Major Themes in Economics

Volume 16 Article 5

Spring 2014

The Final Four Formula: A Binary Choice Logit Model to Predict the SemiFinalists of the NCAA Division I Men's Basketball Tournament

Cameron Fuqua
University of Northern Iowa

The Final Four Formula: A Binary Choice Logit Model to Predict the Semi-Finalists of the NCAA Division I Men's Basketball Tournament

Cameron Fuqua*

ABSTRACT. The NCAA Division I men's basketball tournament is one of the most popular sporting events in America. This paper dissects the tournament and attempts to accurately predict the four semi-finalists ("the final four") using a binary choice logit model. The model does better than any current rating system at predicting these four teams. This paper also examines some common issues about predicting college basketball as a whole. Overall, this paper provides a insights for selection committees, participants in office pools, and coaches to help them achieve their own individual goals.

I. Introduction

This year Warren Buffet offered 1 billion dollars if someone correctly predicted all 63 games of the NCAA Division I Men's basketball tournament, affectionately known as "March Madness." That sounds pretty good, but most estimates put the probability of predicting a perfect bracket at 1 in 128 billion (Woodruff 2012). A 2012 article in Business Insider estimates that 80 to 90 million dollars are gambled legally every year for March Madness. This ranks second to the Superbowl in terms of betting on sporting events (Woodruff 2012). Time magazine also reported the intangible cost to businesses of lost time due to employees being preoccupied with March Madness. Time estimated the cost in 2013 at $134 million, with an average worker watching 1-3 hours of college basketball during work hours in just the first two days (Sanburn 2013). While the promise of $1 billion is quite alluring, this paper does not attempt to perfectly predict the NCAA basketball tournament bracket. It attempts to determine the factors that are most important for a team to reach the semi-finals of the NCAA Division I basketball tournament by using a binomial choice logit model.

*Thank you to Professor Ken Brown, Ken Pomeroy, and Professor Mark Glickman for their assistance and guidance on this paper.

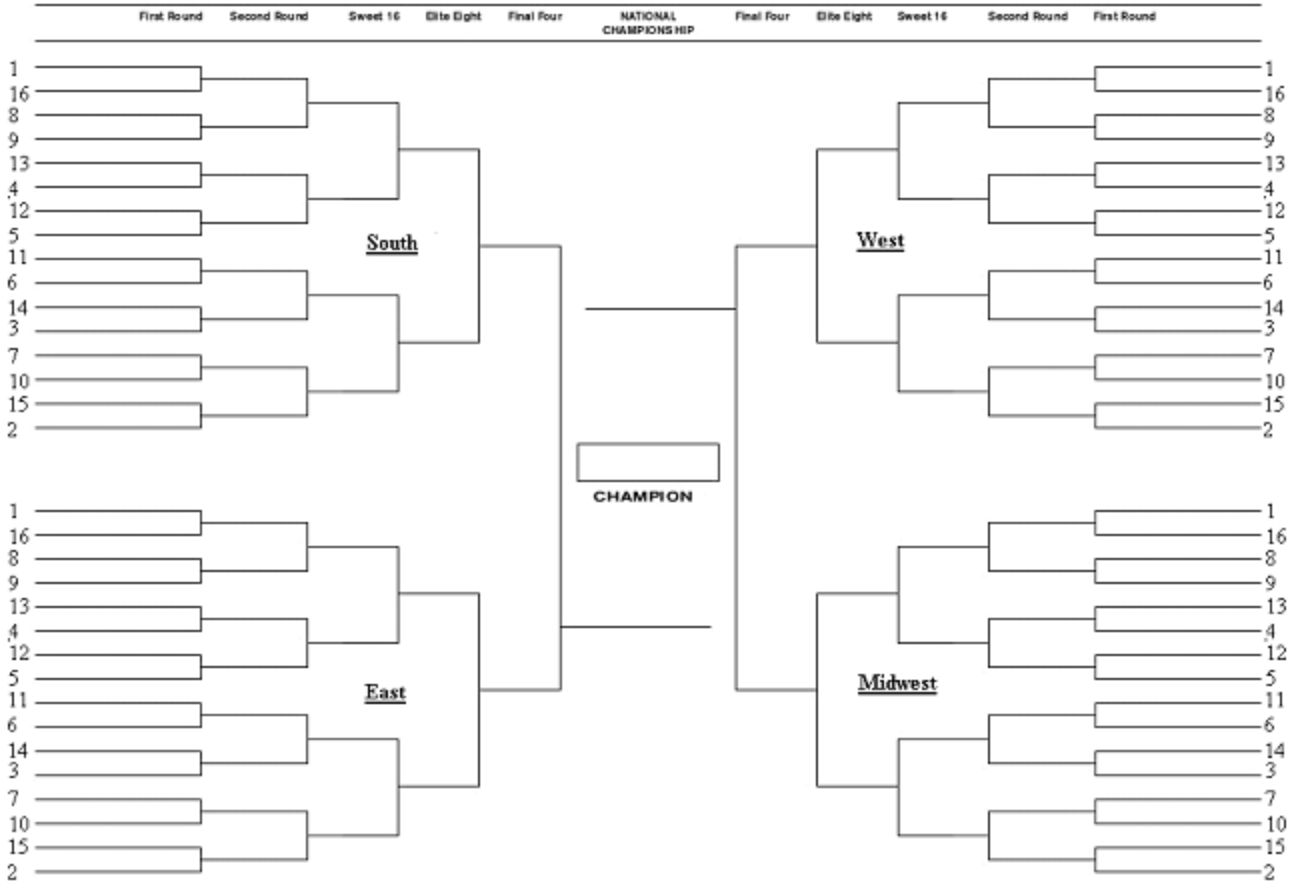
II. The NCAA Basketball Tournament Structure

Tournament Structure and Seeding

The NCAA basketball tournament consists of 68 teams divided into 4 regions playing in a single elimination format. These regions are commonly labelled the Midwest, West, East, and South. Each region is set up so that the 1 seed will not face the 2 seed in its respective region until the regional finals.

There are 4 play-in games called the "First Four". These games consist of the four lowest seeded at-large bids, and the four lowest seeded automatic bids. The seeding and region of these play-in games vary year to year. For 2014 the four play-in games were: 11 seed Tennessee vs 11 Seed Iowa in the Midwest region, 12 Seed Xavier vs 12 Seed North Carolina State in the Midwest Region, 16 Seed Texas Southern vs 16 Seed California Polytechnic University in the Midwest Region, and 16 Seed Mount St. Mary's vs 16 Seed Albany in the South Region. Due to available data, and increased complexity this paper will ignore these games and consider them merely regular season games.

The second round consists of 32 games. The third round is referred to as the regional quarterfinals and consists of 16 games. The fourth round is referred to as the Sweet Sixteen or the regional semi-finals and consists of 8 games. The fifth round is the Elite Eight, or the Regional Finals, and consists of 4 games. The winner of the Elite Eight games advance to the final four. This paper attempts to consistently predict the teams that advance to the final four better than other models.


First Round
Second Round
Sweet 16
Elite Eight
Final Four
NATIONAL CHAMPIONSHIP
Final Four
Elite Eight
Sweet 16
Second Round
First Round
South
West
East
Midwest
CHAMPION
1
16
8
9
13
4
12
5
11
6
14
3
7
10
15
2

Selection Committee

The NCAA Division I men's basketball tournament is controlled directly by the NCAA Division I men's basketball selection committee. This committee is comprised of 10 individuals who serve five year terms. These individuals are athletic directors and conference commissioners in Division I athletics. The NCAA attempts to have these representatives evenly distributed geographically. The primary principle of the selection committee is to "ensure that fair and equitable criteria are used to select the most deserving at-large teams, while also focusing on administering a fair and equitable tournament by creating a nationally-balanced bracket comprised of the most deserving at-large teams and automatic-qualifiers chosen by conferences, while assigning institutions to sites as near to their campuses as possible" (NCAA 2014).

Selection Process

The 68 teams must be selected prior to seeding. There are 32 automatic bids. These 32 bids are each of the individual conference tournament champions. (See Table 1) This leaves 36 teams that will receive an at-large bid determined by the selection committee. The committee relies on strength of schedule, winning percentage, RPI (Ratings Percentage Index, a measure of a team's overall quality), and several subjective factors to properly seed teams (NCAA 2014).

TABLE 1–NCAA Division I Basketball Conferences

Major Conference	Mid-Major Conference	Small Conference
Atlantic Coast	American Athletic	America East
Big 12	Atlantic 10	Atlantic Sun
Big East	Colonial	Big Sky
Big 10	Conference USA	Big South
Pacific-12	Horizon League	Big West
Southeastern	Mid-American	Ivy League[1]
	Missouri Valley	Metro Atlantic
	Mountain West	Mid-Eastern
	West Coast	Northeast
	Western	Ohio Valley
		Patriot League
		Southland
		Southwestern
		Sun Belt
		Summit League

[1]The Ivy League does not have a conference tournament; the regular season champion earns an automatic bid.

Seeding

The selection committee has three goals when seeding teams; it aims to have all four regions evenly balanced, it aims to reduce travel costs for higher-seeded teams, and it aims to limit the possibility of regular season rematches. Attempting to fulfill all three of these criteria often creates tradeoffs in seeding.

The selection committee first ranks the teams from 1 to 68, with 1 being the best and 68 being the worst. This is called true seeding. After seeding teams 1-68, the committee attempts to place teams on an S-curve. This is so each region will be equally balanced. For example the region with the best one seed will have the worst 4 seed. While this is a starting point for the selection committee it does not always hold true. Occasionally, travel expenses, avoiding in-conference matchups and other subjective factors may play a role in seeding teams (NCAA 2014).

III. Who Cares?

There are several useful applications of a predictive NCAA basketball tournament model. The three primary uses are improving seeding efficiency, improving bracket selections for better success in general office pools, and improving a coach's knowledge about what style of basketball increases the chance of tournament success.

Inefficient Seeding

If the NCAA tournament were efficiently seeded, then the final four would always have the highest seeded team from each region (all one seeds). Only once in the history of the tournament have all four number one seeds reached the final four (2008). If all four number one seeds reach the final four the sum of these seeds would be four (1+1+1+1). Over the past eight years the sum of the seeds of the final four has averaged 12.875. This is far above a perfectly efficient bracket.

Second, if a bracket were seeded efficiently, there would be zero upsets. Due to the nature of sports and college basketball it is not likely to have a tournament with zero upsets. While there is no standard number of upsets that is considered acceptable, there are some startling statistics about upsets in the NCAA tournament over the past eight years that may point towards inefficient seeding. On average there are 8.6 second round

upsets. That means 26.9 percent of the time the lower seed wins. Over the second, third, fourth, and fifth rounds in the past eight years, the lower seed won 140 times (29%). Most startling is that in the fifth round alone, the lower seed has won 56.25 percent of the time. While there may not be a rule of thumb to the number of acceptable upsets these figures appear to be high.

Office Pools

The primary purpose of this paper is not to correctly predict all 63 NCAA games of the tournament. But this paper can be a foundation when filling out brackets for an office pool. This paper attempts to determine what statistics are most important to a team's probability of reaching the final four. In most office pools, correctly predicting a final four team is very valuable and the person who correctly predicts the most final four teams will usually win.

Style of Play

Except for 2010, Duke has been performing relatively poorly in the NCAA tournament. In 2012 as a two seed it lost to fifteen-seeded Lehigh (one of only six times a 15 seed has beaten a 2 seed). In 2011 and 2006 as a one seed Duke made the sweet sixteen only to lose to fifth-seed Arizona and fourth-seed LSU respectively. In 2008 as a two seed Duke lost to seventh-seeded West Virginia in only the second round. In 2007 Duke lost in the first round as a six seed to eleventh seeded Virginia Commonwealth.

Why has a school with four national championships, a winning percentage in the tournament of 75%, and 15 final-four appearances been underperforming in the past 8 years?

One possible explanation may be a lack of rebounding prowess. Except for 2010 when Duke won the national championship, it has been 5th, 45th, 12th, 21st, 31st, 26th, and 5th worst in rebounding rate for the tournament years 2006, 2007, 2008, 2009, 2011, 2012, and 2013 respectively. Duke has been in the bottom half of the tournament teams in rebounding in six of the past eight years.

Duke has also struggled in the past 8 year in defensive efficiency. It has ranked 16th, 15th, 16th, 29th, 28th, 12th, 42nd, and 23rd in allowing the most points per 100 possessions among tournament teams from 2006

through 2013. This is far below the level needed for tournament success.

Focusing on a high scoring, volume shooting, non-defensive, non-rebounding style of basketball has not seemed to benefit Duke in recent years and may be one of the reasons for its recent lack of success. This paper provides statistical backing to this claim, and may provide insight for college coaches on which style of basketball is best suited for tournament success.

IV. Current Literature

Despite the current popularity of predicting the outcome of sports and the effort to create an ultimate predictive statistic, there has been very little academic research published on the topic. The research that does exist seems to focus on professional sports, primarily baseball. There are very few published articles on predicting the NCAA tournament. This may be because people who have found credible results may not want to publish their results and instead use them for their own personal good. Also, there may not be any significant results produced due to a lack of available data. There are primarily two approaches to predict the NCAA tournaments. One is through a capture-all statistic such as Ratings Percentage Index, Basketball Power Index, KenPom ratings, and Jeff Sagarin's strength rating. The second is through limited published research.

Commonly Used Power Rankings

ESPN created the Basketball Power Index (BPI) in 2012 in an attempt to accurately predict what teams would receive at-large bids by the selection committee. ESPN claims the BPI adjusts a team's score based on pace of the game, unlike other power rankings. One of the most unique aspects of the BPI is that it accounts for performance due to missing players (injuries, suspensions, etc.). This allows a team not to be punished when it loses a game if their star player has a sprained ankle and is unable to contribute (Oliver 2012).

The Ratings Percentage Index (RPI) is a tool the NCAA has developed to rate several different sports. The selection committee relies heavily on the RPI because of its simplicity. The RPI uses only three components: winning percentage, opponents' winning percentage, and your opponents'-opponents' winning percentage. Winning percentage is

weighted by a factor of .25, opponents' winning percentage by a factor of .5, and your opponents'-opponents' winning percentage by a factor of .25. Winning percentage is weighted for home wins and losses, away wins and losses, and neutral site wins and losses differently. A home win is equal to .6 wins; an away win is equal to 1.4 wins. Conversely, a home loss is equal to 1.4 losses, and a road loss is equal to .6 losses. All neutral site games are weighted as 1 win or 1 loss (NCAA 2014).

Jeff Sagarin created his own power rating index in the 1980's and his metrics have been published in USA Today since 1985. Most of his formula is shrouded in secrecy. Sagarin claims to put weight on margin of victory, an adjustment for blowouts, location of games, and strength of schedule. The exact weights and formulas are unknown (Sagarin 2014).

Like Sagarin, Ken Pomeroy has created his own power rating to predict tournament games. His statistics go back to 2003 and are primarily focused on adjusting outcomes and statistics for the pace of a game. Like Sagarin, his formula has not been released and the exact weights to these statistics are unknown (Pomeroy 2014).

Includes	RPI	BPI	Sagarin	Kenpom
Scoring margin	No	Yes	Yes	Yes
Diminishing returns for blowouts	No	Yes	Yes	No
Pace of game matters	No	Yes	No	Yes
Home/Neutral/Road	Yes	Yes	Yes	Yes
SOS beyond Opponent's opponents' W-L	No	Yes	Yes	Yes
All wins are better than losses (before Opp Adj)	Yes	Yes	No	No
De-weighting games with missing key players	No	Yes	No	No

Published Articles

Shi, Moorthy, and Zimmerman examined the predictive capabilities of current NCAA basketball ranking methods. They predominantly focused on the fact that there is a "glass ceiling" of 75% predictive capabilities. They examined the predictive capabilities of the models employed by Ken Pomeroy, Daryl Morey, John Hollinger, and Dean Oliver. They come to the conclusion that the limited predictive capabilities of the current models is in the choice of variables, not in the models themselves. In theory, selecting the correct variables, may lead to a proverbial busting of the glass ceiling (Shi, Moorthy and Zimmerman 2013).

Most predictive research on basketball has been focused on the NBA. This is because of longer and more consistent schedules. The NBA typically plays a schedule of 82 games with a schedule set by the league. Lee and Berri have recently approached this topic by using production functions to measure positional productivity in the NBA (i.e. Are guards, centers, or forwards more valuable?). They build their model on the premise that "wins in the NBA are determined by how efficiently one scores per possession employed, relative to one's opponent's ability to use possessions efficiently." Lee and Berri calculate the effectiveness of each position on each team. They then use a Cobb-Douglas production function to estimate a log-log econometric model that breaks down the positional quality of each team and how it contributes to wins. Ultimately they discovered that "big men have a greater impact on team wins than small forwards or guards" (Lee and Berri 2008).

On average the higher seed wins 71 percent of the time. This led Carlin (1996) to use seed difference, Sagarin Rating difference, and betting point spreads to predict which teams would reach the final four. His model used two linear regressions. For both models the betting point spread was used as the dependent variable. For their first regression the lone independent variable was the seed difference squared. In the second regression the independent variable was the difference in Sagarin ratings. They applied both of these models to the 1994 NCAA basketball tournament and correctly predicted one of the four regional champions. They state that the model they developed "requires only elementary ideas in probability theory, statistical graphics, and linear regression analysis, and as such should provide an interesting and instructive exercise for students" (Carlin 1996).

V. Model

Except for Carlin, the current literature does not address the issue of predicting the regional champions of the NCAA Division I basketball tournament. Carlin's model is very rudimentary and outdated but can provide a useful foundation for predicting the final four. The model presented in this paper tries to expand on the use of production functions in the NBA as laid out by Lee and Berri and apply the same principles to a binomial logit model. The focus on defensive and offensive efficiencies in this paper is mostly derived from Shi, Moorthy, and Zimmerman.

A binomial logit model can be used to predict success in the NCAA basketball tournament. A logit model result is constrained to a number between 0 and 1. This can be a useful tool in predicting success, as long as success is properly defined. For this model success is defined as winning your respective region and thus making the Final Four. In this logit model the dependent variable is either a 1 for a team reaching the final four or a 0 for teams that do not. Explanatory variables are: points per 100 possessions (PtsPer100Poss) to account for offensive prowess, points per100 possessions allowed (PtsPer100PossAllowed) for defensive effectiveness, rebounding rate (RBSRate) as a measure of ball control, strength of schedule (SOS) in order to normalize statistics based on level of competition, and regional strength (REGSTR) to account for variations in the strength of regions.

$$y = \frac{1}{1 + e^{-[\beta_0 + \beta_1 SOS + \beta_2 PTSPE + \beta_3 DPTSPA + \beta_4 RBSRate + \beta_5 REGSTR]}}$$

VI. Data and Variables

Data

Data were collected on the 510 tournament teams from 2006 to 2013 from BasketballReference.com and randomly cross-referenced against NCAA.org to ensure accuracy. All regressions were run using Gnu Regression, Econometrics, and Time-Series Library (GRETL).

Dependent Variable

The nature of a binomial logit model is that the dependent variable is

binary, taking on a value of either 1 or 0. In this model a 1 will indicate the team made the final four while a 0 will indicate a team did not make the final four.

Strength of Schedule (SOS)

There are many calculations for strength of schedule. For example, NCAA.org typically calculates it based on your opponent's winning percentage and your opponent's-opponents winning percentage. This paper chose to use basketballreference.com's strength of schedule. That calculation is based on an average offensive and defensive NCAA Division I team. If a team typically plays statistically above average teams, then its strength of schedule will be higher. Strength of schedule of 0 would mean over the course of a season the teams you played were a statistically average NCAA Division I team.

The purpose of using strength of schedule statistics is three-fold. Strength of schedule normalizes a team's statistics based on competition level, indicates that a team plays in a more competitive conference, and may indicate that a team has been "battle tested".

There is a difference between scoring 70 points per game against a really good defensive team and scoring 70 points per game against an extremely weak defensive team. In order to normalize a team's statistics, the strength of schedule variable is used.

There are 32 NCAA Division I basketball conferences. The teams that play in the major conferences (See Table 1) typically make the final four more frequently. Teams in one of these six power conferences will innately have a higher strength of schedule and therefore be more likely to reach the final four.

When a team reaches the tournament and has played a very weak schedule it is more susceptible to upsets. It has not been challenged at the highest level yet. Playing a weak schedule is not good preparation for entering a tournament consisting of the 64 best teams in the country.

The lowest strength of schedule for the 512 tournament teams was 2013's Southern University at -10.31. The highest strength of schedule for the 512 tournament teams was 2011's Michigan State at 11.67. The average strength of schedule for the 512 tournament teams is 4.1353.

Points per Possessions Earned (PTSPE)

$$PTSPE = \frac{PointsPerGame}{FGAtt - OffRbs + TO + (.445)(FTAtt)}$$

In order to account for a team's offensive efficiency this model uses the statistic points per possessions earned. The equation used is directly from Lee and Berri's paper on the NBA. Unlike points per game this statistic adjusts for the pace of a game. A team can score a large number of points simply by increasing its number of possessions. By adjusting for the pace of the game this statistic can indicate how effective a team will be when their possessions in a game are limited. As points per 100 possessions increases, so should a team's chance of reaching the final four. The maximum, minimum, and average points per possessions earned for the 512 tournament teams was 1.22 for Missouri in 2012, .93 for Arizona in 2006, and 1.09, respectively.

Points Allowed Per Possessions Allowed (DPTSPA)

$$DPTSPA = \frac{PointsPerGameAllowed}{OppTO + DefRbs + TotalRbs + OppFGMade + (.445)(OppFTMade)}$$

Like points per possessions earned, this equation comes directly from Lee and Berri's work on the NBA. This statistic indicates how effective a team is at preventing their opponent from scoring while holding possessions constant. Even if a team slows a game down in order to prevent high scoring, it does not necessarily mean it is efficient defensively. As a team's points allowed per possession allowed decreases, its chance of reaching the final four will increase. In 2008 Oregon had the worst points allowed per possession allowed of the 512 tournament teams at 1.08. Stephen F. Austin in 2009 had the best points allowed per possession allowed of the 512 tournament teams at .85. The average for the 512 tournament teams from 2006-2013 was .97.

Rebounds Rate (RBSRate)

$$RBSRate = 100(ReboundsPerGame)/((FGAtt - FGMade) + (FTAtt - FTMade) + (OppFGAtt - OppFGMade) + (OppFTAtt - OppFTMade))$$

Rebounds Rate is a statistic that determines what percentage of available rebounds in a game a team gets. Instead of simply using rebounds per game, this statistic adjusts for how many available rebounds there are. Having 40 rebounds per game when there are 100 available rebounds is not as good as having 40 rebounds per game when there are only 50 available rebounds. As rebounding rate increases, a team's chances to reach the final four should increase as well. The minimum rebounding rate for the 512 tournament is 42.54% by West Virginia in 2006. The maximum rebounding rate for the 512 tournament teams is 58.65% by Old Dominion in 2011. The overall average rebounding rate for the 512 tournament teams is 51.99%.

Regional Strength (REGSTR)

$$\sum_{Reg}^{\hat{Y}} = \frac{1}{1 - e^{[\beta_0 + \beta_1 SOS + \beta_2 PtsPer100Poss + \beta_3 PtsPer100PossAllowed + \beta_4 RBSRate]}}$$

Not all regions are created equally. Despite the selection committee's attempt to make all regions equally competitive, it is rarely achieved. This statistic attempts to account for any discrepancies in the overall strength of regions. It also will account for the quality of opponents a team must face in its region to reach the final four. This statistic is created by first running a binary choice logit model excluding regional strength. This results in a predicted probability ($\hat{Y}$) for each of the 512 tournament teams to reach the final four. Then for each individual team I sum the $\hat{Y}$'s for the other 15 teams in its region. For example: In order to calculate the regional strength for 2012 Kentucky (located in the South region) you would sum the $\hat{Y}$'s for each of the other 15 teams in the South region. The highest regional strength was for North Carolina A&T in the Midwest region in 2013 with a regional strength of 1.63. This means that the sum of the $\hat{Y}$'s of the other 15 teams in the Midwest region in 2013 was over 163 percent. The lowest regional strength was for Villanova in 2006 with a regional strength of .304. This would be that the sum of the $\hat{Y}$'s of the other 15 teams in the 2006 Midwest region was just over 30 percent. The average regional strength is .9385.

VII. Econometric Results

Variable	Coefficient	P-Value
Constant	-3.39291	
SOS	.453511	.0033
DPTSPA	-20.2455	.0004
PTSPE	9.64882	.0604
RBSRATE	14.957	.2417
Region Strength	-1.9480	.0303

The primary econometric results to note are the sign of the coefficients, and the p-values of the coefficients.

As expected, the coefficient on strength of schedule is positive. It is also statistically significant at the one percent level. This would indicate that a team from a more prominent conference and is battle tested would have an increased chance to reach the final four. Points allowed per possession allowed has a negative coefficient as predicted. It is also statistically significant at the one percent level. Next, points per possessions earned was hypothesized to have a positive coefficient and this turned out to be true. The coefficient is significant at the five percent level. Rebounding rate was also hypothesized to have positive effect. Due to a large p-value one cannot conclusively determine if rebounding rate is statistically different from zero. In other words, rebounding rate may not influence a team's chance of reaching the final four. Finally, regional strength would be expected to be negative. This is because the tougher a region is and the more difficult the opponents a team faces on its way to the final four, the less likely it is to reach the final four. This hypothesis is confirmed at the five percent level.

Common econometric measures of fit may not be relevant to this model. Adjusted R2 for this model is calculated as .2534. This means that the model is explaining 25 percent of the variation in the dependent variable. Later in this paper it will be apparent that this is a gross underestimation. Also, GRETL produces a measure of fit by determining

the number of cases "correctly predicted". GRETL uses the decision rule that if an observation is greater than .5 that observation will be predicted as a 1. Conversely, GRETL predicts that any observation less than .5 will be a 0. This is not always the best decision rule for this model because a team could have a predicted probability under .5 and still be the best team in its region. For example: In 2013 Louisville would have a predicted probability of reaching the Final Four of 39.2 percent. While GRETL would predict Louisville to not reach the Final Four, in the Midwest region Louisville had the highest predicted probability and therefore would be correctly predicted to reach the Final Four.

One other item to note about the coefficients of the variables is the distinct magnitude difference between points per possessions earned and points allowed per possession allowed. The average of points per possession earned is 1.09 and the average for points allowed per possession allowed is .97. Despite the closeness of these averages, the coefficient of points allowed per possession allowed is nearly double that of points per possession earned. This would indicate that perhaps defense does truly win championships.

VIII. Application to 2007-2012 Tournaments

In order to determine the quality of this model it is best to compare its predictive capabilities to that of other current power ratings. Most pre-tournament data is unavailable for current power ratings. Therefore the comparative analysis is limited to the tournament years 2007-2012. Peter Tiernan of Bracket Science and John Ezekowitz of Harvard Sports Analysis Collective have calculated the predictive capabilities of the previously referred to RPI, BPI (limited information), and KenPom Ratings. (There is no data on the predictive capabilities of Jeff Sagarin's model) They also analyze the predictive capabilities of the "true-seed" method. This method assumes that the higher seed will always win.

A good measure for overall bracket prediction is that of games correctly predicted. This is a measure of how many of the 63 tournament games a model would correctly predict if one were to fill out their bracket pre-tournament. The Fuqua Statistic correctly predicts 42.7 (67.2 percent) games on average. This is first among all other power ratings but only by an average of approximately 1 game per year.

A good predictive indicator of a power rating is how a bracket filled out prior to the tournament would do on ESPN's bracket challenge.

ESPN's bracket challenge awards 1 point for correctly predicting a first round game, 2 points for each sweet sixteen team correctly predicted, 4 points for each elite eight team correctly predicted, 8 points for correctly predicting a final four team, 16 points for correctly predicting a championship game team, and 32 points for correctly predicting the national champion. Using these numbers the Fuqua statistic was third among current power ratings. The best rating system according to this method would be the true seed method. On average the true seed method would score 13 points better per year on ESPN than the Fuqua statistic. The Fuqua statistic struggles to predict the ultimate national champion and therefore takes a hit in this scoring system. KenPom's rating and the true seed method predict three and five national champions respectively over a six year period. The Fuqua Statistic only predicts two national champions. Up to just the final four however the Fuqua statistic would score the most points on ESPN. It is only the games after the final four that it seems other power ratings have an advantage.

One final indicator of predictive success would be its ability to predict matchups. This would mean predicting the first 32 games of the NCAA tournament and then once the next round matchups are determined, predict the next 16 games. For example: In 2012 Duke played Lehigh and Notre Dame played Xavier. The winner of these two games would face each other. This model would have predicted Xavier to beat Notre Dame (Xavier won) and Duke to beat Lehigh (Lehigh won). Then it would have predicted Duke to beat Xavier. By filling out the bracket prior to the tournament this model would have correctly predicted only one game of three correct. If after the first round, however, I examined the Lehigh vs. Xavier matchup my model would have correctly predicted Xavier to advance. Therefore, at predicting matchups this model would have predicted two of the three matchups correctly. The premise of this measure is flawed because one cannot change one's predictions once the tournament begins. However, it seems that this is a popular measure of predictive capabilities. In this category the Fuqua statistic is second to the BPI by .6 percent.

While these overall tournament measures are useful, the purpose this paper is to provide a method of predicting the final four. In this category, over a six year period (24 possible final four teams), KenPom, RPI, and the true seed method would have predicted ten, nine, and ten final four teams respectively. The Fuqua statistc would have correctly predicted 14 final four teams over this same period. This is forty percent better than

any of the current models. It appears that this model is superior to all other current ratings in the category in which it is designed to be best.

	FuquaStatistic	ESPN BPI	NCAA RPI	True Seed	KenPom
Games Correctly Predicted	**256**		238	252	251
Points on ESPN Bracket Challenge	**479**		448[1]	472	474[1]
Points on ESPN Bracket Challenge[2]	623		496	696	625
% of games Correctly Predicted	**67.72**		63.00	66.70	66.40
% of Matchups Correctly Predicted	73.8	74.4		73.0	73.0
Final Four Teams Correctly Predicted	**14**		9	10	10

1 - estimated number based on available data
2 - Calculated up to the final four and not beyond

IX. Issues and Further Research

Shi, Moorthy, Zimmerman claim that there is a "glass ceiling" when it comes to predicting NCAA basketball games. They claim this "glass ceiling" is around 74-75 percent. This model seems to reach this percentage but is unable to break through it. Shi, Moorthy, and Zimmerman attribute this to "the attributes (variables) we and others use."

There are four primary reasons for this ceiling. First, much of the relevant data in a basketball game is unavailable. For example, the number of passes per possession may be relevant to basketball success but no such statistic exists and would be extremely difficult to measure. Second, there are several immeasurable aspects to a basketball game. Specifically, college basketball games involve 18 to 22 year old kids with

large variations in emotion. There is no way to measure the effect on a key player's psyche if his girlfriend breaks up with him the night before. Next, the structure of NCAA Division I basketball creates a challenging setting for predictive analysis. There are 351 teams, and each team has a large range of resources. Each team is also able to select most of its opponents, (only a certain number of conference games are required) and each team only plays roughly 30 games (a small sample size). These factors don't allow for a good statistical sample for modeling. Finally, there is a large level of randomness involved in basketball. There are ten players on a 94 foot by 50 foot court and one wet spot on the floor, one bad call by a referee, or one underinflated basketball, may have a large impact on the game.

Further research is possible but highly tedious. There are two main statistics that could be calculated and may significantly affect a team's probability of reaching the final four. First, a statistic for consistency may be useful. The best way of calculating this would be by taking the standard deviation of statistics. The larger a standard deviation a team has, the more inconsistent a team is, and the less likely they are to reach the final four. Secondly, it may be useful to calculate a team's tournament and game experience. This may be calculated by using the number of tournament minutes played by a team's roster.

There is bound to be a breakthrough in the research of predicting sporting events. As our technology increases we are able to calculate and record more statistics and therefore have more precise models.

X. Conclusion

Is this just another mediocre attempt at predicting the NCAA tournament or have any earth shattering discoveries been made? The simple answer is no. The independent variables found to be significant are commonly agreed on by all basketball pundits. The proverbial glass ceiling determined by Shi, Moorthy and Zimmerman of 73-75 percent prediction accuracy was not broken through. It appears that despite being better than the current power ratings in almost all predictive measures it is only by a slight margin. The only true indicator of this being a superior model is that over the span of six years this model correctly predicted four more Final Four teams than other current power ratings. Overall it may be just another run of the mill model.

References

BasketballReference.com. Accessed April2, 2014. http://www.sports-reference.com/cbb/.

Carlin, Bradley. 1996. "Improved NCAA Basketball Tournament Modeling via Point Spread and Team Strength." *The American Statistician*, 50: 39-43.

Lee, Young, and David Berri. 2008. "A Re-Examination of Production Functions and Efficiency Estimates for the National Basketball Association." *Scottish Journal of Political Economy*, 55: 51-66.

NCAA.com. Accessed April 2, 2014. http://www.ncaa.com/rankings/basketball-men/d1/ncaa-mens-basketball-rpi.

NCAA.com. 2014. "2013-2014 NCAA Division I Men's Basketball Championship Principles and Procedures for Establishing the Bracket." Accessed April 2, 2014. http://www.ncaa.com/content/di-principles-and-procedures-selection.

Oliver, Dean. 2012. "Introducing the BPI." Accessed April 2, 2014. http://espn.go.com/mens-college-basketball/story/_/id/7561413/bpi-college-basketball-power-index-explained.

Pomeroy, Ken. *kenpom.com.* Accessed April 2, 2014. http://kenpom.com/.

Sagarin, Jeff. January 06, 2014. *USAtoday.com.* Accessed April 2, 2014. http://www.usatoday.com/sports/ncaaf/sagarin/.

Sanburn, Josh. 2013. "March Madness Will Cost Employers $134 Million." Accessed April 02, 2014. http://business.time.com/2013/03/19/march-madness-will-cost-businesses-134-million-why-arent-employers-concerned/

Shi, Zifan, Sruthi Moorthy, and Albrecht Zimmerman. Accessed April 2, 2014. "Predicting NCAAB match outcomes using ML techniques-some results and lessons learned." http://dtai.cs.kuleuven.be/events/MLSA13/papers/mlsa13_submission_12.pdf.

Woodruff, Mandi. 2012. "You've Got A 1 in 35 Billion Chance At Filling Out The Perfect March Madness Bracket". Accessed April 2, 2014. http://www.businessinsider.com.au/youve-got-a-1-in-35-billion-chance-at-filling-out-the-perfect-march-madness-bracket-2012-3.