Deloitte.

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK | Gabelli School of Business

8th Annual March Data Crunch Madness 2021

TABLE OF CONTENTS

Deloitte.

FORDHAM | Gabelli School of Business

Problem Statement

The term 'March Madness' was first used in reference to basketball by an Illinois high school official, Henry V. Porter, in 1939.

There have been 80 NCAA tournaments between 1939 and 2019. Kentucky has the most NCAA tournament appearances (58) and NCAA tournament wins (129)

Again, it's Kentucky leading the way. The Wildcats have 129 NCAA tournament wins, for an average of 2.2 wins per appearance. The Tar Heels are right behind with 126 wins, or 2.5 per appearance.

This year we have a lot of uncertainty for 2021 March Madness, and we are going to predict the winner using current and historical data sets.

Initial Hypothesis

Based on our initial hypothesis, we believe **Gonzaga** will win the tournament.

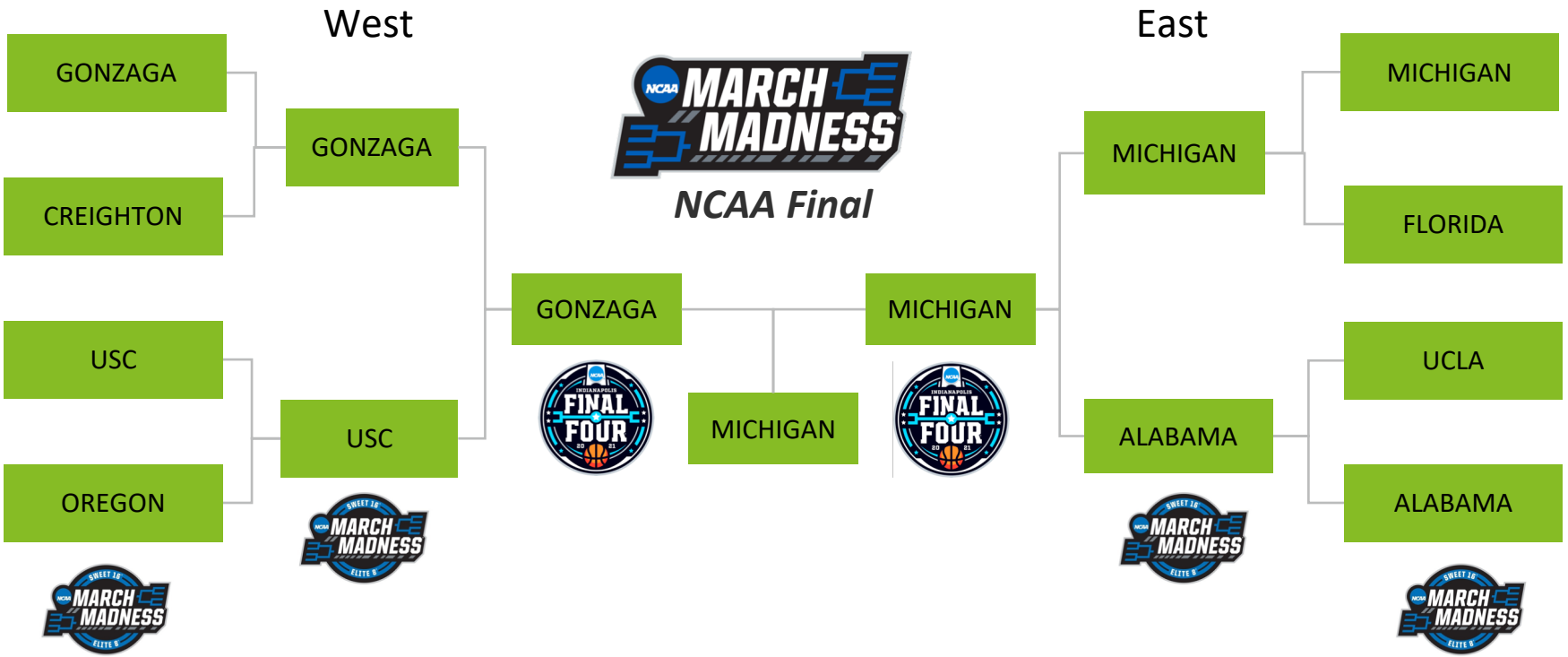
They lead in both overall **scoring per game** and **field goal percentage**. Guess the adage of "offense wins championships" is true.

Based on the provided **historical** and data that we pulled from ESPN **current** data we would pick Gonzaga to win it all.

According the problem statement, we still need **machine learning models** to better predict the winner of 2021 March Madness.

South and Midwest Division :
Sweet Sixteen to the NCAA Final Prediction

West and East Division :
Sweet Sixteen to the NCAA Final Prediction



NCAA Final Four to the Winner Prediction

Using our model prediction:
Michigan has **72.8%** probability
of winning when facing Houston

Introducing the Data

Initially using data from 2002 until 2020 to train and test until the recent 2021 data is released

01 **Game Data**
Game id, host name and latitude and longitude and score

02 **KenPom Data**
Efficiency, four factor data, tempo, etc.

03 **Coaching Data**
Coach name, career wins, season wins, NCAA tournament appearances

04 **Team Location Data**
Latitude and longitude of team1 and team2 home base location

05 **Poll Data**
AP preseason/final polls, coaches' preseason/final polls

06 **RPI Data**
A quantity used to rank sports teams based upon a team's wins and losses

Feature Selection

indirect-impact data which indicates the differences between two teams including:

Coaching experience

This refers to whether the coach has ever led a team to the Final Four

Offense efficiency / defense efficiency

While these were given initially as separate features, we went on to combine them under the banner of log 5

External ESPN data

We collected external ESPN Data to validate our results, including each team's Point Per Game (PPG)

Historical scores / rankings

We calculate historical score and found out the ability for one team to create a score gap over others (iterative ranking score)

Log 5 probability

A formula invented by Bill James to estimate the probability that team A will win a game, based on the true winning percentage of Team A and Team B. P(W) = (A - A B) / (A + B - 2A*B)

External factors impacts, distance

We calculated each team's overall distance to the stadium they would be playing at

Rating System for Historical Scores

Point Spread of Each Game

- We created a rating system according to each game's point spread

team	spread	opponent
F Dickinson	6.0	Prairie View
Belmont	11.0	Temple
N Dakota St	4.0	NC Central
Arizona St	9.0	St John's
Minnesota	10.0	Louisville
LSU	5.0	Yale

Teams' Rating Based on Point Spread

- Rating system represents the ability of each team to perform a higher score gap

team	rating
North Carolina	14.219692
South Carolina	14.094053
Michigan	13.797574
Florida	13.587028
Connecticut	13.537197
...	...
N Colorado	-9.965363
Prairie View	-10.516966

Prediction Models

Logistic Regression

Provides a measure of a coefficient size is, as well as its direction of association

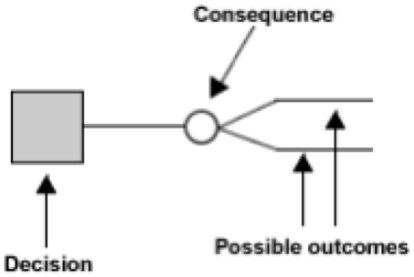
According to the feature selection, we also calculate the coefficient of each feature

$$p = \frac{1}{1 + e^{-y}}$$

Decision Tree

The classification process easy to visualize and understand

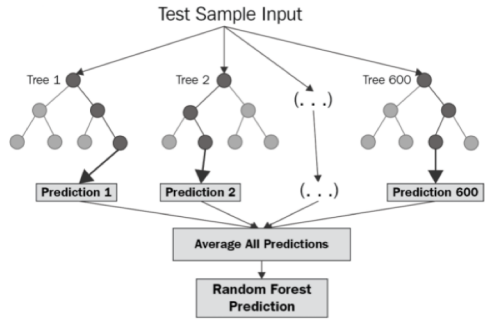
The accuracy of model heavily affected by the decision of splitting features



Random Forest

The random sampling technique used in the selection of the optimal splitting features

Provides the highest accuracy, but time-consuming



Model results

Overall Performance

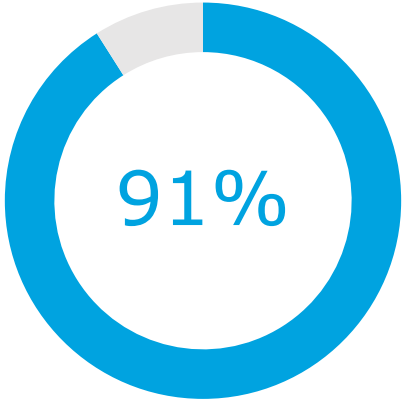
based on 2019 performance

	Logistic Regression	Random Forest	Decision tree
Area under Curve (AUC)	0.91 by cross validation	0.85	0.80 by cross validation
Log loss	0.48	0.5	0.55
Probability Results	Not Contain 1,0	Contain 1,0	Contain 1,0
Cross Validation	Yes	No , time consuming	Yes

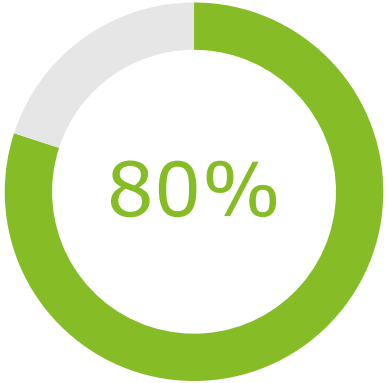
Model results

Area under Curve (AUC)

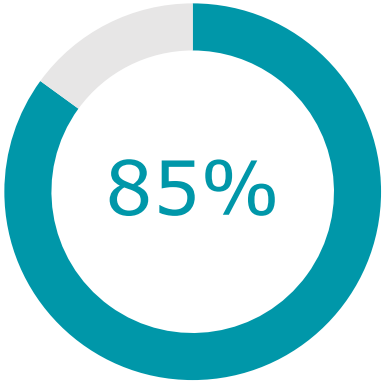
based on 2019 performance



Logistic Regression



Decision Tree



Random Forest

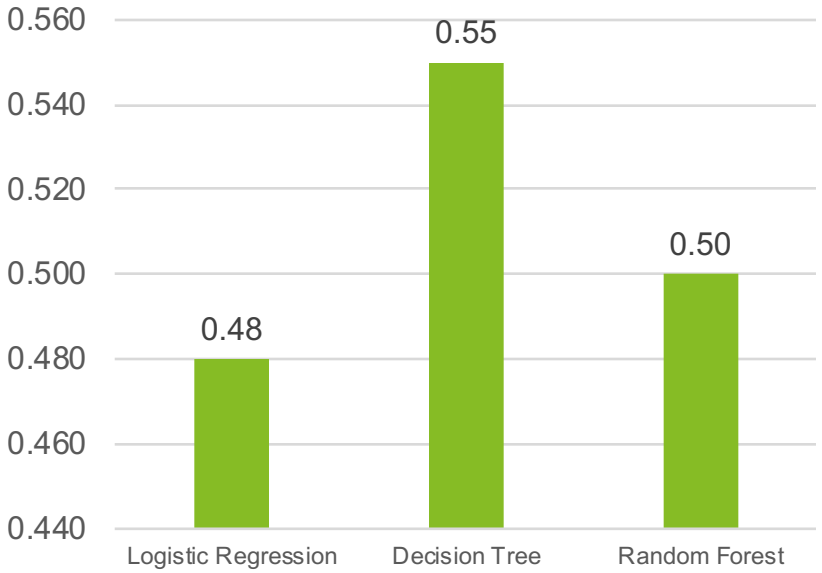
Model results

Log Loss

based on 2019 performance

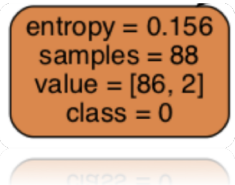
Log Loss is a measure of uncertainty (entropy), so a low Log Loss means a low uncertainty/entropy of your model.

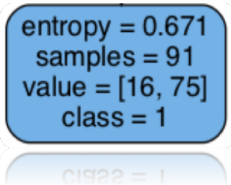
A lower log loss value means **better** predictions.



Comparing Winners and Losers

Decision Tree Nodes Testing



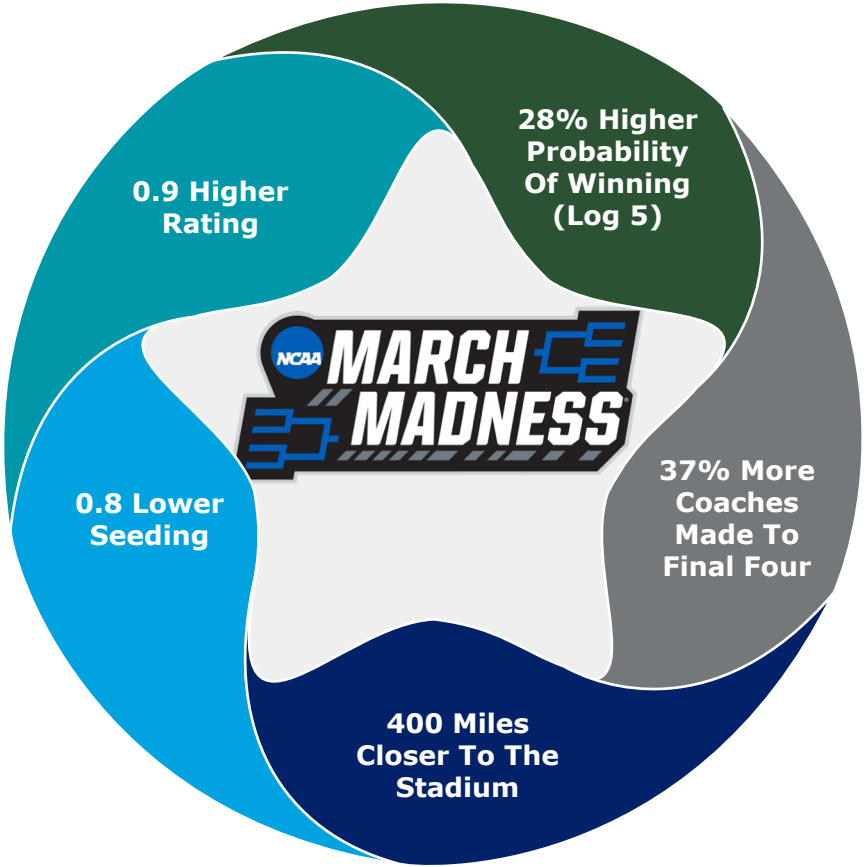


Team 1 **Losing** Node Example	Team 1 **Winning** Node Example
log 5 prediction <= **15.5%** team 1's rating no better than team 2	log 5 prediction >= **76.7%** team 1 200 miles closer to the stadium
The total sample size is **88** The accuracy of this node is **0.977**	The total sample size is **91** The accuracy of this node is **0.824**
The model predicted team 1 as **losing**	The model predicted team 1 as **winning**

Portrait of Winning Team

What It Takes to Win

By Analyzing 2002-2019 data, we observed that on average a winning team has:



Conclusion & Future Improvement

We believe that we did a phenomenal job with our feature engineer and model selection because we strictly focus on the comparison of two teams.

Initially through a point spread/score gap that was later aided by the addition of current season data.

We adopted an external dataset from ESPN that highlighted the current season's team statics which we then used to validate our models and predictions that were built on the provided historical date.

One major improvement that we would like to see is the data drilled down to the player level.

We believe that star players are crucial in the final games, especially in the strong teams under final game pressure. We want to include start payer information as one feature when predicting final game results.

An example at the professional level would be if Lebron James was out for an extended period, the Lakers would potentially be missing out on 30 points a night. Is this number something a team can recover or compensate for or does one player truly make a team.

Deloitte.

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK | Gabelli School of Business

Q & A

References

- Iterative Strength Rating: https://blog.collegefootballdata.com/talking-tech-bu
- Log 5 Wikipedia: https://en.wikipedia.org/wiki/Log5
- Decision Trees for Decision Making: https://hbr.org/1964/07/decision-trees-for-decision-making

Deloitte.

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK | Gabelli School of Business

Thank you.

8th Annual March Data Crunch Madness 2021

Let's Trade Stocks!

Yinzhe Lu, Jiacheng Zhang, Yue Wang, Timothy Welles