Predicting the 2016 NCAA Men's Basketball Tournament

Tyler Olson Tom Tran Alex Zajichek
CS: 4980 Spring 2016

May 5, 2016

Table of contents

The Problem


First Round
MARCH 17–18
Second Round
MARCH 19–20
Sweet 16
MARCH 24–25
Elite Eight
MARCH 26–27
Final Four
APRIL 2
Final Four
APRIL 2
Elite Eight
MARCH 26–27
Sweet 16
MARCH 24–25
Second Round
MARCH 19–20
First Round
MARCH 17–18
11 Vanderbilt (19-13)
11 Wichita State (24-8)
16 Holy Cross (14-19)
16 Southern (22-12)
First Four
DAYTON
MARCH 15–16
11 Michigan (22-12)
11 Tulsa (20-11)
16 Florida Gulf Coast (20-13)
16 Fairleigh Dickinson (18-14)
2016 NCAA TOURNAMENT
FINAL FOUR
HOUSTON
APRIL 2 AND 4
CHAMPIONSHIP
APRIL 4
SOUTH
EAST
WEST
MIDWEST
1 Kansas (30-4)
16 Austin Peay (18-17)
8 Colorado (22-11)
9 UConn (24-10)
5 Maryland (25-8)
12 South Dakota State (26-7)
4 California (23-10)
13 Hawai'i (27-5)
6 Arizona (25-8)
11 Vanderbilt/Wichita State
3 Miami (Fla.) (25-7)
14 Buffalo (20-14)
7 Iowa (21-10)
10 Temple (21-11)
2 Villanova (29-5)
15 UNC Asheville (22-11)
1 Oregon (28-6)
16 Holy Cross/Southern
8 Saint Joseph's (27-7)
9 Cincinnati (22-10)
5 Baylor (22-11)
12 Yale (22-6)
4 Duke (23-10)
13 UNC Wilmington (25-7)
6 Texas (20-12)
11 Northern Iowa (22-12)
3 Texas A&M (26-8)
14 Green Bay (23-12)
7 Oregon State (19-12)
10 VCU (24-10)
2 Oklahoma (25-7)
15 Cal State Bakersfield (24-8)
1 North Carolina (28-6)
16 Florida Gulf Coast/FDU
8 USC (21-12)
9 Providence (23-10)
5 Indiana (25-7)
12 Chattanooga (29-5)
4 Kentucky (26-8)
13 Stony Brook (26-6)
6 Notre Dame (21-11)
11 Michigan/Tulsa
3 West Virginia (26-8)
14 Stephen F. Austin (27-5)
7 Wisconsin (20-12)
10 Pittsburgh (21-11)
2 Xavier (27-5)
15 Weber State (26-8)
1 Virginia (26-7)
16 Hampton (21-10)
8 Texas Tech (19-12)
9 Butler (21-10)
5 Purdue (26-8)
12 Little Rock (29-4)
4 Iowa State (21-11)
13 Iona (22-10)
6 Seton Hall (25-8)
11 Gonzaga (26-7)
3 Utah (26-8)
14 Fresno State (25-9)
7 Dayton (25-7)
10 Syracuse (19-13)
2 Michigan State (29-5)
15 Middle Tennessee (24-9)
Sports Illustrated
NCAA is a registered trademark of the National Collegiate Athletic Association.

Kaggle Competition

March Machine Learning Mania 2016

1. Use provided data and other sources to create a predictive model
2. Estimate the probability that team i beats opponent j for all 2278 combinations of the 68 tournament teams in 2016
3. Get scored based on the actual results of the tournament

The Data

- Subset of the original data
  - 2003/2004 season through 2015/2016 season
- Separate regular season and tournament data
- Traditional basketball statistics
  - Field goals made, field goals attempted, free throws made, free throws attempted, rebounds, etc.

Related Work

- Purpose of Prediction
  - Betting odds, selection, performance, outcomes
- Data selection
  - Quality v.s. quantity, correlated statistics, regularization, contaminated data
- Model development and evaluation
  - Primarily supervised learning methods
  - Classification accuracy, predictive binomial deviance, AUC

Our Approach

- For each tournament matchup, model the outcome of the game with the two teams' regular season information
- Two considerations for response:
  1. Win/Loss (1 or 0)
  2. Point Differential (team i's score - opponent j's score)

Preliminary Variable Selection

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

Models Considered

1. Bayesian Linear Regression (BLR)
2. Logistic Regression (LR)
3. Bootstrap Linear Regression (BLS)
4. Random Forest (RF)
5. Generalized Boosted Regression (GBM)
6. Neural Network (NNET)

Notation

Let

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

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

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

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

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

Bayesian Linear Regression

Prior Distributions [Cowles 2013]

$$w_m \sim Normal(0, 10^6) \leftarrow \text{Uninformative Prior}$$

$$Y_{PD_{ijk}} | \mathbf{w}, \mathbf{x_{ijk}} \sim Normal(\mu_{Y_{PD_{ijk}}} = \mathbf{w}^{\mathbf{T}} \mathbf{x_{ijk}}, \sigma^2_{Y_{PD_{ijk}}})$$

Predictive Distribution

The distribution for a new prediction was then obtained via `R2OpenBUGS` [Sturtz 2005].

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

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

Logistic Regression

Let

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

Then,

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

Methodology :

1. Model all possible subsets of predictors
2. Choose model with *lowest* AIC (Akaike's Information Criterion) [Ledolter 2006]
3. Estimate the probability of team i beating opponent j

$$\widehat{Y_{W_{ijk}}} = P(\widehat{Y_{W_{ijk}} = 1})$$

Bootstrap Least-Squares Regression

Let

$$Y_{PD_{ijk}} = \text{team i's score - opponent j's score}$$

Then

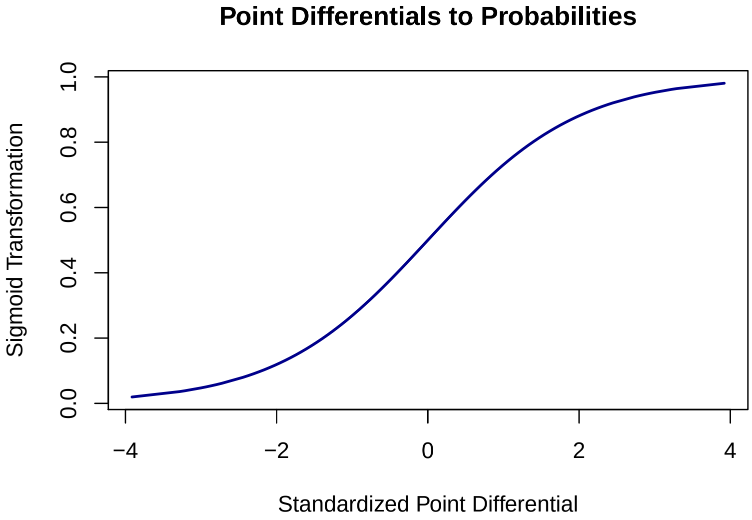
$$Y_{PD_{ijk}} = \mathbf{w}^{\mathbf{T}}\mathbf{x_{ijk}} = w_0 + w_1 x_{ijk1} + \ldots + w_{16} x_{ijk16} + \epsilon_{ijk}$$

Methodology :

1. Find LS estimates for each of 100,000 bootstrap samples
2. Average LS estimates over all bootstrap models
3. Convert predicted point differentials to probabilities via the sigmoid function [Turner 2015]:

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

Bootstrap Least-Squares Regression



Random Forest

Let

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

Methodology :

- ensemble technique, refinement of bagged trees
- at each tree split, a random sample of m features is drawn and considered for splitting
- $m = \sqrt{p}$ where p is the number of features

Predicted class probability = mean predicted class probabilities of the trees or by votes

[Breiman 2001]

Generalized Boosted Regression

Let

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

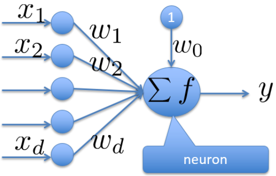
$$P(Y_{W_{ijk}} = 1) = \frac{1}{1 + e^{-\mathbf{w}^\mathbf{T}\mathbf{x}_\mathbf{ijk}}}$$

Methodology :

- ensemble of weak prediction models
- gradient descent algorithm
- at each stage $1 < m < M$, improve $F_m(x)$ by fitting $h(x)$ to the residual $y - F_m(x)$
- add $h(x)$ to the current model: $F_{m+1}(x) = F_m(x) + h(x)$

[Ridgeway 2007]

Neural Networks



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

Methodology :

- stochastic gradient descent
- back propagation
- one hidden layer

[Yang 2016]

Methods of Evaluation

1. Predictive Binomial Deviance [Kaggle 2016]

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

*Scoring measure used in Kaggle competition

2. Percent of correct picks by match-up

3. ESPN Bracket Scoring [ESPN 2016]

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

*13.02 million brackets submitted this year

Results

Scores

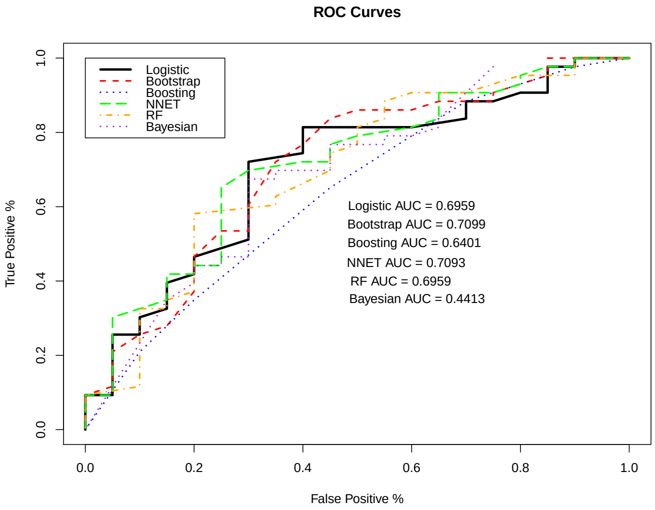
	BLR	LR	BLS	RF	GBM	NNET
PBD	1.682	.5613	.6084	.5873	.6770	.5696
Matchup %	65.08	71.43	71.43	74.60	69.84	73.02
ESPN	360	870	1380	1140	590	770

Percentiles

	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

*MCMC did not converge

ROC curves and AUC



Citations

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

[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.

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

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

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

Citations

[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.