3	-	-	-	-	0.583 12	-	-	-	-	0.714 7	-	-	0.900 20	-	-
4	-	-	-	0.143 7	-	-	-	-	-	-	0.667 9	0.800 20	-	-	-
5	-	-	-	-	-	-	-	-	-	-	0.550 20	-	-	-	-
6	-	-	-	-	-	-	-	-	-	0.700 20	-	-	-	-	-
7	-	-	-	-	-	-	-	-	0.650 20	-	-	-	-	-	-
8	-	-	-	-	-	-	-	0.600 20	-	-	-	-	-	-	-

The posterior of β is then $f(\beta|y,X) \propto f(\beta)\prod_{i=1}^{n}\exp(y_i x_i'\beta)[1+\exp(x_i'\beta)]^{-1}$. Simple importance resampling algorithm is used in drawing samples from the posterior, with proposal density $f(\beta|y,X)$ and the resampling weight $\prod_{i=1}^{n}\exp(y_i x_i'\beta)[1+\exp(x_i'\beta)]^{-1}$.

The Bayesian estimates of the coefficients associated with the 13 variables are given below. These are used in each of the rounds of March Madness to predict the winning team.

FGM	AdjO	AdjD	ASM	SAGSOS	Pyth	X3PM	FTA
-0.075	14.103	-14.242	-0.059	4.415	2.1527	0.331	0.621

ORPG	DRPG	APG	PFPG	ATRATIO
0.015	0.934	0.342	-0.114	-0.011

The accuracy of bracketing March Madness 2014/5 season by Bayes logistic linear model with self-consistency for each round is given in Table 2.

Table-2. Accuracy of Each Round of Bayes Model 2015 March Madness

Round of 64	Round of 32	Sweet 16	Elite 8	Final 4	Championship
78.1%	56.3%	50%	25%	25%	0

Source: NCAA 2015

The double and single scoring accuracy rate for March Madness 2015 are calculated below:

Double scoring accuracy: 83/192 = 43.2%

Single scoring accuracy: 40/63= 63.5%

The accuracy of bracketing March Madness 2015/16 season by Bayes logistic linear model with self-consistency for each round in this tournament is given in Table 3.

Table-3. Accuracy of Each Round of Bayes Model 2016 March Madness

Round of 64	Round of 32	Sweet 16	Elite 8	Final 4	Championship
75%	62.5%	75%	50%	50%	0%

Source: NCAA 2016

The double and single scoring accuracy rate for March Madness 2016 are calculated below:

Double scoring accuracy: 100/192=52.08%

Single scoring accuracy: 43/63=68.25%

5. MACHINE LEARNING METHOD: SUPPORT VECTOR MACHINE

The classification problem of classifying a team as to whether they would win or lose could also be approached by the support vector machine (SVM), a supervised method for finding a decision boundary in the space of the margin of game predictors to classify the outcome of each game. During the training process, the SVM aims to find a hyperplane (i.e. a geometric margin) that maximizes the width of the gap between the two categories, win and lose. The resulting classifier can then be used to determine whether a new game is a win or lose based on the margin of predictors. This resulting hyperplane can be defined by a subset of examples, called support vectors, which "mark" the boundary between classes. Mathematically, an SVM classifier

is in a form of $\sum_{i=1}^{m} \alpha_i y_i K(x, x_i) + b$ with $K(x, x_i) = x_i' x$, and α_i $(i = 1,2, \dots, m)$ are the parameters minimizing the criterion $\frac{1}{2}\sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - c\sum_{i=1}^{m} \alpha_i y_i$ subjected to $\sum_{i=1}^{m} \alpha_i = 0$ and $0 \leq \alpha_i y_i \leq c$ for some predetermined tuning parameter c.

Data for this method was collected based on seasonal averages of all teams playing in March Madness and the results from March Madness for the years 2007-2008 through 2013-2014 on the 14 variables that are given in Section 2. This was for the training process.

To account for possible nonlinearity of the classifier, we consider the kernel with Gaussian radial basis in the space of the margins of game predictors $K(u, v) = \exp(-\gamma||u - v||^2)$, where γ is a tuning parameter. The training process indicates that the optimal choice of γ is 0.007 in terms of cross-validation error, and that of cost c is 1. Our classifier turns out to be a linear combination of 397 different Gaussian radial functions $K(x, x_j)$ with an intercept -0.198.

After the training process, the model was used to predict the results of March Madness in 2015 before Round 64 began. The accuracy of bracketing March Madness 2014/5 season by SVM is given in Table 4.

Table-4. Accuracy of Each Round of SVM Model 2015 March Madness

Round of 64	Round of 32	Sweet 16	Elite 8	Final 4	Championship
87.5%	75%	75%	75%	50%	0%

Source: NCAA (2015, 2016)

The double and single scoring accuracy rate for March Madness 2015 are calculated below:

Double scoring accuracy: 116/192 = 60.4%

Single scoring accuracy: 50/63= 79.4%

The model was next used to predict the results of March Madness in 2016 and the results for each round are given in Table 5.

Table-5. Accuracy of Each Round of SVM Model 2016 March Madness

Round of 64	Round of 32	Sweet 16	Elite 8	Final 4	Championship
65.6%	68.8%	62.5%	50%	50%	0%

Source: NCAA (2015, 2016)

The double and single scoring accuracy rate for March Madness 2016 are calculated below:

Double scoring accuracy: 95/192=49.5%

Single scoring accuracy: 40/63= 63.5%

The University of North Carolina lost in the last second of the Championship game to Villanova who scored a 3 point shot. If North Carolina had won the game, the double scoring accuracy would have been 66.1% and the single scoring accuracy would have been 65.1%.

6. RANDOM FOREST METHOD

Prediction of a game outcome may be deemed as a classification problem. Based on margin of the game predictors of two competing teams, we need to classify the outcome of the game into two possible categories: win or lose. This could be done by classification tree method in data mining.

Classification trees work similarly to regression trees except the residual sum of squares is no longer a suitable criterion for splitting the nodes. Instead one may use Gini index $1-\sum p_{ik}^2$ to measure the impurity of the nodes, where p_{ik} is the observed proportion of wins or loses within node i. Figure 1 gives an illustration of a typical classification tree. where 1 means win, and 0 means lose.

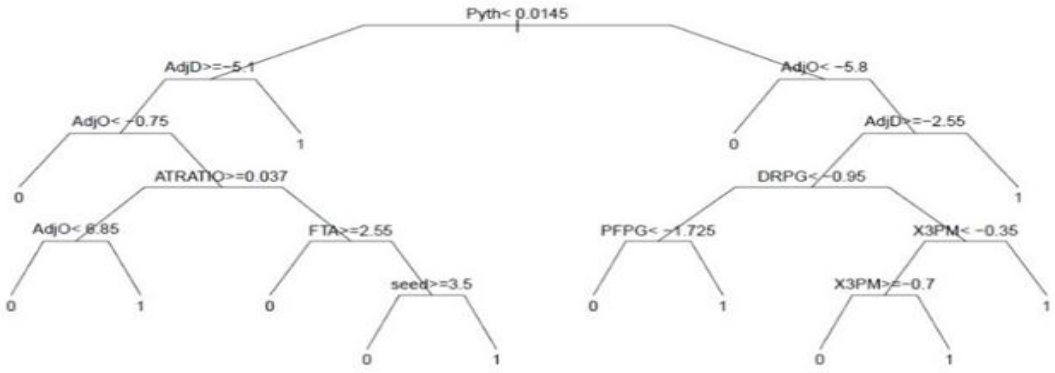

Figure-1. Illustration of Classification Tree

Source: (NCAA, 2015, 2016; Pomeroy, 2015, 2016; Sagarin, 2015, 2016)

Our random forest grows 500 classification trees. To classify a new game outcome from margin of two team predictors, it puts the input them down each of the trees in the forest. Each tree gives a classification, which may be counted as a "votes" for that category (win or lose). The forest then chooses the classification having the most votes (over all the trees in the forest).

There is a trade-off between small and large number of predictors used in splitting each node in trees. The out-of-bag error is considered as the criterion in our selection of the optimal number of predictors in our study. Each tree in our random forest is grown as follows: Bootstrap with replacement all the training data, randomly select 5 predictors out of the 15 predictors at each node for splitting the node. No pruning is taken during the growth of each tree.

Figure 2 graphs the importance of all the predictors measured by mean decrease in Gini index. The larger value indicates larger importance

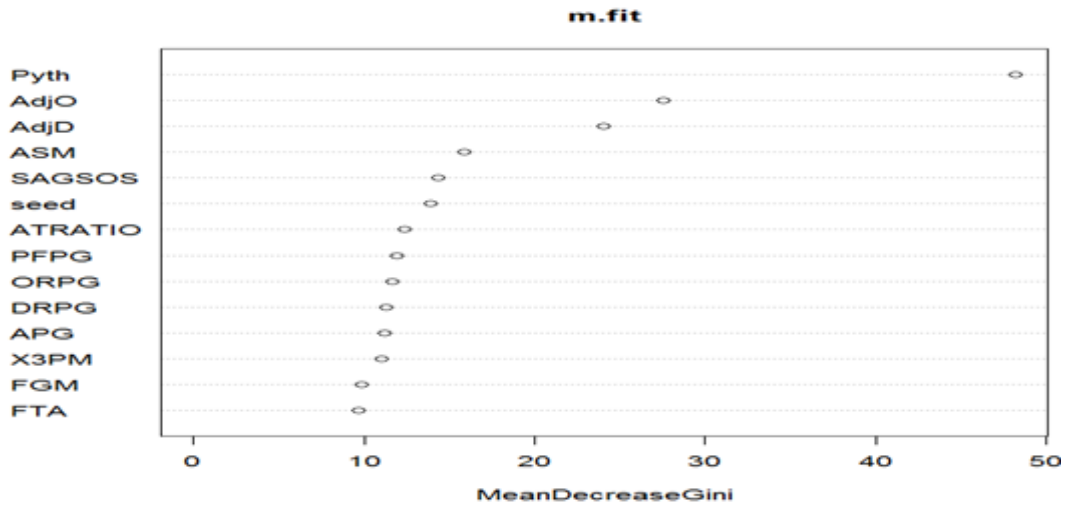

Figure-2. Gini Index of Each Predictor

Source: (NCAA, 2015, 2016; Pomeroy, 2015, 2016; Sagarin, 2015, 2016)

The bracketing accuracy by Random Forest Method for March Madness 2014/5 season is given in Table 6.

Table-6. Accuracy of Each Round of Random Forest Model 2015 March Madness

Round of 64	Round of 32	Sweet 16	Elite 8	Final 4	Championship
78.1%	68.8%	62.5%	75%	25%	0

Source: NCAA (2015, 2016)

The double and single scoring accuracy rate for March Madness 2015 are calculated below:

Double scoring accuracy: 107/192 = 57.3%

Single scoring accuracy: 45/63= 71.4%

The bracketing accuracy by Random Forest Method for March Madness 2015/16 season is given in Table 7.

Table-7. Accuracy of Each Round of Random Forest Model 2016 March Madness

Round of 64	Round of 32	Sweet 16	Elite 8	Final 4	Championship
78.1%	56.3%	75%	50%	50%	0%

Source: NCAA (2015, 2016)

The double and single scoring accuracy rate for March Madness 2016 are calculated below:

Double scoring accuracy: 99/192=51.6%

Single scoring accuracy: 43/63=68.25%

It is noted that in 2016, University of North Carolina lost in the last second of the game to Villanova who scored a 3 point shot. If North Carolina would have won, the double scoring accuracy for this model would have been 68.2% and the single scoring accuracy, 69.8%.

7. CONCLUSIONS

For the 2015 Championship predictions, the machine-learning method did the best in the double scoring system, getting 60.4% of all the points, followed by the random forest method at 57.3% of the points, and the Bayes model at 43.2% of the points. For the 2016 Championship prediction, the Bayes model actually was a little ahead of the other two methods, getting 52.08% of the double points, with the random forest method obtaining 51.6% of the points and the machine-learning method obtaining 49.5% of the points. It is noted however, that if the University of North Carolina had won the game (and it was only lost in the last 0.5 seconds) then the random forest method

would have obtained 68.2% of the points under the double scoring system and the machine-learning method would have obtained 66.1% of the points. The Bayes model would still have had only 51.6% of the points in the double scoring system.

Overall, it appears that both the machine-learning method and the random forest method do slightly better than the Bayes model using probability self-consistency. We will continue to follow this. It is also noted that predicting an entire bracket is harder than predicting individual games such as in the research by Shi *et al.* (2013) and Magel and Unruh (2013). This is because in developing a bracket, all rounds after the first round rely on predictions from the previous rounds.

Future work could involve adding an indicator variable as to whether a team made it to the Sweet Sixteen Round in the past year to all the models, or similar types of indicator variables could be added to the models. Historical probabilities could also be calculated as to what is the probability that a team will make it to Round 32 this year if a team made it to Round32 last year. This could also be done with other rounds, such as the Sweet Sixteen round. The prior could be based on these historical probabilities instead of the historical probabilities associated with seeds.

Funding: This study received no specific financial support.

Competing Interests: The authors declare that they have no competing interests.

Contributors/Acknowledgement: All authors contributed equally to the conception and design of the study.

REFERENCES

Barra, A., 2014. Is March madness a sporting event-or a gambling event? Available from www.theatlantic.com [Accessed March 21, 2014].

Breiter, D. and B. Carlin, 1997. How to play office pools if you must. Chance, 10(1): 5-11.

Magel, R. and S. Unruh, 2013. Determining factors influencing the outcomes of college basketball games. Open Journal of Statistics, 3(4): 225-230.

March Madness, 2014. Available from http://en.wikipedia.org/wiki/march_madness.

NCAA, 2015, 2016. Basketball statistics. Available from http://www.ncaa.com.

Pomeroy, K., 2015, 2016. Pomeroy's ratings. Available from http://www.kenpom.com.

RPI, 2015. Available from http://www.collegerpi.com.

Sagarin, J., 2015, 2016. Basketball statisitics. Available from http://www.usatoday30.usatoday.com/sports/sagarin.htm.

Shen, G., S. Hua, X. Zhang, Y. Mu and R. Magel, 2015. Predicting results of March madness using the probability self-consistent method. International Journal of Sports Science, 5(4): 139-144.

Shi, Z., S. Moorthy and A. Zimmermann, 2013. Predicting NCAAB match outcomes using ML techniques – some results and lessons learned. ECML / PKDD 2013 Workshop on Machine Learning and Data Mining for Sports Analytics, arXiv:1310.3607v1 [cs.LG] 14Oct. 2013.

West, B.T., 2006. A simple and flexible rating method for predicting success in the NCAA basketball tournament. Journal of Quantitative Analysis in Sports, 2(3): 3-8.

West, B.T., 2008. A simple and flexible rating method for predicting success in the NCAA basketball tournament: Updated results from 2007. Journal of Quantitative Analysis in Sports, 4(2): 6-8.