Big Dance: Analysis of Bracketology

Kara Cocco, David Agard and Joseph Nolan

DEPARTMENT OF MATHEMATICS & STATISTICS

NKU NORTHERN KENTUCKY UNIVERSITY

Introduction

For some fans, filling out their bracket for the National Collegiate Athletic Association (NCAA) men's college basketball tournament can be a sport in itself. An array of variables may affect the outcome for any given game, but for this research the main purpose is to determine which variables are the most important in predicting the outcome for the "favorite, or higher seeded team. This project addressed two main goals:

- ✓ Create a model to estimate the probability of victory for the higher seeded team in any given game
- ✓ Develop strategies to create brackets that yield the best chance to win the annual bracket challenge

Data/Variables

Data from the 1985-2013 NCAA men's basketball seasons were used in this study. The following variables were considered for each team:

- Team Seed
- Win/Loss Percentage
- Points Per Game (PPG)
- Opponent Points Per Game (O-PPG)
- Margin of Victory (MOV)
- Strength of Schedule (SOS)
- Simple Rating System (SRS)
- Offensive & Defensive SRS

Each of the above variables will be incorporated into the model as a "difference" between the higher and lower seeded team

Based upon work by Dean Oliver, author of *Basketball on Paper*, four additional variables were considered (data available only from 2001-2013):

- Effective Field Goal Percentage
- Free Throw Percentage
- Turn Over Percentage
- Offensive & Defensive Rebound Percentage

Data were obtained from the following sources:

- http://statsheet.com/mcb
- http://www.allbrackets.com
- http://www.basketball-reference.com

Method/Results

Backward stepwise logistic regression (with significance level 0.05) was used to form a model based on data from 1985-2010. The remaining three years of data were withheld and used for validation.

Logistic Regression Table

Predictor	Coef	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-0.13028	0.272			
MOV_Diff	0.05058	0.007	1.05	1.01	1.09
SOS_Diff	0.40735	0.000	1.50	1.42	1.59
Seed_Diff	0.21901	0.000	1.24	1.17	1.32
WL%_Diff	0.14929	0.000	1.16	1.13	1.19

Figure 1: The final predicted model determined that the 'difference' columns for the following variables: Margin of Victory, Strength of Schedule, Seed, and Win/Los Percentage are statistically significant.

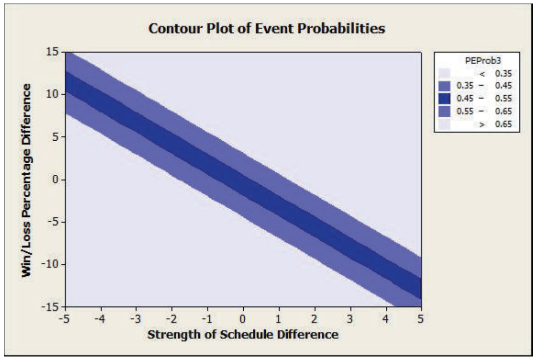
Odds Ratio Example: The odds of correctly predicting the outcome for the higher seeded team are between 1.17 and 1.32 times larger when the seed difference increases by 1.

Validation Results: Using years 2011-2013 data, the model correctly assigned a greater than 50% probability to 146 out of 189 eventual winners. This results in a 77.25% accuracy rate. In particular, the model correctly predicted 29 upsets.

Figure 2: A contour plot of the event probabilities based on seed difference of 1 with win/loss percentage difference vs. strength of schedule difference from the high seed to the low seed.

Example 1: Based on the model, when the seed difference is 1 and there is a 0 difference in the win/loss percentage, as well as strength of schedule, the probability of the higher seeded team winning is estimated to be 52.51%. When the lower seed has a win/loss percentage 1% higher, the chance for a win for the higher seed drops to 48.42%.

Example 2: Based on the model, when the seed difference is 1 and the "favorite" is rated to have a 10 point higher strength of schedule, with a -10 point difference in win/loss percentage, the probability of winning is 92.28%. Whereas, with the same 10 point difference in strength of schedule, but +10 point difference in win/loss percentage difference, the "favorite" has a 99.68% probability of winning.



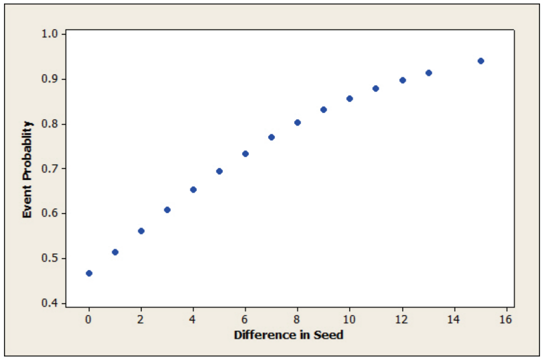


Figure 3: Simple scatterplot using a single variable ONLY of 'difference in seed', and comparing that to the event probability. Logistic regression determined event probabilities based on difference in seeds.

Example 1: A seed difference of 1 gives the "favorite" a 51.84% event probability of winning.

Example 2: A seed difference of 7 gives the "favorite" a 76.39% event probability of winning.

Example 3: A seed difference of 12 gives the "favorite" an 88.96% event probability of winning.

Modern Approach: An alternative approach considers (additionally) the variables suggested by Oliver. In this model margin of victory was replaced by turnover percentage with the other selected variables remaining the same. Model validation resulted in a similar accuracy rate.

Strategy for Bracket

The model estimates the probability of the event:
The higher seed wins → P

Iterative Approach:

1. Predict first round based on model favorite.
2. Use first round model winners to predict second round
3. Repeat for each additional round.

Single Entry Bracket:

1. If P < .5 Then, pick the upset
2. If P ≥ .5 Then, keep higher seed

Multiple Entry Brackets:

Bracket #1

1. Use single entry approach

Bracket #2

1. If P < .35 Then, pick the upset
2. If .35 < P < .65 Then, flip a coin
3. If P ≥ .65 Then, keep higher seed

Bracket #3

1. If P < .40 Then, pick the upset
2. If .40 < P < .60 Then, flip a coin
3. If P ≥ .60 Then, keep higher seed

Bracket #4

1. If P < .45 Then, pick the upset
2. If .45 < P < .55 Then, flip a coin
3. If P ≥ .55 Then, keep higher seed

Potential Modification: Use a biased coin reflecting estimated probability for the higher seed.

Call me in 2014: (859) 555 - 1234

Future Work

- ✓ Compare model performance against syndicated prognosticators (e.g. Sagarin)
- ✓ Model a quantitative response like difference in game score

Acknowledgements: This research was a product of a UR-STEM summer undergraduate research experience supported by *NKU FORCE: Focus on Occupations, Recruiting, Community, and Engagement.*