2016 March Machine Learning Mania
Forecasting the 2016 NCAA Men's Basketball Tournament

Tyler Olson, Tom Tran, Alex Zajichek

Abstract

This paper implemented machine learning and statistical modeling methods to predict the outcome of the 2016 NCAA Men's Basketball Tournament-known as "March Madness". After a strategic set of initial predictors were chosen, the following models were explored: Bayesian linear regression, bootstrap least-squares regression, logistic regression, random forest, generalized boosting regression, and neural networks. Stemming from Kaggle's annual *March Machine Learning Mania*, each model had the common purpose of predicting the probability of a given team *i* beating opponent *j*. A threshold was set at 0.5 to classify a game as a 1 (win) or a 0 (loss). The predictive binomial deviance (Kaggle's measure of model performance), classification accuracy, and ESPN's bracket scoring were used for model evaluation. The logistic regression model produced the best predictive binomial deviance at the 80.2^{nd} percentile in Kaggle's competition, random forest classified the largest percent of tournament match-ups correctly at 74.60%, and the bootstrap least-squares model produced the best bracket according to ESPN's bracket scoring at the 99.6^{th} percentile out of approximately 13.02 million submitted brackets for the 2016 tournament. Play-in games were not considered in the model building or evaluation.

1 Introduction

The NCAA Men's basketball tournament is an annual, single-elimination competition traditionally involving 64 teams and 63 games throughout March and April. The winning academic institution is crowned national champion, and the amalgamation of unpredictability, high stakes action, and drama that fills each game is why the American public has fallen in love with "March Madness." Due to this level of interest across the United States, predicting the eventual winners of each tournament game, which involves filling out a tournament bracket, has evolved into a popular social activity. Basic probability has shown that the task of correctly predicting the winners of all 63 match-ups is virtually impossible, so naturally, this elusive objective has found its way into the field of machine learning. Our goal is to use data-driven machine learning algorithms to construct models that will better predict the outcomes of these games. Kaggle holds a "March Machine Learning Mania" competition each spring, in which historical tournament data dating back to 1985 is provided. This data, along with other publicly available information sources, will be used to develop six different models for the purpose of predicting the probability of a given team *i* beating opponent *j*.

2 Related Work

Purpose of Prediction

As the NCAA basketball tournament has grown in popularity and profitability, March Madness has evolved from a sporting event into a gambling event. Approximately 3 billion dollars were wagered on tournament games in 2015, with 40 million Americans participating in office pools, online bracketology competitions, and casino betting. Berkowitz et al. examined the relationship

between additional betting lines and price accuracy in wagering markets [Berkowitz 2014]. When the totals line and money line were both included with the sides line, which reflects the market expectations of the point differential between the two opponents, became more accurate. The total points scored by both teams as well as the a priori probability of each team winning the game both contributed to this increase in accuracy. Carlin recognized the value of contrarian picks, and proposed a model that identified favorable yet underbet teams, in order to provide a positive ROI [Carlin 2005]. This model incorporated ratings including Vegas lines and tournament seeds, team statistics, and an underappreciation statistic. However, the results were quite underwhelming, so contrarian strategies may not be appropriate for accurate prediction within this domain. The sides line and the over/under are the two most popular wagers in sports gambling, so Kain and Logan developed a seemingly unrelated regression structure that incorporated both features [Kain 2011]. Although the sides line was an accurate predictor of the margin of victory, the over/under struggled to predict the sum of scores. Therefore, reliance on sports wagering markets for the purpose of predicting the outcome of games can be unreliable.

Tournament selection and seeding is the earliest catalyst of the madness that occurs throughout March. The inclusion and subsequent ranking of 68 tournament eligible teams limits the potential matchups and paths to the championship game. Thus, once teams are selected, assigned seeds can potentially play a significant role in a team's tournament success. The Offensive-Defensive Model (ODM), introduced by Govan, Langville, and Meyer in 2009, generated team ratings according to the seasonal games that have already occurred [Govan 2009]. A matrix was used to store the number of points generated by a particular team in a particular matchup, and offensive and defensive ratings were calculated according to a team's ability to produce more points and allow less points. While the simplicity of this model is desirable, ignoring other aspects of the game could be leaving prediction accuracy on the table. Kvam and Sokol were much more ambitious, and looked to improve upon existing ranking systems [Kvam 2006]. Using a combined logistic regression/Markov chain (LRMC) model, predictions of tournament outcomes were produced with basic in-game information as the input data. The logistic regression model calculated the probability that a team with a margin of victory of a certain number of points at home is better than its opponent. The Markov chain model assigned each team to an individual state, and then ranked each team. Predictions were made according to the higher-ranked team. Kvam and Sokol discovered that predicting outcomes according to their rating system was much more accurate than predicting according to the AP, USA Today, RPI, Sagarin, and Massey rankings. Based on this result, the margin of victory seems to be an appropriate predictor of success in the tournament.

College basketball experts and novice bracketologists alike are interested in the potential statistical performance of a team or an individual player, for the purpose of evaluating upcoming games once the tournament has started. Given a particular team and its opponent, analysts enjoy predicting the approximate box-scores statistics before the game has even occurred. Whether a certain player will score 30 points or his team will outrebound the opponent by 10, this information is quite indicative of the potential flow of the game. Statisticians have studied the relationship between team wins and an individual player's statistics in an NBA setting, with the goal of measuring each player's marginal product [Berri 1999]. Other researchers have proposed the use of cumulative win probabilities over the duration of a game in order to measure both team and individual player performance [Bashuk 2012]. The purpose here was to calculate a win probability index (WPI) that was indicative of how the game arrived at the final score, not only on the final score of the game. WPI and the cumulative win probabilities (CWP) were predicted according to play-by-play data, as well as the conference matchup, location, and strength of schedule. Related to the work of Bashuk, Chen, Huang, and Thompson modeled each possession throughout the course of a game, with the number of points scored and a vector of players on the court as the model parameters [Chen 2011]. Logit and probit functions were used to estimate a given player's skill rankings according to one-point, two-point, and three-point offense and defense. Given these player rankings, the model was able to predict points per possession with a fair amount of accuracy. Both Bashuk and Chen were reliant on data that described each possession throughout the course of a game. Although the two models performed well, this data is extremely hard to come by for NCAA games.

Filling out a tournament bracket is what makes March Madness so appealing for Americans, and each participating individual has the ultimate goal of correctly predicting all 67 winners. Therefore, the outcome of the game is the most interesting and valuable result. The most obvious way to predict the outcome of a game is by using binary classification algorithms, which was exhaustively tested

by Beckler, Wang, and Papamichael [Beckler 2009]. Linear regression, support vector machine (SVM), logistic regression, and artificial neural networks (ANN) were all used to predict the outcome of NBA games. Linear regression had the highest classification accuracy of 0.7009, followed by logistic regression: 0.6876, SVM: 0.0791, and ANN: 0.6536. All four simple approaches performed fairly well within the specific problem domain of predicting a win or loss. Carlin first identified the advantage of simplicity when he was working to predict regional champions using point spreads and the relative strength of teams fifteen years earlier [Carlin 1994]. He produced significant results using only basic probability theory and linear regression analysis. However, researchers have found that simple models have a limit regarding the level of accuracy that can be achieved. In an attempt to maximize accuracy, Loeffelholz used in-game statistics, such as field goals made/attempted, offensive/defensive rebounds, and steals/block/turnovers from NBA games to train feed-forward, radial basis, probabilistic, and generalized regression neural networks [Loeffelholz 2009]. The optimal subset of features was identified using the knowledge of domain experts and signal-to-noise ratios. The best networks predicted the winning team with an accuracy of 0.7433, which improves upon the accuracy of domain experts (0.6867). If the predicted outcome of a game is desirable, the combination of a variety of machine learning techniques and in-game statistics has been proven to do the job quite effectively.

Data Selection

When it comes to algorithmic prediction of sporting events, the quality, quantity, and consistency of the data is what ultimately brings about accuracy or unreliability. The information used to construct the model(s) plays a pivotal role, whether betting odds, game outcomes, player performance, or team rankings are the goal of the forecast. Berkowitz et al. understood the potential impact of additional sources of information on the accuracy of pricing in gambling markets, and found evidence that the totals line and the money line, along with the sides line, improved the generation of overall betting odds [Berkowitz 2014]. In this case, the addition of more data, which was deemed appropriate for the forecasting environment, allowed the model to become more accurate. The relationship between data features can also have an effect on model performance. Each player's marginal product complements a team's overall production in basketball, and Berri's research involved quantitatively determining the relative value of players [Berri 1999]. He proposed that a player's value could be calculated using a linear combination of points scored, field goal attempts, offensive rebounds, defensive rebounds, assists, steals, blocks, personal fouls, turnovers, and team wins, divided by the team totals for the nine statistics mentioned. While more advanced APBRmetrics such as per-minute production, per-minute team tempo factor, and per-minute defense offered seemingly appropriate information, the correlation present between the traditional statistics and advanced metrics caused a decline in the overall performance of the model. This balance between the quantity and quality of data features needs to be maintained in order to develop useful models.

Information describing in-game team statistics, individual statistics, and tournament seeding has been previously used to predict the outcome of March Madness games. Magel and Unruh examined which of these statistics were the most significant in determining winners and losers, and measured factor significance using logistic and least squares regression [Magel 2013]. They found that the differences between four team statistics: assists, free throw attempts, defensive rebounds, and turnovers, offered the most pertinent information. The use of team rankings and seeds as predictors has also undergone analysis. Motivated by the fact that more than 70% of all teams in the Elite Eight since 1985 have been seeded three or higher, Jacobson et al. studied the relationship between historical win percentages of high-seeded teams and fourth, fifth, and sixth round tournament wins [Jacobson 2009]. Since there was an insignificant difference between these win percentages in the later rounds, sole reliance on the past success/failure of teams according to their tournament seed for the purpose of prediction was discouraged. Yuan et al. built upon the findings of Magel, Jacobson, and other researchers by consolidating various team and player level archival data into more than 30 different models of performance metrics [Yuan 2015]. This group ultimately discovered that their most sucessful models had two distinguishing characteristics: the incoporation of sufficient regularization and the absence of data contamination. Historical data including tournament results as well as regular-season results frequently caused the overfitting of models to results from a particular season, which led to poor predictive performance. This type of data contamination is commonplace in public data sources, and finding isolated regular season information is an issue many researchers face.

Model Development and Evaluation

A wide variety of machine learning algorithms have been implemeneted and tested by research groups interested in predicting betting odds, player performance, and point differentials of NCAA and NBA basketball games.

- Linear Regression [Beckler 2009, Schwetman 1996, Carlin 1994]
- Support Vector Machine (SVM) [Beckler 2009]
- Logistic Regression [Beckler 2009, Lopez 2015, Schwetman 1996, Parker 2010]
- Artificial Netural Network (ANN) [Beckler 2009, Loeffelholz 2009]
- K-Means Clustering [Beckler 2009]
- Outlier Detection [Beckler 2009]
- Markov Models [Strumbeli 2012]
- Offensive-Defensive Model (ODM) [Govan 2009]
- Network Diffusion Model [Melo 2012]
- Logistic Regression/Markov Chain (LRMC) [Kvam 2006]
- K-Nearest Neighbor (KNN) [Hoegh 2015]
- Markov Logic Networks (MLNs) [Orendorff 2007]

In the context of tournament forecasting, there are three forms of evaluation that were commonly used to assess the predictive accuracy of proposed models.

- Classification Accuracy [Beckler 2009, Loeffelholz 2009, Schwetman 1996, Yuan 2015]
- Predictive Binomial Deviance Function [Lopez 2015, Yuan 2015]
- AUC [Yuan 2015]

3 The Proposed Work

Data

The majority of the data used was provided by Kaggle. It consists of detailed game information dating back to 2003. This, along with other sources, were used to carry out the analysis. The majority of our time was spent doing data manipulation to get the data in the form needed to conduct our analysis.

Plan

This paper caries out the following:

1. Research and explore the data to strategically come up with a starting set of basketball statistics that effectively influence the outcome of games.
2. Learn multiple models with a common goal of predicting the outcome of a basketball game.
3. Evaluate the performance of the models.

4 Methodology

Preliminary Variable Selection

The following are the 16 features used in the modeling process:

Variable	Team *i*	Opponent *j*
Seed	w_1	w_{16}
Pythagorean Expectation[1]	w_2	w_9
Effective Field Goal%[3]	w_3	w_{10}
Points per Possesion[5]	w_4	w_{11}
Economy[4]	w_5	w_{12}
Free Throw %	w_6	w_{13}
Rating Percentage Index[2]	w_7	w_{14}
Win %	w_8	w_{15}

Each of the eight listed variables are included for both teams in a match-up. Superscripts reference equations in appendix.

Models

Let

$$Y_{W_{ijk}} = \begin{cases} 1 & \text{if team } i \text{ beats opponent } j \\ 0 & \text{if opponent } j \text{ beats team } i \end{cases}$$

$$\widehat{Y_{W_{ijk}}} = P(\widehat{Y_{W_{ijk}}} = 1) = \text{predicted probability that team } i \text{ beats opponent } j$$

$$\mathbf{w}^T = (w_0, w_1, ..., w_{16}) \leftarrow \text{model parameters}$$

$$\mathbf{x}_{ijk} = k^{th} \text{example of team } i \text{ playing against opponent } j$$

$$Y_{PD_{ijk}} = (\text{team } i\text{'s score - opponent } j\text{'s score}) \text{ is called the } \textit{point differential}.$$

(1) Bayesian Linear Regression (BLR)

$Y_{PD_{ijk}}$ is modeled against the set of chosen predictors via a Bayesian linear regression model. The `R2OpenBUGS` package in `R` is used for Markov Chain Monte Carlo simulation [Sturtz 2005].

Cowles suggests the following prior distributions for a simple MCMC [Cowles 2013].

$$w_m \sim Normal(0, 10^6) \leftarrow \text{Uninformative Prior}$$
$$\tau = \frac{1}{\sigma^2_{Y_{PD_{ijk}}}} \sim Gamma(\alpha = 0.001, \beta = 0.001)$$
$$Y_{PD_{ijk}} | \mathbf{w}, \mathbf{x_{ijk}} \sim N(\mathbf{w}^T \mathbf{x}_{ijk}, \sigma^2_{Y_{PD_{ijk}}})$$

For a predicted point differential $\widehat{Y_{PD_{ijk}}}$, the following predictive distribution is of interest.

$$f(\widehat{Y_{PD_{ijk}}} | \mathbf{x_{ijk}}) = \int_{\mathbf{w}} f(\widehat{Y_{PD_{ijk}}} | \mathbf{w}, \mathbf{x_{ijk}}) f(\mathbf{w} | \mathbf{Y_{PD}}) d\mathbf{w}$$

The outcome of the game directly depends on this value since a positive differential indicates seed *i* wins a game, and a negative point differential indicates they lose. Once the posterior distribution is obtained, the posterior probabilities can be found.

$$P(\widehat{Y_{W_{ijk}}} = 1) = P(\widehat{Y_{PD_{ijk}}} > 0 | \mathbf{x_{ijk}})$$

(2) Bootstrap Least-Squares Regression (BLS)

In this approach, point differential is also modeled.

$$Y_{PD_{ijk}} = \mathbf{w^T x_{ijk}} + \epsilon_{ijk} = w_0 + w_1 x_{ijk1} + ... + w_{16} x_{ijk16} + \epsilon_{ijk}$$

where ϵ_{ijk}'s are independent and identically distributed.

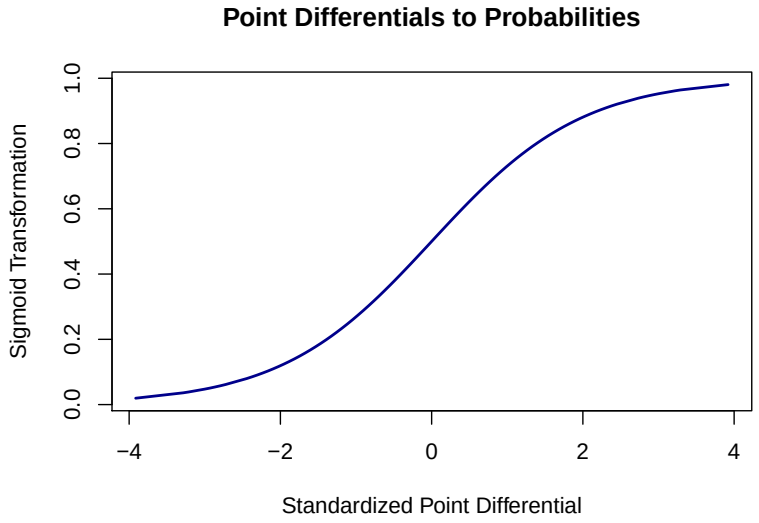
Under these assumptions, a bootstrap technique is performed to estimate the true $\mathbf{w}$. This model is used to predict point differentials, $\widehat{Y_{PD_{ijk}}}$, which leads to win probabilities via the sigmoid function [Turner 2015].

$$\widehat{Y_{W_{ijk}}} = P(\widehat{Y_{W_{ijk}}} = 1) = \frac{1}{1 + e^{-\widehat{Y_{PD_{ijk}}}}}$$

This conversion is intuitive since

$$P(\widehat{Y_{W_{ij}}} = 1) \begin{cases} < 0.5 & \text{if } \widehat{Y_{PD_{ijk}}} < 0 \\ = 0.5 & \text{if } \widehat{Y_{PD_{ijk}}} = 0 \\ > 0.5 & \text{if } \widehat{Y_{PD_{ijk}}} > 0 \end{cases}$$

meaning that negative point differentials predict low probabilities to win, and vice versa.



(3) Logistic Regression (LR)

This approach models the outcome of a game directly. For a given matchup,

$$log\left(\frac{P(Y_{W_{ijk}} = 1)}{P(Y_{W_{ijk}} = 0)}\right) = \mathbf{w^T x_{ijk}} = w_0 + w_1 x_{ijk1} + ... + w_{16} x_{ijk16}$$

$$P(Y_{W_{ijk}} = 1) = \frac{1}{1 + e^{-\mathbf{w^T x_{ijk}}}}$$

In this setting, the subset of **w** that minimizes the Akaike's Information Criteria (AIC) for *n* = number of examples, *SSE* = sum of squares for error, and *p* = number of features in the model is identified,

where,

$$\text{AIC} = nlog\left(\frac{SSE}{n}\right) + 2(p+1)$$

$\widehat{Y_{W_{ijk}}}$ is found based on the final model [Ledolter 2006].

(4) Random Forest (RF)

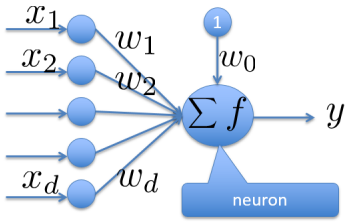
Random Forest (RF) is a refinement of bagged trees. The model is trained by growing many classification trees (e.g. 1,000). Each tree is built as follows: For *N* observations in the training set, sample *N* cases with replacement. At each tree split, a random sample of *m* features is drawn and considered for splitting. The optimal *m* is proved to be $\sqrt{p}$, where *p* is the number of features. Each tree is grown to the largest extent possible without any pruning [Breiman 2001]. The predicted class probability of a test sample is computed as the mean predicted class probabilities of the trees in the model. The class probability of a single tree is the proportion of samples of the same class in a leaf for the training set. All the variables in the final model are kept. The importance of each variable from the model can be retrieved [Breiman 2001].

(5) Generalized Boosting Regression (GBM)

$$P(Y_{W_{ij}} = 1) = \frac{1}{1 + e^{-\mathbf{w^T x_{ijk}}}}$$

As a boosting technique, Generalized Boosting Regression (GBM) is an ensemble model of weak learners. The weak learners are trained as follows: At each stage $1 < m < M$ of training, improve $F_m(x)$ by fitting $h(x)$ to the residual $y - F_m(x)$. The new weak learner $h(x)$ is then added to the current model: $F_{m+1}(x) = F_m(x) + h(x)$. The implementation of AdaBoost in the `R` package `gbm` adopts AdaBoost's exponential loss function, but uses Friedman's gradient descent algorithm. The objective is to to find a regression function, $\hat{F}(x)$, that minimizes the loss function, $\psi(y, F)$: $\hat{F}(x) = argmin_\rho \sum_{i=1}^{n} \psi(y_i, \rho)$ [Ridgeway 2007].

(6) Neural Networks (NNET)



$$P(Y_{W_{ij}} = 1) = \frac{1}{1 + e^{-\mathbf{w^T x_{ijk}}}}$$

Stochastic Gradient Descent is used to minimize the loss function

$$E(\mathbf{w}) = \frac{\mathbf{1}}{\mathbf{2}} \sum_{\mathbf{i=1}}^{\mathbf{n}} ||\mathbf{y}(\mathbf{x_i}, \mathbf{w}) - \mathbf{t_i}||^{\mathbf{2}}$$

by the following back propagation:

$$w^{(\tau+1)} = w^{(\tau)} - \eta \bigtriangledown E(w^{(\tau)})$$

For this simple dataset, neural networks with only one hidden layer is trained. In addition to the number of neurons, a weight decay parameter for regularization is controlled. Cross validation accuracy is used to select the optimal model. The final model has one hidden node and weight decay of one [Yang 2016].

Model Evaluation

There are three methods of evaluation used for each of the six models.

1. Predictive Binomial Deviance

$$PBD = \frac{-1}{n}\sum_{i=1}^{n} Y_{W_{ijk}} log(\widehat{Y_{W_{ijk}}}) + (1 - Y_{W_{ijk}}) log(1 - \widehat{Y_{W_{ijk}}})$$

The PBD is the measure used by Kaggle to score competitors, in which a *smaller* value is better. It takes into account the predicted probabilities as well as the actual result of the game, and penalizes heavily for being "both confident and wrong" [Kaggle 2016].

2. Classification Accuracy

The percentage of games classified correctly that *actually* occurred is found. Although a tournament bracket is generally completed before any games occur, this will give a good measure of how well the models can classify a single game.

3. ESPN Bracket Scoring

For each model, a tournament bracket is simulated. These brackets are scored and ranked according to ESPN's bracket scoring measure, which allocates points per correct pick in each round. For the 2016 NCAA tournament, there were approximately 13.02 million brackets submitted through ESPN.com [ESPN 2016].

Round	1	2	3	4	5	6
Points per pick	10	20	40	80	160	320

5 Results

	BLR	LR	BLS	RF	GBM	NNET
PBD	1.682	.5613	.6084	.5873	.6770	.5696
Matchup %	65.08	71.43	71.43	74.60	69.84	73.02
ESPN	360	870	1380	1140	590	770

Table 1: Raw scores produced by each evaluation method for the six models.

	BLR	LR	BLS	RF	GBM	NNET
PBD	1.4	80.2	44.1	57.7	30.8	74.3
ESPN	3.5	84.5	99.6	98.1	32.0	68.3

Table 2: Percentiles of PBD and ESPN bracket scores for Kaggle and ESPN.com, respectively.

Discussion

The logistic regression model had the best predictive binomial deviance, ranking in the 80.2^{nd} percentile in the Kaggle competition. The random forest produced the largest accuracy by classifying 74.60% of the 2016 tournament games correctly. The bootstrap least-squares model produced a bracket score of 1380, ranking in the top 0.4% of all brackets submitted through ESPN.com.

The goal of the Kaggle competition is not only to classify the outcome of games, but to do so with the highest level of confidence. This system penalizes heavily for being incorrectly confident, leading to a trade-off between the potential for a better predictive binomial deviance, and playing it safe with

less-extreme probabilities. Logistic regression and neural networks were the preferred models for success in the Kaggle competition, but were beat by other models in the ESPN bracket scoring.

The bracket score was determined by filling out a 2016 NCAA tournament bracket based solely upon predicted probabilities obtained for the match-ups, which were then scored according to ESPN's bracket scoring measure. Here, the magnitude of the predicted probabilities did not play an important role, but only whether it exceeded the threshold of 0.5 to classify a game as 1 (win) or 0 (loss). The bootstrap least-squares and the random forest models provided the top brackets, both predicting the national champion correctly.

References

[Bashuk 2012] Mark Bashuk, "Using Cumulative Win Probabilities to Predict NCAA Basketball Performance," *MIT Sloan Sports Analytics Conference*, 2012.

[Bassett 1996] Gilbert W. Bassett, "Predicting the Final Score."

[Beckler 2009] Matthew Beckler, Hongfei Wang and Michael Papamichael, "NBA Oracle," CMU Classwork.

[Berri 1999] David J. Berri, "Who is Most Valuable? Measuring the Players Production of Wins in the National Basketball Association," Nanage. Decis. Econ. 20: 411-427, 1999.

[Breiman 2001] Leo Breiman, "Random Forests", University of California-Berkeley, 2001

[Carlin 1994] Bradley P. Carlin, "Improved NCAA Basketball Tournament Modeling via Point Spread and Team Strength Information"

[Carlin 2005] Jarad B. Niemi, Bradley P. Carlin and Jonathan M. Alexander, "Identifying and Evaluating Contrarian Strategies for NCAA Tournament Pools."

[Chen 2011] Leland Chen, Joseph Huang and Ryan Thompson, "Bayesian Skill Ranking," 2011.

[Cowles 2013] Mary Kathryn Cowles, *Applied Bayesian Statistics: With R and OpenBUGS Examples*, Springer Texts in Statistics, 2013.

[ESPN 2016] "ESPN Tournament Challenge", *NCAA Tournament Challenge Bracket*, ESPN, 2016.

[Govan 2009] Anjela Y. Govan, et al, "Offense-Defense Approach to Ranking Team Sports," *Journal of Quantitative Analysis in Sports*, 2009.

[Gupta 2015] Ajay Andrew Gupta, "A new approach to bracket prediction in the NCAA Men?s Basketball Tournament based on a dual-proportion likelihood," *Journal of Quantitative Analysis in Sports*, De Gruyter, Vol. 11(1), Pages 53-67, March 2015.

[Hoegh 2015] Andrew Hoegh, Marcos Carzolio, Ian Crandell, Xinran Hu, Lucas Roberts, Yuhyun Song and Scotland C. Leman, "Nearest-neighbor matchup effects: accounting for team matchups for predicting March Madness," *Journal of Quantitative Analysis in Sports*, 2015.

[Jacobson 2009] Sheldon H. Jacobson, et al, "Seeding in the NCAA Men?s Basketball Tournament: When Is A Higher Seed Better?", *The Journal of Gambling Business and Economics*, Vol. 3, No. 2, pp. 63-87, 2009.

[Jurasinski 2014] Gerald Jurasinski, Franziska Koebsch, Anke Guenther and Sascha Beetz (2014). flux: Flux rate calculation from dynamic closed chamber measurements. R package version 0.3-0. https://CRAN.R-project.org/package=flux.

[Kaggle 2016] "March Machine Learning Mania 2016.", *Evaluation*, Kaggle, 11 February 2016.

[Kain 2011] Kyle J. Kain and Trevon D. Logan, "Are Sports Betting Markets Prediction Markets? Evidence from a New Test," January 2011.

[Koenker 2010] Roger Koenker and Gilbert W. Bassett Jr. "March Madness, Quantile Regression Bracketology, and the Hayek Hypothesis", *Journal of Business and Economic Statistics*, Vol. 28, Iss. 1, 2010.

[Kubatko 2007] Justin Kubatko, et al, "A Starting Point for Analyzing Basketball Statistics", *Journal of Quantitative Analysis in Sports*, 2007.

[Kvam 2006] Paul Kvam and Joel S. Skol, "A Logistic Regression/Markov Chain Model for NCAA Basketball", *Naval Research Logistics*, Vol. 53, Pages 788-803, 2006.

[Ledolter 2006]B Abraham and J Ledolter, "Introduction to Regression Modeling", Duxbury Press, 2006.

[Loeffelholz 2009] Bernard Loeffelholz, Earl Bednar and Kenneth W. Bauer, "Predicting NBA Games Using Neural Networks," *JQAS*, 2009.

[Lopez 2015] Michael J. Lopez and Gregory J. Matthews, "Building an NCAA men's basketball predictive model and quantifying its success," *JQAS*, 2015.

[Magel 2013] Rhonda Magel and Samuel Unruh, "Determining Factors Influencing the Outcome of College Basketball Games," *Open Journal of Statistics*, Vol. 3 No. 4, Pages 225-230, 2013.

[Melo 2012] Pedro O. S. Vaz De Melo, Virgilio A. F. Almeida, Antonio A. F. Loureiro, and Christos Faloutsos, "Forecasting in the NBA and Other Team Sports: Network Effects in Action," *ACM Transactions on Knowledge Discovery from Data*, Vol. 6, No. 3, Article 13, October 2012.

[Orendorff 2007] David Orendorff and Todd Johnson, "First-Order Probabilistic Models for Predicting the Winners of Professional Basketball Games," *JQAS*, 2007.

[Page 2007] Garritt L. Page, Gilbert W. Fellingham and C. Shane Reese, "Using Box-Scores to Determine a Position's Contribution to Winning Basketball Games," *Journal of Quantitative Analysis in Sports*, Volume 3, Issue 4, Article 1, 2007.

[Parker 2010] Ryan J. Parker, "Modeling Basketball's Points per Possession With Application to Predicting the Outcome of College Basketball Games," College of Charleston.

[R 2015] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[Ridgeway 2007] Greg Ridgeway, "Generalized Boosted Models: A guide to the 'gbm' package", 2007.

[Ruiz 2015] Francisco J. R. Ruiz and Fernando Perez-Cruz, "A generative model for predicting outcomes in college basketball", *Journal of Quantitative Analysis in Sports*. Volume 11, Issue 1, Pages 39-52, February 2015.

[Schwertman 1996] Neil C. Schwertman, Kathryn L. Schenk and Brett C. Holbrook, "More Probability Models for the NCAA Regional Basketball Tournaments," *The American Statistician*, Vol. 50, No. 1., pp. 34-38, February 1996.

[Sokol 2006] Joel S. Sokol and Paul Kvam, "A Logistic Regression/Markov Chain Model For NCAA Basketball," *Naval Research Logistics*, 53, 2006.

[Sokol 2012] Joel S. Sokol, Mark Brown, Paul Kvam and George Nemhauser, "Insights from the LRMC Method for NCAA Tournament Prediction," *MIT Sloan Sports Analytics Conference*, 2012.

[Stanke 2012] Luke Stanke, "Can Statistical Models Out-predict Human Judgment? Comparing Statistical Models to the NCAA Selection Committee," *MIT Sloan Sports Analytics Conference*, 2012.

[Strumbelj 2012] Erik Strumbelj and Petar Vracar, "Simulating a basketball match with a homogeneous Markov model and forecasting the outcome," *International Journal of Forecasting*, 28, 532-542, 2012.

[Sturtz 2005] Sturtz, S., Ligges, U., and Gelman, A. R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software*, 12, 1-16, 2005.

[Turner 2015] Scott Turner, "Net Prophet", *Kaggle Competition: From Point Spreads to Win Percentage*, 20 February 2015.

[Yang 2016] Tianbao Yang, Neural Networks [Powerpoint slides]. Spring 2016.

[Yuan 2015] Lo-Hua Yuan, Anthony Liu, Alec Yeh, Aaron Kaufman, Andrew Reece, Peter Bull, Alex Franks, Sherrie Wang, Dmitri Illushin and Luke Bornn, "A mixture-of-modelers approach to forecasting NCAA tournament outcomes," *Journal of Quantitative Analysis in Sports*, 2015.

Appendix

$$\text{Pythagorean Expectation} = \frac{\text{PointsFor}^{13.91}}{\text{PointsFor}^{13.91} \times \text{PointsAgainst}^{13.91}} \tag{1}$$

$$\text{RPI} = (\text{WP} \times 0.25) + (\text{OWP} \times 0.5) + (\text{OOWP} \times 0.25) \tag{2}$$

$$\text{EFG\%} = \frac{2FGM + 0.5 \times 3FGM}{FGA} \tag{3}$$

$$\text{Economy} = \text{AST+ STL - TO} \tag{4}$$

$$\text{Possessions} = FGA + 0.475 \times FTA - ORB + TO \tag{5}$$