Building Perfect Tournament Brackets with Data Analytics

Christopher Hagmann, Dr. Nan Kong

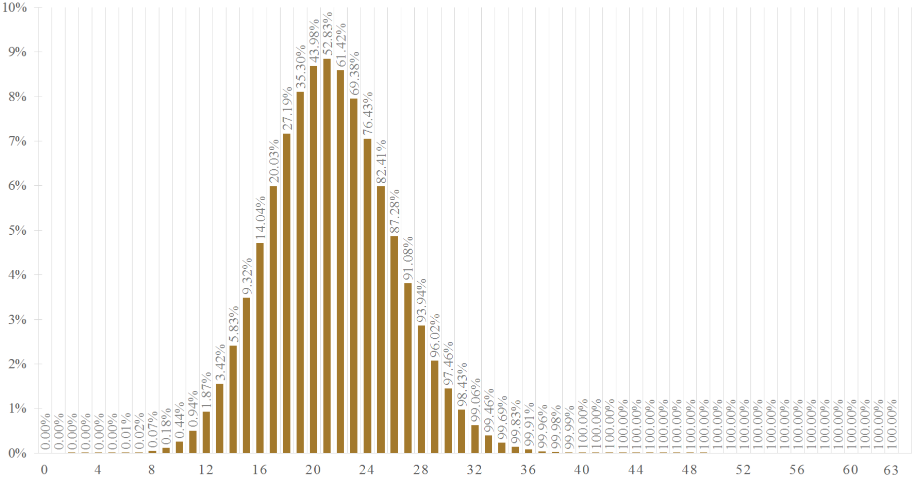
Purdue University

November 12, 2014

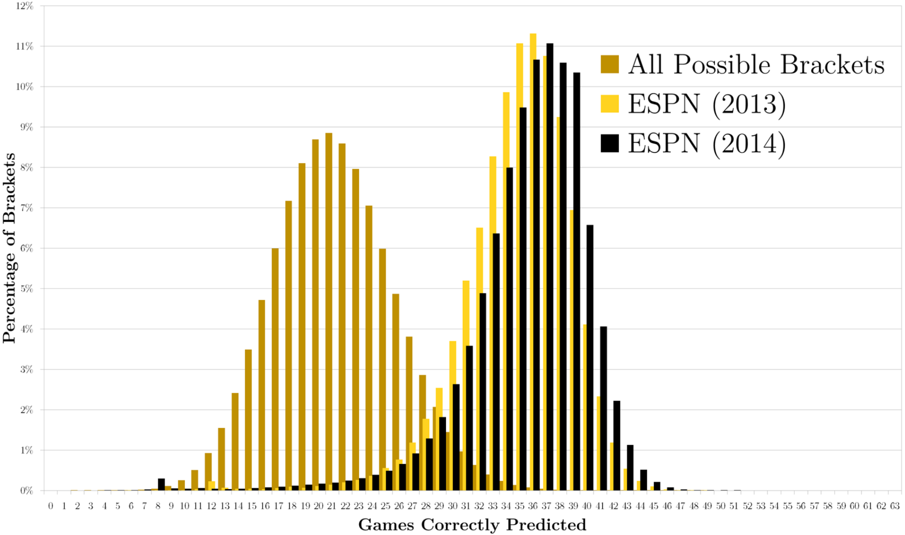
NCAA D-I men's college basketball tournament

- Every March, the National Collegiate Athletic Association (NCAA) has collegiate basketball tournaments
- The most prestigious involve Men's Division I (D-I) "March Madness" tournament
- Last year, $1,000,000,000 was offered for correctly predicting all 63 match outcomes

Distributions of Possible Brackets



Distributions of ESPN Tournament Brackets



Tempo-Free Statistics

- Statistics based on possessions instead of games
- Allows for better comparisons as the tempo, or number of possessions a team has in a game, varies greatly among players

Example

Suppose Player A and Player B both average 20 points per game, but the tempo of their respective teams is 60 and 80 possessions per game. This mean Player A is averaging 2 points every 6 possessions while Player B is only averaging 2 points every 8 possessions. Put this way, it is easier to see that Player A is more effective at converting possessions to points than Player B.

Tempo-Free Statistics

Definitions

Adjusted Offensive Efficiency (AdjO)

The adjusted offensive efficiency is an estimate of the offensive efficiency (points scored per 100 possessions) a team would have against the average D-I defense.

Adjusted Defensive Efficiency (AdjD)

The adjusted defensive efficiency is an estimate of the defensive efficiency (points allowed per 100 possessions) a team would have against the average D-I offense.

Pythagorean expectation (Pyth)

The Pythagorean expectation is a descriptive statistic that combines a team's AdjO and AdjD and is an estimate a team's expected winning percentage against an average D-I team.

Tempo-Free Statistics

Ballpark feel

	Best	Worse
AdjO	125.7	97.3
AdjD	86.9	108.3
Pyth	.9540	.3552

Picking Chalk

Chalk Brackets

This is a term in sports betting that has its origins back to the days that betting odds were written of chalkboards and refers to always choosing the higher rank team to win. As each team in the tournament bracket is explicitly ranked, using this methods yields a unique bracket that is call the chalk bracket. While this is a very simple idea, it usually fares well in small bracket pools.

Log5

Log5 is a method of estimating the probability one team will beat another based on their true winning percentages. In basketball, the Pythagorean expectation is used. It gets its name from it appearance to logit functions and because $p_L \equiv .500$

$$p_{a,b} = \frac{\frac{p_a}{1-p_a}}{\frac{p_a}{1-p_a} + \frac{p_b}{1-p_b} * \frac{p_L}{1-p_L}} = \frac{p_a(1-p_b)}{p_a(1-p_b) + p_b(1-p_a)}$$

Log5

Notable Properties

- If $p_A = 1$, Log5 will always give A a 100% chance of victory
- If $p_A = 0$, Log5 will always give A a 0% chance of victory
- If $p_A = p_B$, Log5 will always return a 50% chance of victory for either side
- If $p_A = 1/2$, Log5 will give A a $1 - p_B$ probability of victory.

Interdependence of Games

There are two ways to look at the likelihood of a team being predicted to advance

Conditional Calculate the probability of a team advancing to the next round conditional on all possible games that could occur.

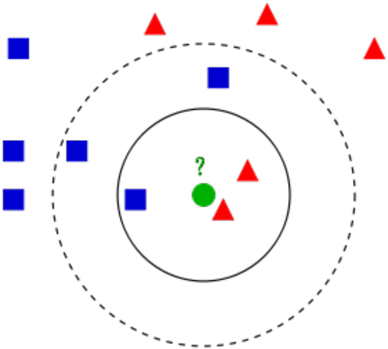
Independent Calculate the probability based on who is actually in the match, independent of what has happened or what will happen latter.

Proposed Method

- We propose a new method for predicting the perfect bracket.
- This involve using k-Nearest Neighbor (kNN) to cluster a team's current opponent with their previous opponents to determine a modified winning percentage.
- This new winning percentage is used in the Log5 to detering the teams odds of winning

kNN Algorithm

- kNN is a method of regression used to determine a value of a property based on the value of this property in *k* neighbors
- The point being examined is a team's current opponent.
- the possible neighbors are the team's that they have already played
- The property is the outcome of the game (W/L)



kNN Algorithm

- The outcome is weighted based on distance
- This returns a win and loss score
- This is repeated for opponent
- These are aggregated to determine a new winning percentages

$$p_a = \frac{W_a + L_b}{W_a + W_b + L_a + L_b}$$

Finding k

- The data for the 2003 - 2013 season was provided by Ken Pomeroy and the above methods were applied using k values of between 3 and 25. 25 is the shortest season played, and therefore, the largest neighborhood similar to all teams.
- It was found that $k = 25$ using the conditional method had the highest average number of games corrected predicted, followed by $k = 25$ using the independent method.
- Second was $k = 8$

Note for Comparisons

- kNN method performed poorly when compared to standard Log5 method in 2003-2013, average with $k = 25$ around 39.5 games correct compared to around 41.1 for Log5.
- Parameter in Pythagorean exception was fitted to 2003-2013 data using Log5 as measure.
- As such 2003-2013 was training set for both.
- Chalk averaged 40.9 games correct over this time period.

Results for 2014 Tournament

Bracket	Games Correct
FiveThirtyEight	40
Dick Vitale	41
President Obama	40
"Chalk"	39
Conditional kNN (k=25)*	42
Independent kNN (k=25)*	40
Log5 (Both Methods)	38

*NOTE: $k = 25$ was in fact the best k

Possible Extensions

- Add additional attributes in the distance formula
- Allow for “upsets”
- Compare to other regression/classification methods

Acknowledgment

Special thanks to Ken Pomeroy for providing me with data

Questions?

Thank You!