THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

# Does how far you've come determine how far you'll go? An analysis of travel distance within the NCAA Tournament

Daniel Horlick, Calleigh Miller, Katie Webb, and David Bruss

## INTRODUCTION

- March Madness is notoriously unpredictable, given the chance of a perfect bracket is nearly one in 9.2 quintillion (Chris Katje, 2023)
- Should a tournament team consider the distance between their campus and the neutral game site as a possible factor for winning?
- Can a predictive model accurately determine whether a team will win or lose a tournament game based off the distance traveled?
  - The top seeded teams do not have to travel very far. What is the outcome when we control for these large seed differences?

## LITERATURE REVIEW

- Previous studies on the NCAA Tournament have analyzed factors that influence wins that include experience of the team, strength of the team, and coaching (Pfifer et al., 2019). The studies found that strength of a team was more useful in predictions than experience
- Another study developed at the University of Southern Indiana, compiled data from point differentials, location of the game, and date of the game, to create a "PageRank" calculation that attempted to eliminate biases in predictions (Matthews et al., 2021). Their model was found to be useful in seeding process
  - Both studies mentioned failed to focus on travel distance

## METHODOLOGY

- Data was collected from both the 2019 and the 2022 seasons
- Distance traveled determined by physical distance between university and game location
- Power 6 includes the following conferences: ACC, Big East, Big Ten, Big 12, Pac-12, SEC
- A pre-covid and post-covid variable was included to measure if results varied with the pandemic
- Automatic bid includes teams who earned a spot by winning a conference tournament
- Games with a seed difference greater than 7 were omitted as the team with the better seed are highly likely to win regardless
- Binary logistic regression model was conducted using "Wins" as the dependent variable

## DESCRIPTIVE STATISTICS

Table 1.
*Descriptive Statistics for Independent Variables*

| Conference | *n* | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| Power6 | 154 | 0 | 1 | .69 | .462 |
| AutoBid | 154 | 0 | 1 | .36 | .481 |
| Seed | 154 | 1 | 16 | 6.89 | 4.12 |
| DiffInSeed | 154 | -7 | 7 | .00 | 3.83 |
| Round | 154 | 2 | 68 | 42.13 | 24.86 |
| PostCovid | 154 | 0 | 1 | .49 | .50 |
| DistComparedtoOpp | 154 | -2804 | 2804 | 1.67 | 883.72 |
| DistSchooltoGame | 154 | 0 | 3022 | 978.59 | 693.19 |

## CORRELATIONS

- The table shows Pearson correlation coefficients, highlighting significant correlations with an asterisk
- No variables were too highly correlated
- No evidence of multicollinearity in our final model
- Note: the "DistComparedtoOpp" variable is the difference between the school's and the opponent's "DistSchooltoGame" value

Table 2.
*Independent Variables Correlations*

| | POW | AB | SEED | DIFS | ROU | COV | DCO |
|---|---|---|---|---|---|---|---|
| Power6 | | | | | | | |
| AutoBid | -.389* | | | | | | |
| Seed | -.433* | .129 | | | | | |
| DiffInSeed | -.152 | .064 | .464* | | | | |
| Round | -.341* | -.028 | .690* | .000 | | | |
| PostCovid | -.023 | -.058 | .074 | .000 | .018 | | |
| DistanceComparedtoOpp | .008 | -.115 | .011 | .021 | .002 | .002 | |
| DistSchooltoGame | .048 | -.117 | -.027 | .022 | .054 | .057 | .633* |

* $p < .01$

## LIMITATIONS/FUTURE RESEARCH

- Limited time = Limited data
  - Only 2022 and 2019 March Madness tournaments are included
- The maximum difference in seed of 7 was simply picked based on upset percentages in the first round
- Teams must only travel for rounds of 68, 64, 16, and 4
  - Future research could get a larger data set of games in those rounds to see if travel affects the following game
- Possible research topic: investigate how Last Four In teams play when they have three games and two flights in less than five days
- Possible research topic: isolating second game in tournament and evaluate performance after one day of no travelling

## IMPLICATIONS

- Distance is not a significant factor for winning in March Madness, even when removing large seed differences from the data set
  - Bracket committee should not sacrifice in order to put higher seeds closer to their school
  - Ex. Putting highest 2 seed with highest 1 seed
- Teams could be accustomed to away games by the end of the season
  - Anywhere from 10-18 away games are played in a season by each team
  - Neutral March Madness sites allow for each team to have plenty of fans attend
- Despite this, the 59.1% of games were correctly predicted in the classification table
  - The unpredictability of March Madness makes that extra 9.1% valuable

## RESULTS

- Our final model from a binary logistic regression resulted in a predictive accuracy of 59.1%
- The model was almost equally as accurate when producing either outcome (wins and losses)

Table 3.
*Binary Regression Classification Table*

| | | **Predicted** Win | | |
|---|---|---|---|---|
| **Observed** | | 0 | 1 | Percentage Correct |
| Win | 0 | 44 | 33 | 57.1 |
| | 1 | 30 | 47 | 61.0 |
| **Overall Percentage** | | | | **59.1** |

## FINAL MODEL

- The distance variables, which we hypothesized would impact winning odds, had 0 effect on winning. Each of the distance variables resulted in 1.000 log odds ratios (no effect)
- Power6 resulted in a log odds ratio of 2.254, indicating that if a team is a part of the Power6, they are twice as likely to produce a win
- Similarly, AutoBid had a log odds ratio of 1.370, indicating that if a team received an automatic bid to the tournament, they are 37% as likely to win

Table 4.
*Binary Regression Final Model on the Dependent Variable Wins*

| Variable | Coefficient | Coefficient significance | Log Odds |
|---|---|---|---|
| (Constant) | -.973 | .170 | .378 |
| Power6 | .813 | .064 | 2.254 |
| AutoBid | .315 | .418 | 1.370 |
| Seed | .024 | .753 | 1.024 |
| DiffInSeed | -.019 | .740 | .981 |
| Round | .003 | .815 | 1.003 |
| DistSchoolxGame | .000 | .964 | 1.000 |
| PostCovid | .015 | .964 | 1.015 |
| DistanceComparedtoOpp | .000 | .428 | 1.000 |
| DistSchooltoGame | .000 | .964 | 1.000 |
| | | Model Significance | .766 |
| | | Cox & Snell $R^2$ | .031 |
| | | Naegelkerke $R^2$ | .042 |

[a]Significant at the .01 level
[b]Significant at the .05 level