March Madness Prediction Models

A Case Study on Predicting the Men's Division 1 NCAA Basketball Tournament

STOR 320-001 Group 8:
Chris, Alyson, Caroline, Sam and Nimalan

WHAT HAPPENED TO THIS YEAR'S UNC BASKETBALL SEASON?

- 2022 Runner-ups
- 2022-2023 Preseason #1
- Failed to reach this year's tournament
- So what determines tournament success?

Proposed Question

- Can round by round models be created that accurately predicts whether each team in the NCAA Division I men's basketball tournament should win or lose using only regular season statistics?

Our Data

- Analyzed the public dataset “College Basketball Dataset” posted on Kaggle by Andrew Sunberg, which consisted of...
  - Division I college basketball seasons from 2013- 2019
  - Regular season statistics and elimination rounds
  - 22 numeric variables
  - 3 categorical variables

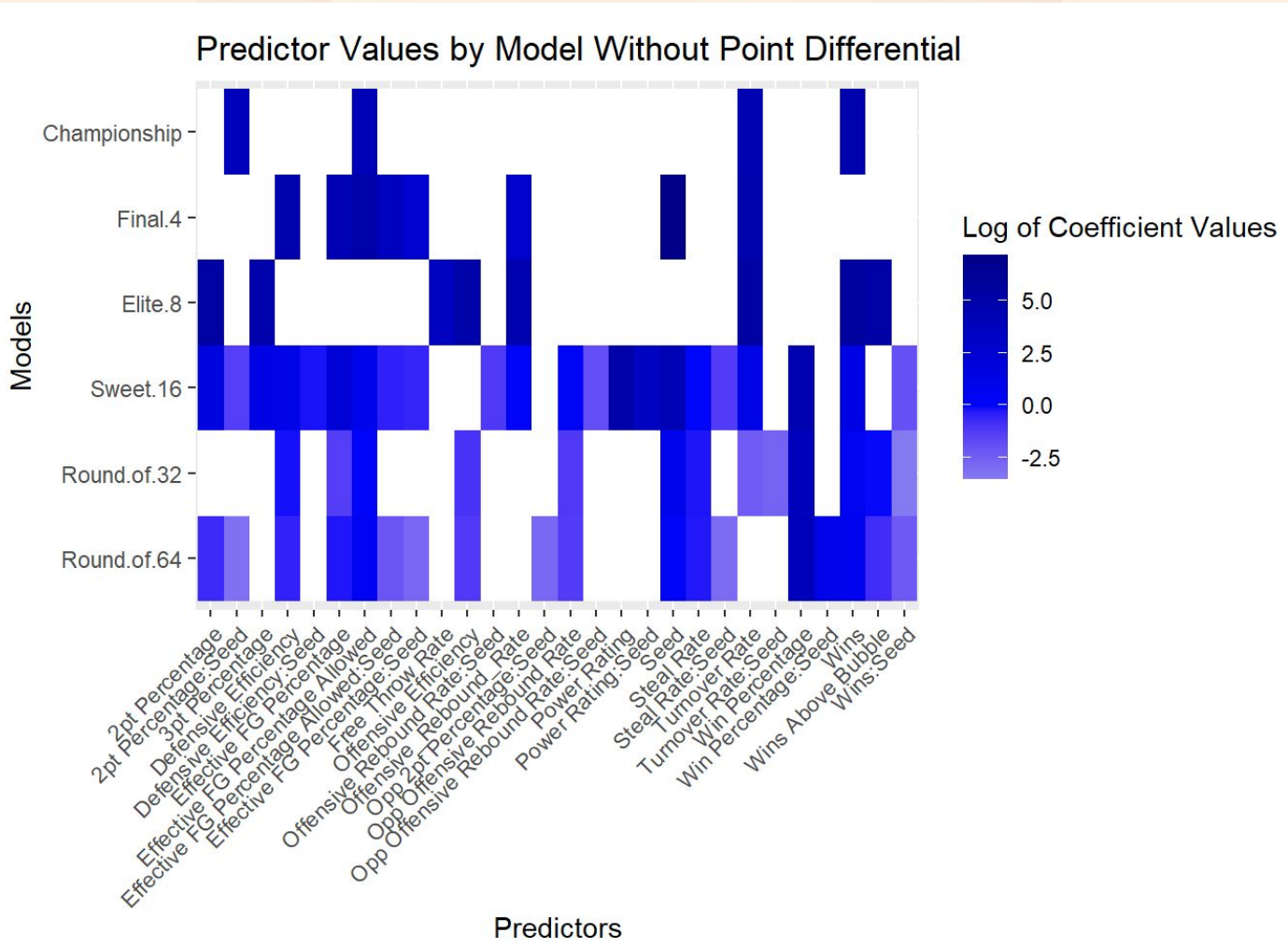
- Filtered the data to include only Round of 64 teams
- Added the variable “Win Percentage” for the regular season
- Manually created a new post season dataset that included binary win/loss variables for each round

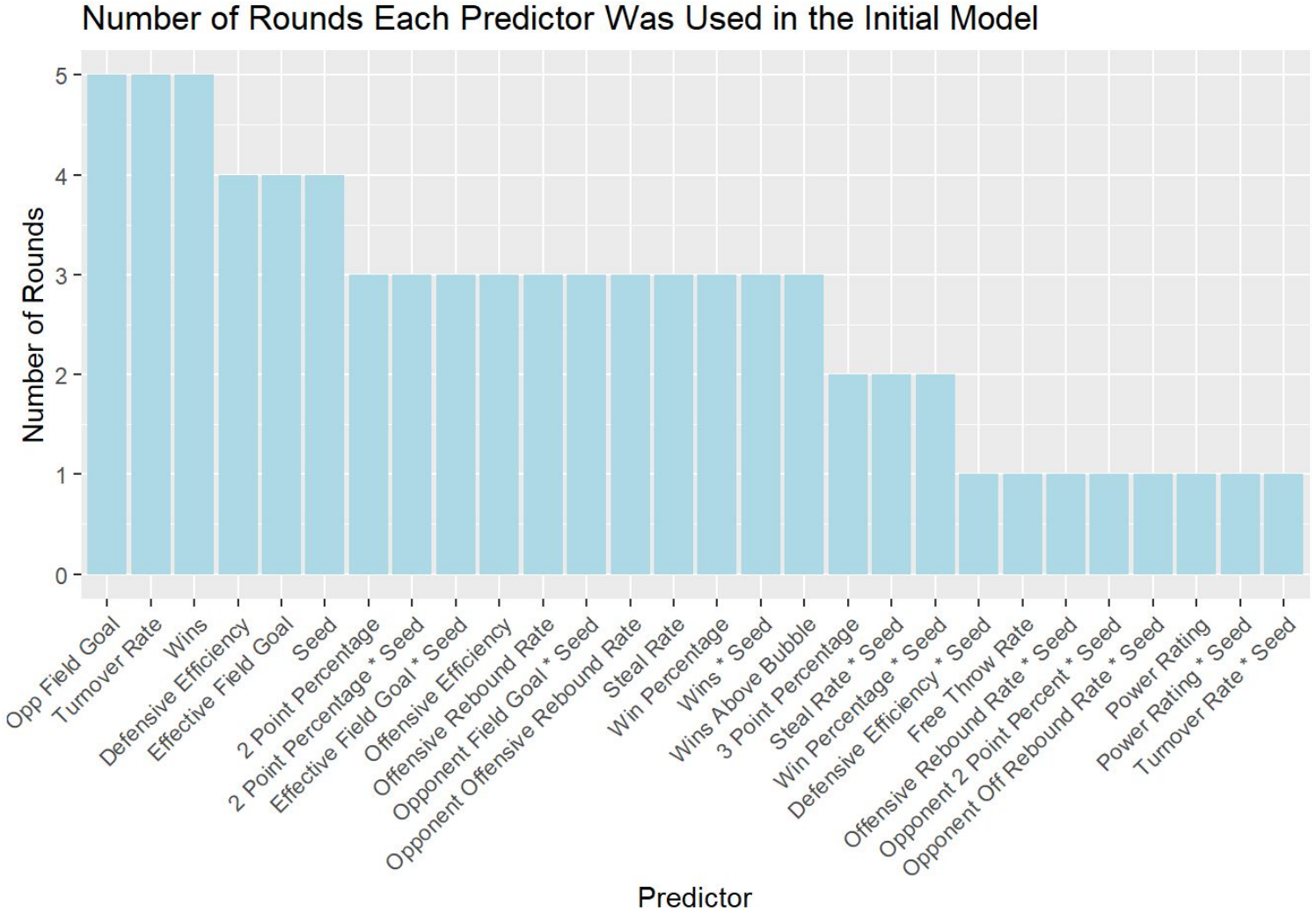
Sample of Our Dataset

Team	Year	Conference	Win Percentage	Defensive Efficiency	Seed	Round of 64 Opponent	Round of 64 W or L	Round of 64 PD	Round of 32 Opponent	Round of 32 W or L	Round of 32 PD
Akron	2013	MAC	0.7812500	94.4	12	VCU	0	NA	NA	NA	NA
Albany	2013	AE	0.6857143	98.7	15	Duke	0	NA	NA	NA	NA
Arizona	2013	P12	0.7714286	92.2	6	Belmont	1	17	Harvard	1	23
Belmont	2013	OVC	0.7741935	96.2	11	Arizona	0	NA	NA	NA	NA
Bucknell	2013	Pat	0.8181818	93.1	11	Butler	0	NA	NA	NA	NA
Butler	2013	A10	0.7428571	93.0	6	Bucknell	1	12	Marquette	0	NA
California	2013	P12	0.6363636	92.3	12	UNLV	1	3	Syracuse	0	NA
Cincinnati	2013	BE	0.6470588	88.3	10	Creighton	0	NA	NA	NA	NA
Colorado St.	2013	MWC	0.7272727	98.4	8	Missouri	1	12	Louisville	0	NA
Colorado	2013	P12	0.6363636	90.2	10	Illinois	0	NA	NA	NA	NA

Created models to predict tournament success

- Created a model for every round of the tournament using StepAIC to identify important variables in predicting that round
- Validated each model using n-Fold Cross Validation
- Note: These models were done only using regular season statistics

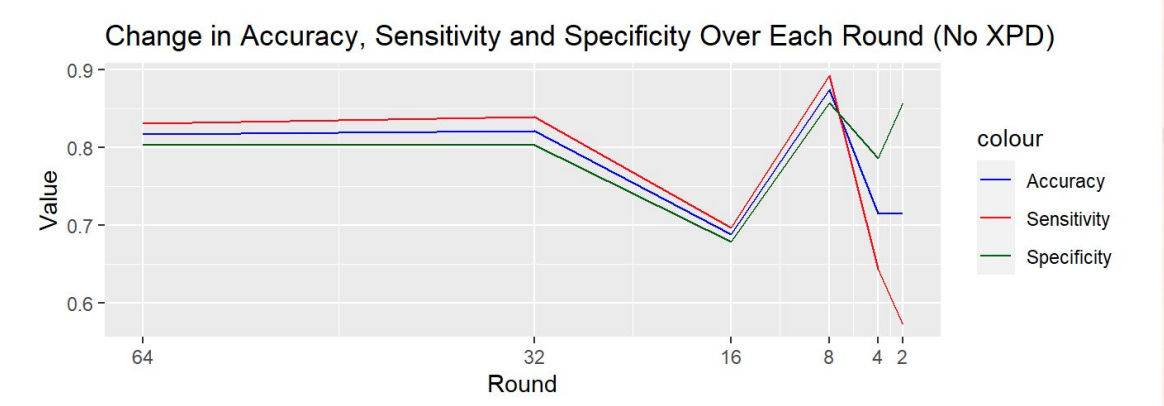

Predictor Values by Model Without Point Differential
Championship
Final.4
Elite.8
Sweet.16
Round.of.32
Round.of.64
Models
2pt Percentage
2pt Percentage:Seed
3pt Percentage
Defensive Efficiency
Defensive Efficiency:Seed
Effective FG Percentage
Effective FG Percentage Allowed
Effective FG Percentage Allowed:Seed
Effective FG Percentage:Seed
Free Throw Rate
Offensive Efficiency
Offensive Rebound Rate:Seed
Offensive Rebound_Rate
Offensive Rebound_Rate:Seed
Opp 2pt Percentage:Seed
Opp Offensive Rebound Rate
Opp Offensive Rebound Rate:Seed
Power Rating
Power Rating:Seed
Seed
Steal Rate
Steal Rate:Seed
Turnover Rate
Turnover Rate:Seed
Win Percentage
Win Percentage:Seed
Wins
Wins Above Bubble
Wins:Seed
Predictors
Log of Coefficient Values
5.0
2.5
0.0
-2.5


Number of Rounds Each Predictor Was Used in the Initial Model
Number of Rounds
0
1
2
3
4
5
Opp Field Goal
Turnover Rate
Wins
Defensive Efficiency
Effective Field Goal
Seed
2 Point Percentage
2 Point Percentage * Seed
Effective Field Goal * Seed
Offensive Efficiency
Offensive Rebound Rate
Opponent Field Goal * Seed
Opponent Offensive Rebound Rate
Steal Rate
Win Percentage
Wins * Seed
Wins Above Bubble
3 Point Percentage
Steal Rate * Seed
Win Percentage * Seed
Defensive Efficiency * Seed
Free Throw Rate
Offensive Rebound Rate * Seed
Opponent 2 Point Percent * Seed
Opponent Off Rebound Rate * Seed
Power Rating
Power Rating * Seed
Turnover Rate * Seed
Predictor

Results

- Our model's have an average sensitivity value of 74.55%
- Our model's have an average specificity value of 79.76%
- Our model's have an overall average accuracy value of 77.16%

Round	Sensitivity	Specificity	Accuracy
64	0.8303571	0.8035714	0.8169643
32	0.8392857	0.8035714	0.8214286
16	0.6964286	0.6785714	0.6875000
8	0.8928571	0.8571429	0.8750000
4	0.6428571	0.7857143	0.7142857
2	0.5714286	0.8571429	0.7142857



Conclusion

- Using statistical analysis and machine learning techniques, we created highly accurate models for predicting team performance and tournament outcomes
- These models have the potential to benefit...
  - sports analysts
  - sports fans
  - tournament organizers

- Our model has the potential to be modified to suit any sports competition worldwide

Works Cited

David, Juan Paolo. "'Karma Gets Ya over Time' - Preseason No. 1 UNC Set to Miss March Madness 2023, and Fans Take over with Hilarious Memes and Reactions." *Sports News*, Sportskeeda, 13 Mar. 2023, https://www.sportskeeda.com/basketball/news-karma-gets-ya-overtime-pre-season-no-1-unc-set-miss-march-madness-2023-fans-take-hilarious-memes-reactions.

National Collegiate Athletic Association. "Browse Every NCAA Bracket since 1939 with Stats and Records." NCAA.com, NCAA.com, 16 Mar. 2021, https://www.ncaa.com/basketball-men/d1/every-ncaa-bracket-1939-today-tournament-stats-records.

Sundberg, Andrew. "College Basketball Dataset." Kaggle, 16 Mar. 2021, https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset.