Mining College Basketball Data

Sam List, Andrew Cox, Silas Adheke
Faculty Mentors: Joseph Nolan & David Agard
Department of Mathematics and Statistics, Northern Kentucky University

Background

An unwritten rule recently discussed on Kentucky Sports Radio suggests the first team in a college basketball game to score 71 will win the game 95% of the time. College basketball has a big influence on Universities and is popular with fans across the country. This project examines this and related potential ideas to see if there is any score in the game more significant than the 71 point mark. It also develops a model in an attempt to accurately predict the win probability from the home team perspective at any point during the course of the game.

Data / Methods

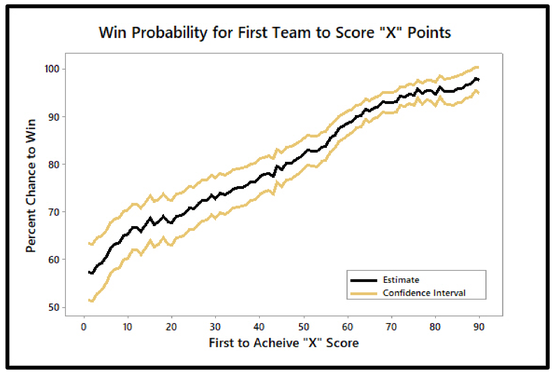
This study focused on all eligible NCAA division-I basketball teams. Data were collected for a random sample of 39 teams. All regular season and conference tournament games were included for each team (N = 845 games). Play-by-play for each game was scraped from ESPN's website using SAS. Vegas lines were obtained and merged into the data for use in predicting winner. SAS code was implemented to identify the first to each score within a game.

Final data sets were constructed from the home team perspective and included the following variables:

- Elapsed Time
- Home and Visitor Score Differential
- Vegas Line
- First to Reach Scoring Thresholds
- Winner

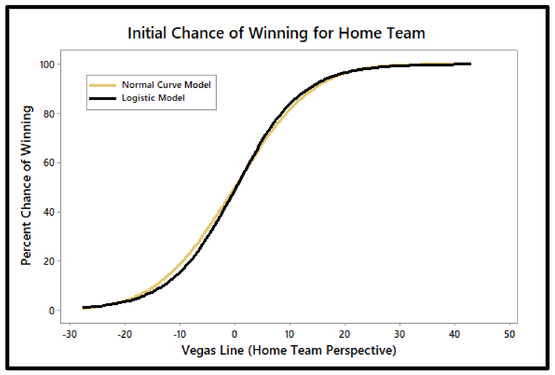
Using the final model we generated a scatter plot using the first to reach a score and who eventually won the game. We then used a normal distribution to compare the Vegas line spreads to the chance of winning using the game data we collected. Next we constructed a contour plot that examined the difference in score given an elapsed time to determine a probability of winning the game. Finally we used the game data from a NKU basketball game and the data provided on ESPN to compare our model which used a combination of the Vegas line, point differential, and the elapsed time to the ESPN win probability model.

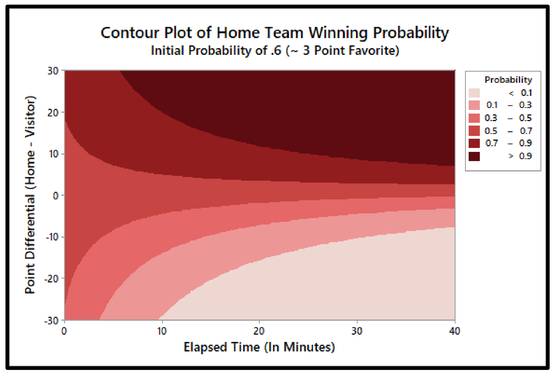
Results



The figure to the left displays confidence interval bands for the probability that a team to reach a certain score during the course of the game will go on to win that game. For example the team to reach 71 points first is estimated to have a 90.9 to 95.2 percent chance to win the game. Note that this interval includes as plausible the statistic given on Kentucky Sports Radio.

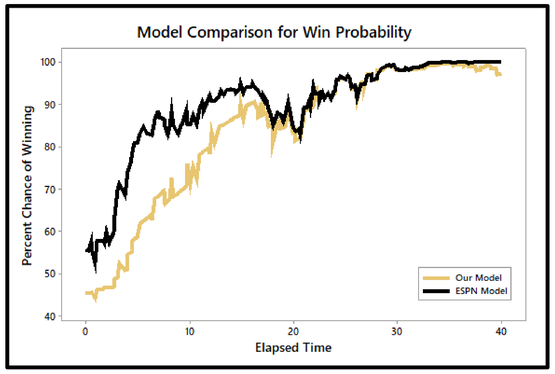
The figure to the right displays two approaches for estimating the initial pregame probability of the home team winning, based on the Las Vegas point spread. The yellow curve uses a normal distribution model for the error between the point spread and the actual difference at the end of the game. The second approach (black curve) uses a logistic regression model to predict the home team's initial chance of winning.





Another win probability model uses a logistic regression with several independent variables: elapsed time in the game, the current differential in score between teams (and their interaction), and the aforementioned initial probabilities for the home team winning the game. The figure to the left presents a contour plot using a fixed initial win probability of .6 (representing a 3 point favorite).

The model to the right is an attempt to recreate the win probability model displayed on ESPN. This example shows the Northern Kentucky University victory over Wright State University in the 2019 Horizon League Championship game. The graph is presented from NKU's perspective. At the start of the game, NKU got off to a hot start and had a 14-point lead 10 minutes into the game. This gets reflected by the rapid increase in the win probability. After 15 minutes our model is indistinguishable from the ESPN model. The primary distinction between the models is in the initial probability and that is because Vegas had Wright State as a 2-point favorite going into the game.



Conclusions

The rule of 71 is not a myth; the first team to reach 71 points has at least 90.9 percent chance of being the eventual winner and perhaps as high as a 95.2 percent chance. It remains plausible that the 95 percent discussed on Kentucky Sports Radio is reasonable.

Probability of winning is associated to initial probability, elapsed time, and point differential. Our model from halftime correctly picks the eventual winner 80 percent of the time.

Our Model is at least partially consistent with the model presented by ESPN data analysts.

Future Work

Future study might include:

- A more detailed comparison to ESPN's model.
- Incorporate additional variables to the initial probability model.
- Create interactive graphics comparable to ESPN

References

1. Berkman, J. (n.d.). The Complete List of NCAA Division 1 Colleges (Most Recent). Retrieved June 13, 2019, from https://blog.prepscholar.com/complete-list-of-division-1-colleges-by-state
2. Carlin, B. P. (1996), "NCAA Basketball Tournament Modeling via Point Spread and Team Strength Information," *The American Statistician, 50, 39-43.*
3. Gormley, S. (2017, April 02). The Rule of 71... Retrieved June 13, 2019, from http://kentuckysportsradio.com/basketball-2/the-rule-of-71/
4. NCAA BASKETBALL SCORES AND ODDS ARCHIVES. (n.d.). Retrieved June 13, 2019, from https://www.sportsbookreviewsonline.com/scoresoddsarchives/ncaabasketball/ncaabasketballoddsarchives.htm
5. NCAAM College Basketball Scores - NCAAM Scoreboard. (n.d.). Retrieved June 13, 2019, from https://www.espn.com/mens-college-basketball/scoreboard

Acknowledgements: This research was a product of a 2019 UR-STEM summer undergraduate research experience supported by Northern Kentucky University's Center for Integrative Natural Science and Mathematics.
Special thanks to the Burkardt Consulting Center for use of office space and computers. We would also like to thank Jacob Englert for his work on the SAS coding to obtain information for play-by-play.