Wharton SPORTS ANALYTICS Student Research Journal

2023, Spring Edition

wsb.wharton.upenn.edu/student-research-journal

Predicting March Madness Cinderella Teams

Allen Sha, *W'23*
The Wharton School of the University of Pennsylvania, Philadelphia PA, USA

Noah Hyman, *W'23*
The Wharton School of the University of Pennsylvania, Philadelphia PA, USA

Joshua Sewell, *W'23*
The Wharton School of the University of Pennsylvania, Philadelphia PA, USA

Aron Ramsey, *W'23*
The Wharton School of the University of Pennsylvania, Philadelphia PA, USA

Advisor: Abraham Wyner, PhD
The Wharton School of the University of Pennsylvania, Philadelphia PA, USA

Abstract

This study investigates the challenge of predicting which low-seeded teams will make it to the Sweet 16 in the NCAA Championship tournament. The authors utilized power scores and simulations to predict Cinderella teams and examined the factors that contribute to their success. The study found that power scores alone were insufficient to identify Cinderella teams, and that favorable matchups and luck also played significant roles. Identifying the factors that contribute to the success of Cinderella teams is important, as predicting these outcomes is notoriously difficult and has been an ongoing challenge for sports analysts. While historical odds may still be effective in predicting outcomes, further research is needed to refine methods for predicting Cinderella teams and improve the accuracy of future predictions. The findings offer valuable insights into the methods and challenges of predicting outcomes in sports, and highlight the need for continued research in this area.

key words: basketball, NCAA tournament, underdogs, point differential, strength of schedule,

Introduction

Every March, thousands of statisticians and sports analysts attempt to predict which college basketball teams will make a run for the NCAA Championship. This is an incredibly difficult task as upsets and statistically improbable outcomes happen each year. In just the last five years, two 16 seed teams, UMBC and FDU, took down 1 seed powerhouses. A common challenge is attempting to predict what low-seeded teams will make a deep run in the tournament. Sports analysts have used a variety of methods to approach this problem, using regressions based on team attributes, KenPom metrics, and many others. These methods have been largely inconclusive. We sought to predict March Madness Cinderellas using power scores and simulations to find low-seeded teams that have an outsized probability of making the Sweet 16. This method factors in a team's performance during the regular season as well as the difficulty of their matchups in the tournament. Our goal was to determine what causes certain teams to go further than they should given their seed.

Methods

Data

We used 2 datasets from Kaggle:
https://www.kaggle.com/competitions/march-machine-learning-mania-2023/data
The file MRegularSeasonCompactResults contained all men's NCAA Division I regular season games from 1985, when the tournament took its current form, through 2023, which was 181,683 games. The file MNCAATourneySeeds contained all the teams and their seeds in every tournament since 1985. For our analyses, we only looked at the past 5 years when the tournament was played (2018, 2019, 2021, 2022, 2023).

Methodology

We first sought to predict the Cinderellas of March Madness. We defined Cinderellas as teams seeded 8 or lower making to the Sweet 16. Since 2018, there have been 18 such teams. To do this, our group attempted to fit a power score for every team in a given season. Then, given the year's bracket, match up the respective teams and run a simulation using their power scores and drawing a random normal variable.

Power Score Model

The first step was to clean and rearrange the data into a new matrix fit for a regression. Every row in our data frame consisted of a game between two NCAA division I teams. The variables of interest to us were Team 1 and Team 2 (Team number is randomized), Location (Home, Away, or Neutral), and Point Differential (Team 1 points - Team 2 points). The set up was as follows:

- For every team T_n in game g, let $X_{Tn} = 1$ if they are team 1
- For every team T_n in game g, let $X_{Tn} = -1$ if they are team 2
- Let $L_H = 1$ if team 1 is Home, $L_N = 1$ if location is N. Let L_H & $L_N = 0$, if team 1 is Away
- Let Y = point differential between team 1 T_n and team 2 T_n in game g

Next, we fit a linear regression on Y point differential using those team 1, team 2, and location:

$$Y = \beta_1 X_{T1} + \beta_2 X_{T2} + \ldots + \beta_n X_{Tn} + \beta_{n+1} L_H + \beta_{n+2} L_N + \varepsilon$$

The result of this regression was a list of coefficients or power scores for every team in the season that represented each team's expected points above average and accounted for the strength of their schedule and court location advantages. Note that in 2023, there were 363 teams which means there were 363 team coefficients. For the simulation, we filtered these down to just the power scores of the 64 teams who made the tournament. Additionally, unlike other sports, basketball is unique in that the point differential can sometimes misrepresent how close a game was. Teams can quickly fall behind or ahead if the score differential is large. To account for this in a simple way, we capped the maximum point differential at 30 points, which shrunk the coefficients of the best teams because they were not rewarded for outstandingly dominant performances. In the future, we could explore other ways to account for this such as transforms on the differential or instead use a logistic regression on the outcome of which team won. Figure shows the result of the power score regression for 2023.

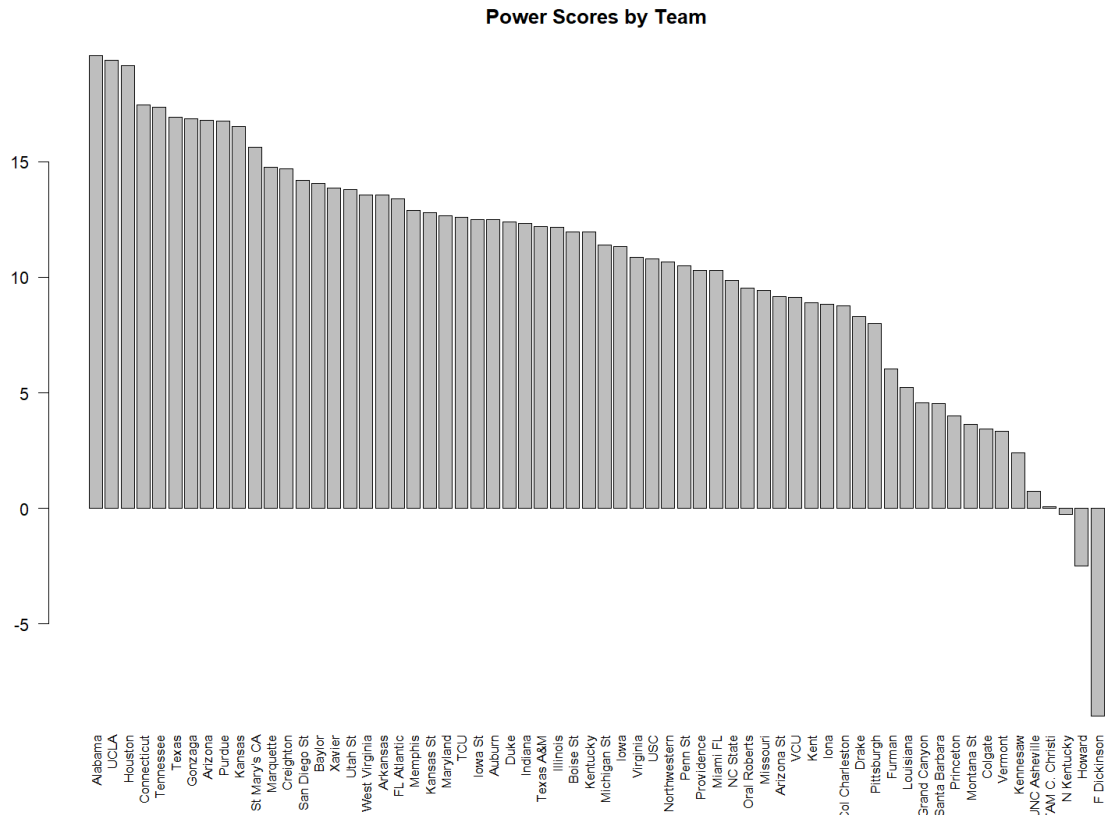

Figure 1: 2023 NCAA Bracket Teams ranked in descending order by power score

Top 10 Teams			Bottom 10 Teams		
Team	Seed	Power Score	Team	Seed	Power Score
Alabama	X01	19.59	Princeton	X15	4.00
UCLA	Z02	19.37	Montana St	W14	3.65
Houston	Y01	19.15	Colgate	Y15	3.43
Connecticut	Z04	17.46	Vermont	W15	3.35
Tennessee	W04	17.37	Kennesaw	Y14	2.40
Texas	Y02	16.91	UNC Asheville	Z15	0.75
Gonzaga	Z03	16.86	TAM C. Christi	X16b	0.08
Arizona	X02	16.79	N Kentucky	Y16	-0.27
Purdue	W01	16.77	Howard	Z16	-2.50
Kansas	Z01	16.51	F Dickinson	W16a	-8.98

Figure 2: Top 10 and bottom 10 teams by power score

Simulations

After obtaining the power scores for every team, we created a vector that represented the bracket. At the start of round 1, there are 64 teams so our vector looks like (1,16,8,9,5,12,4,13…50,63), where each pair of elements represents a matchup in round 1. Note that the numbers were ordered so that the winner of each matchup would be directly next to the winner of the appropriate next pair of seeds in order to reflect the bracket. This step is vital as we are just as interested in determining the importance of the schedule of opponents as we are in the importance of the teams' own power score.

For each matchup, we determined the winner by subtracting team 2's power score from team 1's and then drawing a randomized normal variable with a mean of 0 and the standard deviation of the power score model. If Y_m is greater than 0, then team 1 is given the win.

$$Y_m = \beta_{T1} - \beta_{T2} + \varepsilon \text{ ; where } \varepsilon \text{ is rnorm (n = 1, mean = 0, sd = } \sigma)$$

Without drawing an error term, the equation would always favor the team with a higher power score. After determining the winners from round 1, we then repeated this step with the winners to simulate the teams that would make it to the Sweet 16.

We ran this simulation 10,000 times. A simulation is useful here because we are interested in capturing how the different potential matchups from the first and second round may impact a team's probability of making it to the Sweet 16.

Results

The simulation outputted the predicted probability for every team to make it to the Sweet 16. These probabilities are shown in descending order in the below figure for the 2023 tournament.

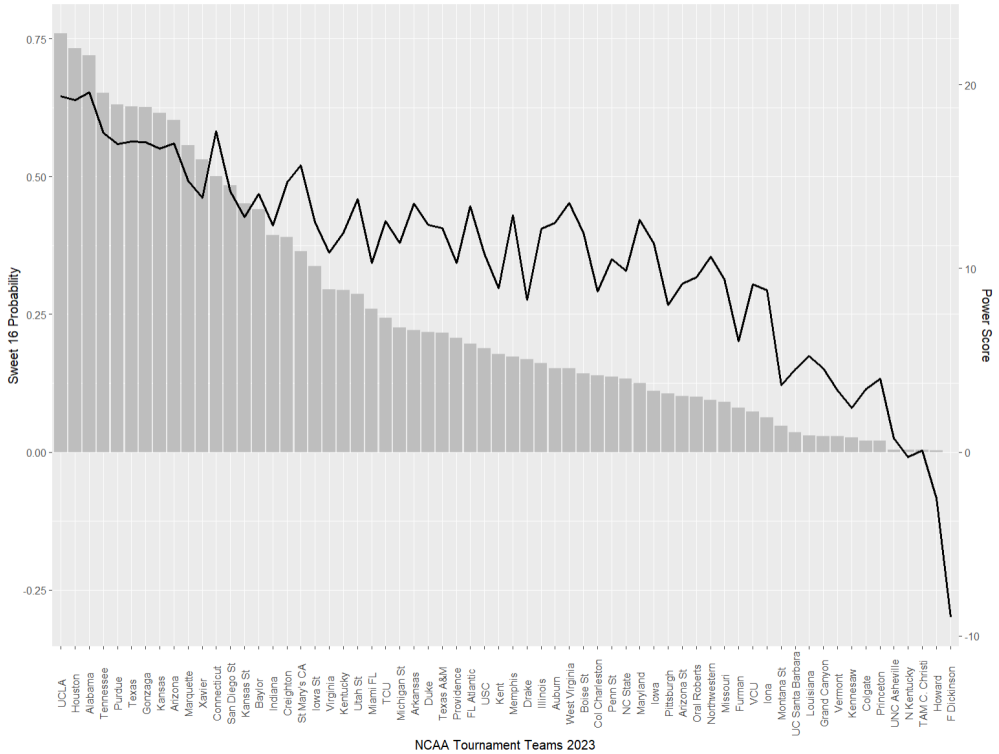

Figure 3: Simulation results for making it to the Sweet 16 Round (bars, left axis) and Power Scores (line, right axis) for 2023 NCAA teams.

While there is a general correlation with probability and team's power score, it is not perfect as seen by the spikes in power score. For example, Alabama had the highest power score of all the teams but placed third in highest probability of making it to the Sweet 16. The reason for this is because UCLA and Houston have more favorable matchups in the first two rounds.

Figure 4: Simulated Probabilities for NCAA 2023 East Division Bracket (Circled teams are the actual winners at each round)

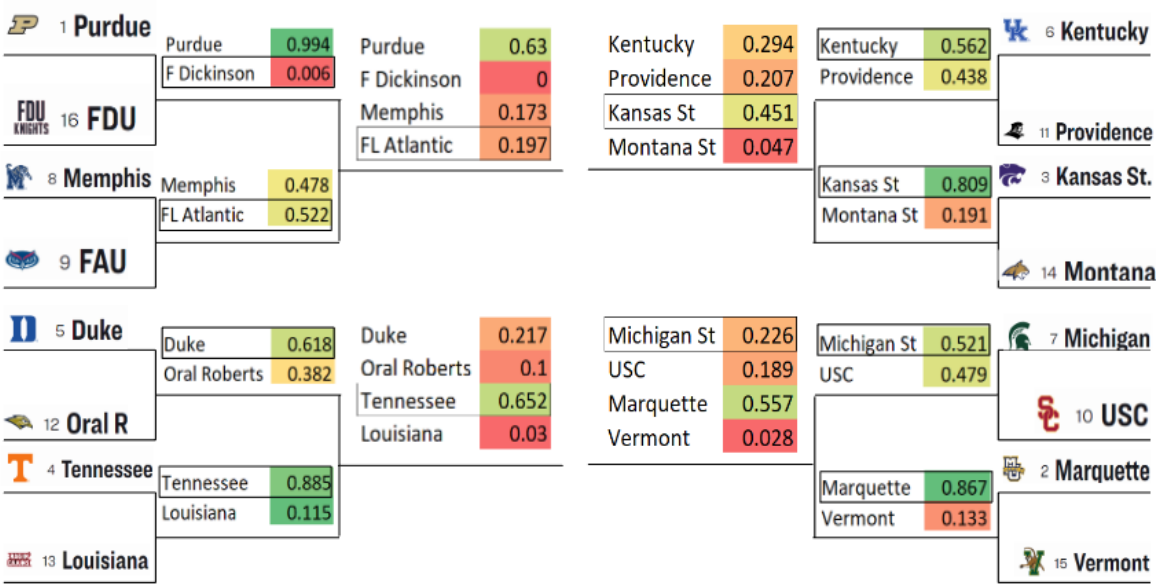

Figure 4 is an example of the simulation for just the East division from this year's NCAA tournament and the probabilities of each team to make it through the first two rounds. Note that generally the higher seeded teams are predicted to be more likely to win, except for one case where ninth seed FAU was simulated to beat the eighth seed Memphis (FAU went on to be the only Cinderella from this group). The exact probabilities are also useful because we can see just how much of a coin flip certain games are, such as Michigan St. vs. USC where there is less than a 5% differential in predicted win probability.

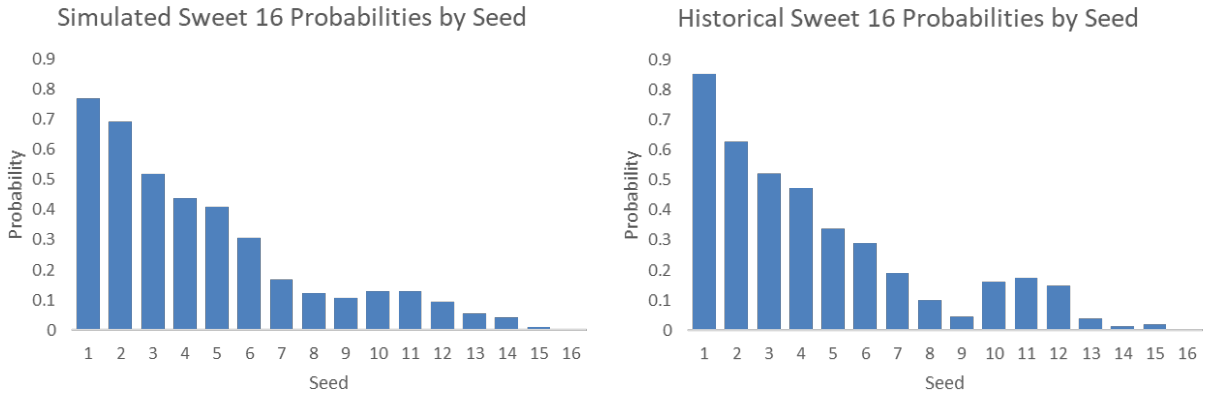

Figure 5: Simulated and historical probabilities of making it to the Sweet 16 by seed

The graph above on the left shows the simulation of the past five years grouped by seeds. When we compare this to the actual historical odds plot on the right, we see that it closely resembles the shape. There are slight differences: our model gives less chance to first seeds and a higher chance to eighth and ninth seeds, and historically, tenth and eleventh seeds have done better than what our model gives them credit for. However, in general, our model does reflect historical odds and tendencies. For example, tenth and eleventh seeds are correctly predicted to have a higher probability of making it to the Sweet 16 than eighth and ninth seeds. This is because eighth and ninth seeds generally have to match up with top seeds in the second round which greatly reduces their chances.

Analysis

To assess the accuracy of our model, we used a logistic loss function on all 8-16 seeds for the past 5 years. We inputted our predicted percentages and compared it against whether or not the teams were actually Cinderellas (with values 1 and 0). We used the equation:

$$-\frac{1}{n}\sum\{(y_i=1)*\log(P_i) + (y_i=0)*\log(1-P_i)\}$$

Our simulated predictions on the last five March Madness tournaments got a log loss of 0.313. We then calculated the log loss for the historical average percentage chance that each seed makes it to the Sweet 16. This generated a log loss of 0.316, meaning that our model was slightly better at predicting than the historical averages. However, this is a very small difference and indicates that our model is similar to basing Cinderella prediction off of the average odds of reaching the Sweet 16. This shows the difficulty of predicting March Madness outcomes. One potential reason that our model lacked accuracy is that we capped the point differentials in our power scores to 30. While this was meant to produce more accurate scores that did not overvalue blowouts, it is possible that this decreased the predictive power of our simulations.

While our model was not significantly better than the average rates, we also wanted to observe the 18 Cinderellas from the past 5 years and check if our model showed anything unique about them (note that this is a relatively small sample size). In other words, are past Cinderellas predicted to have a higher probability of making it by our model than their seed has on average?

Season	Seed	Team	Prob. above seed average	P.S. above seed average
2023	9	FL Atlantic	9.13%	1.779
2023	15	Princeton	1.02%	2.150
2023	8	Arkansas	9.71%	1.312
2022	8	North Carolina	2.57%	-0.678
2022	15	St Peter's	-0.42%	0.188
2022	10	Miami FL	-4.84%	-2.385
2022	11	Iowa St	3.95%	-0.227
2022	11	Michigan	2.57%	1.378
2021	11	UCLA	5.97%	1.517
2021	8	Loyola-Chicago	5.66%	2.577
2021	11	Syracuse	4.00%	1.150
2021	12	Oregon St	-1.34%	-1.824
2021	15	Oral Roberts	-0.36%	-1.467
2019	12	Oregon	4.97%	2.298
2018	11	Syracuse	-5.39%	-0.710
2018	9	Kansas St	-3.71%	-1.176
2018	11	Loyola-Chicago	0.85%	-0.290
2018	9	Florida St	7.45%	0.205
		Average	**2.32%**	**0.322**

Figure 6: Cinderellas probabilities and power scores above seed average predicted by model

From figure 6, we can see that 12/18 of the Cinderellas did have an above seed average probability to make it to the Sweet 16, with the average being 2.32% higher. Additionally, the right hand column shows the respective power score for each team. However, the correlation for the power score is more inconclusive. While the average is slightly positive, only 10/18 had an above average seed power score. Some of these teams such as North Carolina, Iowa St, and Loyola-Chicago (2018) had below average power scores but above average probability. This suggests that having favorable bracket matchups with weaker teams may actually matter more than a team's own power score.

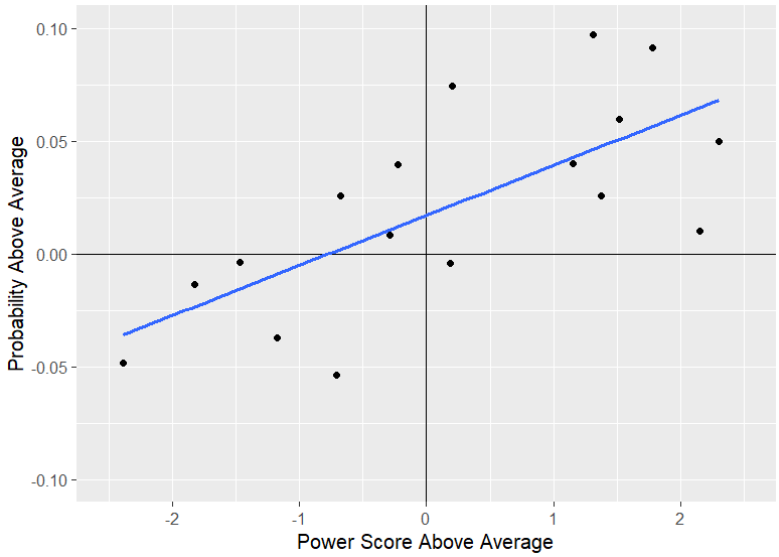

Figure 7: Scatter plot and trendline of power score vs probability above seed average

To look more closely at this relationship, the equation of the trendline in figure 7 is y = 0.0164 + 0.0212x where x is power score above average and y is measured in % probability of making it to the Sweet 16 above average by our model. Each point represents one of the 18 Cinderellas. It is interesting to note that even when the power score is average (x = 0), the y-int is 1.64%, meaning these Cinderellas are still predicted to have a 1.64% higher probability than average despite being average in power score. This reflects a significant boost due to the overvalued teams these Cinderellas generally faced.

We would also like to highlight that our model did exceptionally well predicting Cinderellas for 2023. There were three teams this year that fit our criteria: Princeton (a 15 seed), Florida Atlantic (a 9 seed), and Arkansas (an 8 seed). Our model predicted that Florida Atlantic had a 19.7% chance of reaching the Sweet 16, while the historic rate of 9 seeds reaching the Sweet 16 is 4.7%. This means our model predicted that FAU had a 15% higher chance of reaching the Sweet 16 than an average 9 seed, which is a significantly outsized probability. FAU is currently playing in the Final Four, meaning they far surpassed expectations for a 9 seed. Our model also predicted that Arkansas had a 22.1% chance of making the Sweet 16, which is 12% higher than the 8 seed average of 10.1%. Again, our model successfully predicted an outsized probability of a lower seed making the Sweet 16. A potential reason for their outsized probabilities of making the Sweet 16 is once again favorable matchups, particularly in the second round. The potential matchups for FAU in the second round were Purdue and Fairleigh Dickinson, and the potential matchups for Arkansas were Kansas and Howard. Purdue and Kansas had the two worst power scores of the one seeds and had power scores worse than several two, three, and four seeds. Howard and Farileigh Dickinson had the two worst power scores in the entire tournament. Given the seeds of FAU and Arkansas, they had relatively easy matchups, especially in the second round. These findings indicate that easy matchups may be the biggest indicator of success for lower-seeded teams.

In the absence of a firm conclusion in terms of predicting Cinderellas, we also tried to use a trait-based approach to see if we could find a commonality in terms of the team stats of Cinderella teams. Despite a low sample size, we ran a logistic regression with a dataset of all “cinderella eligible teams”, teams seeded 8 and over, in every tournament since 2013. We used the logistic model to try and predict the binary cinderella flag variable, which noted if they advanced to the sweet 16. Likely due to the sample size issue, this model did not produce any significant nor intelligent results.

Conclusion

Using power scores of each team and running through thousands of simulations gave us reasonable probabilities for lower seeded teams to make it to the Sweet 16. However, these odds were not significantly better than just using historical odds given the log loss of our predictions was 0.313 and the historicals was 0.316. This shows the difficulty of predicting Cinderellas with just power scores using point differential, strength of schedule, and court location. In terms of positively identifying historical Cinderellas, there were some promising results. 12/18 past Cinderellas did have an above seed average probability of 2.32% of making it to the Sweet 16. One component of this is because of the team’s power score. Cinderellas had higher power scores on average, but only slightly at 0.322. Favorable matchups may account for a larger component of a Cinderella’s predicted success. Given a team has to beat two teams en route to the Sweet 16, if they face easy matchups, or in other words low power score teams, the team will have a much higher probability to make the Sweet 16. For example, using the past 18 Cinderellas, given a power score that is average for their seed, our model still would have predicted a 1.64%

higher chance than average to make the Sweet 16. Lastly, other than a team's power score and matchups, there is a large factor of luck. Certain teams could be playing better or worse than their power score and anomalies of huge upsets can always happen, such as FDU beating Purdue this year which our model predicted to only have a 0.6% chance of happening.

Future Improvements

One method to improve our model would be to introduce dynamic power scores. These power scores would change throughout the tournament as teams advance, providing a more up-to-date representation of a team's strength. For example, after Fairliegh Dickinson beat Purdue, their power score would increase significantly and impact the prediction for their second-round game. This could potentially lead to more accurate predictions.

Another improvement is to include more inputs in the simulation to account for factors beyond what we included in our power scores. Our power scores account for point differences, strength of schedule, and home court advantage, but clearly there are more aspects that impact the outcome of a basketball game. For example, incorporating offensive and defensive ratings would highlight favorable and unfavorable matchups for teams that have strengths or weaknesses in certain areas.