NCAA March Madness Data Crunch Competition Report
Gabelli School of Business

Jennie Le

Overview: What is March Madness ?

- The NCAA Division I men’s basketball tournament
- Formed by 68 Division I level college basketball team
- The competition is divided by several rounds into:
  - First Four
  - First and Second Round
  - Sweet 16/Elite Eight
  - Final
- The prediction is to fill out the championship bracket !!!!

Problem Statement

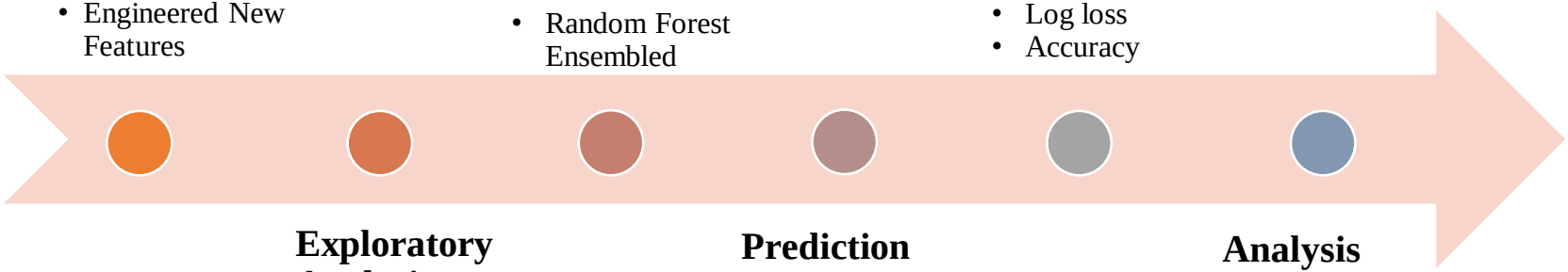
What are we proposed to do?

- Descriptive analytics with data visualization on data from 2002 to 2019.
- Prediction model building on data from 2002 to 2019.

Why is this important?

- A reference for folks who would like to make a bet on the game!!!
- Used by NCAA teams to make improvement for their performance.
- Used by NCAA to adjust their marketing strategies such as ticket price sales.

Methodology Diagram



Data Preprocessing
- Data Split
- Clean Data
- Engineered New Features

Exploratory Analysis
- Visualization

Feature Selection
- Highy Correlated Variables Removal
- Random Forest Ensembled

Prediction Model Building
- Logistic regression
- Random forest
- Linear discriminant analysis (LDA)
- Gradient Boosting
- Support Vector Classifier (SVC)

Prediction Model Evaluation
- Log loss
- Accuracy

Analysis Based on Prediction
- Visualization

Data Preprocessing

Derived Novel Features

- Difference = Team1 Feature N – Team2 Feature N
- Ratio (Quotient) = Team1 Feature N / Team2 Feature N
- Teamwork Score (Ability) = 0.8 x Team arate + 0.2 * Team Adjde
- Win Rate = Wins / (Wins + Losses)

Split Data (Used in prediction model building)

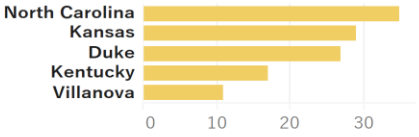
- Training Data: 2002 – 2018; Test Data: 2019
- Data needed prediction: 2020

TOP NCAA TEAM BY SEED NUMBER

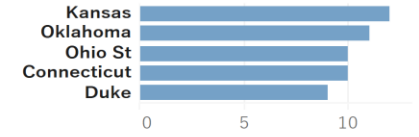
The committee will create a "seed list" (i.e. rank of the teams in "true seeds" 1 through 68) which is used to assess competitive balance of the top teams across the four regions of this national championship. Additionally, the seed list reflects the sequential order with which teams will be placed in the bracket

The dashboard represents team frequency assigned for each seed from 2002 to 2019

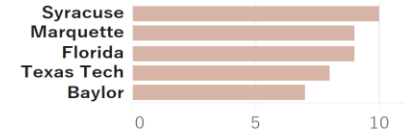
SEED 1



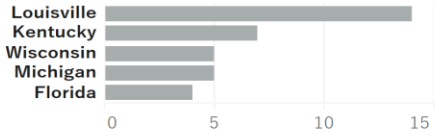
SEED 2



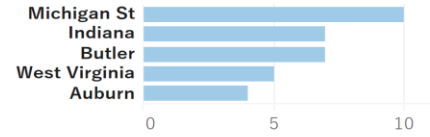
SEED 3



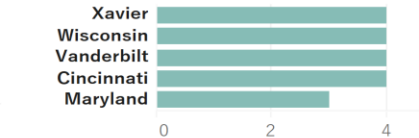
SEED 4



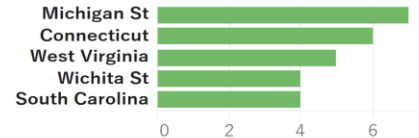
SEED 5



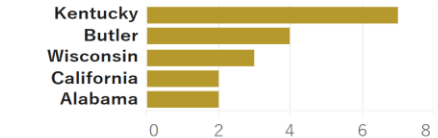
SEED 6



SEED 7



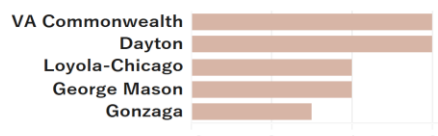
SEED 8



SEED 9

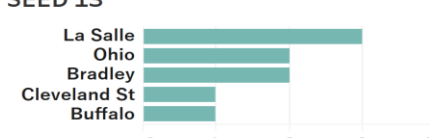
SEED 10

SEED 11



SEED 12

SEED 13



SEED 14

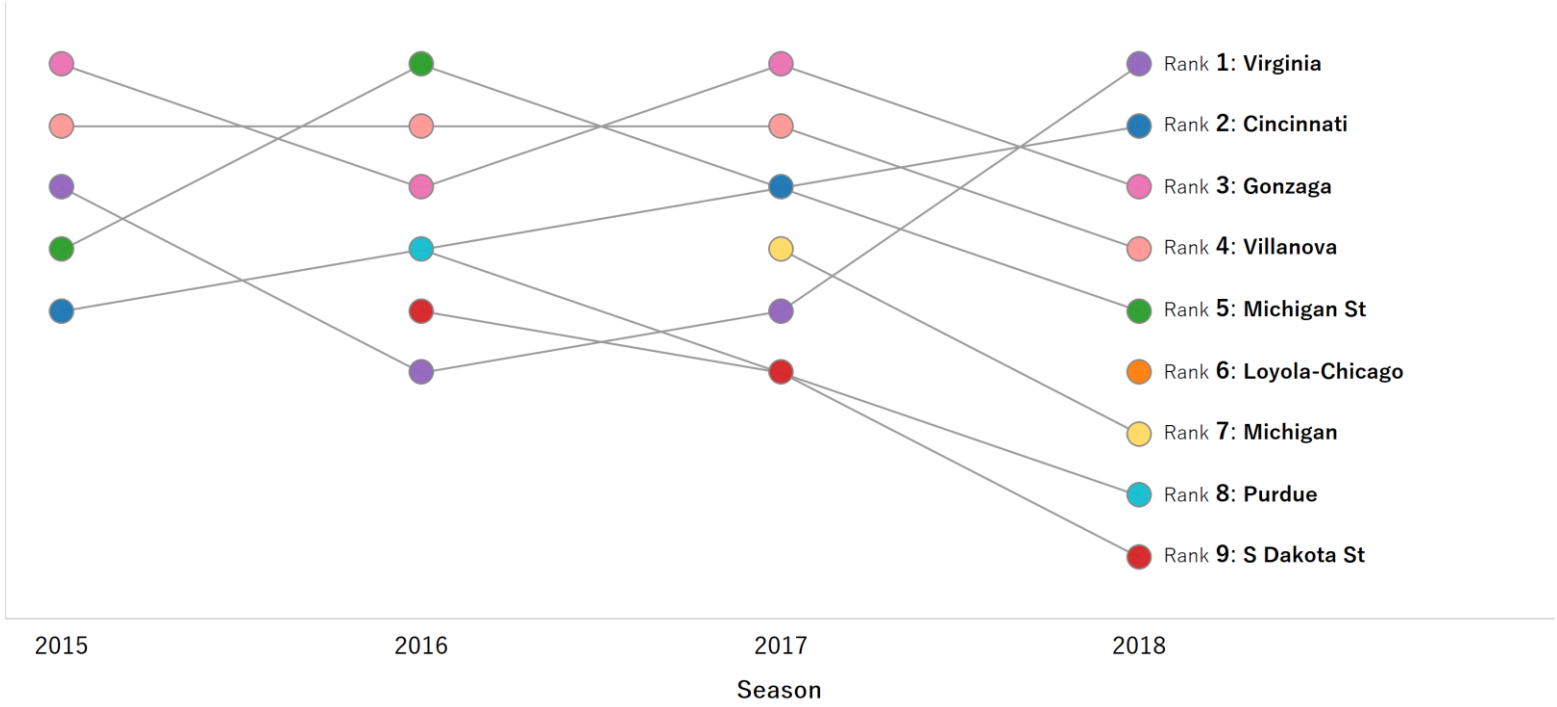
SEED 15

SEED 16

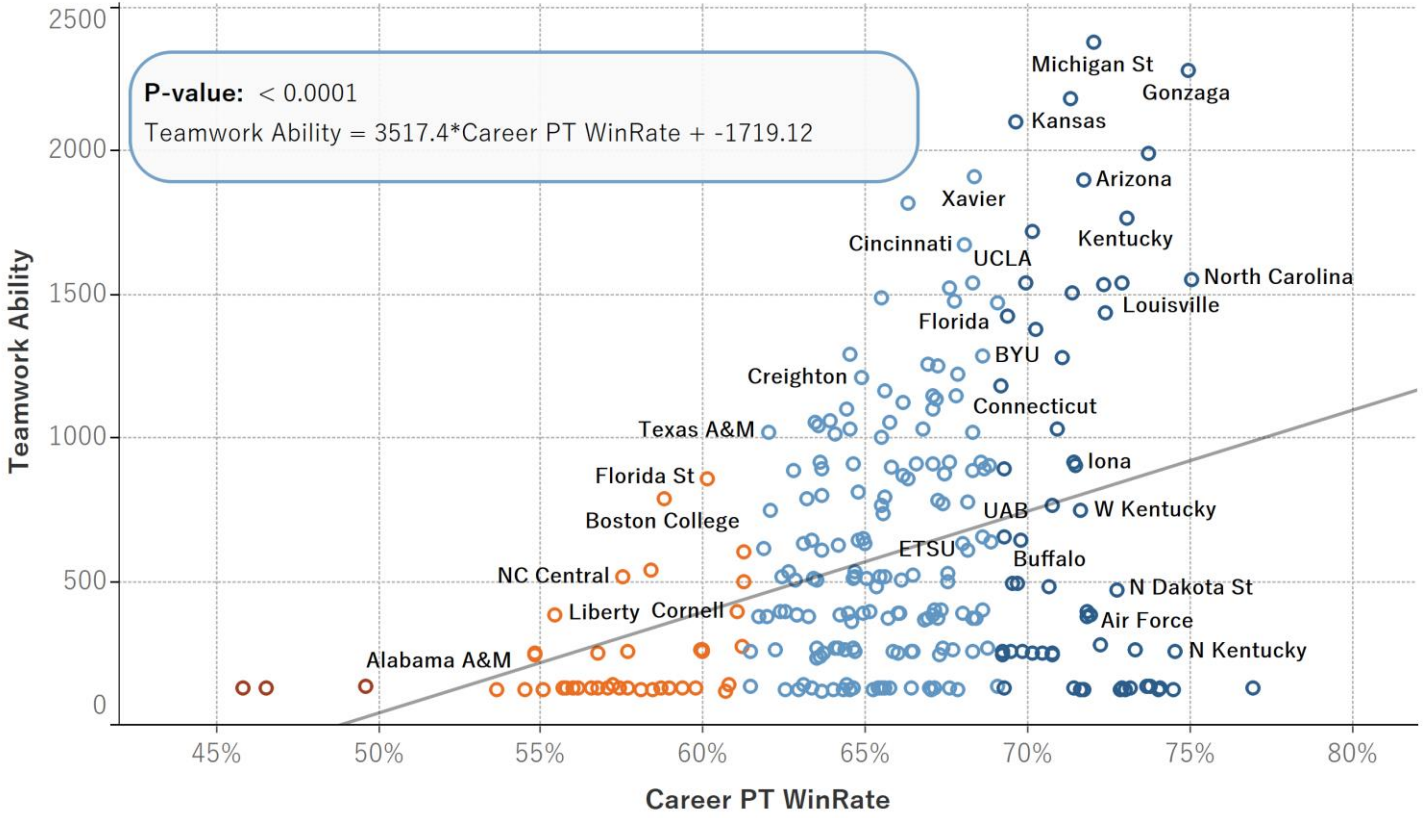
Top 10 Team Season Win Ranking (2018)

Teams are **ranked** based on the number of **wins per season**.

The dashboard represents how each team perform from 2015 to 2018 based on their season wins. In 2018, **Virginia** can be considered as **the strongest team.**



Can **teamwork** impact **the coach's number of wins per season** ?



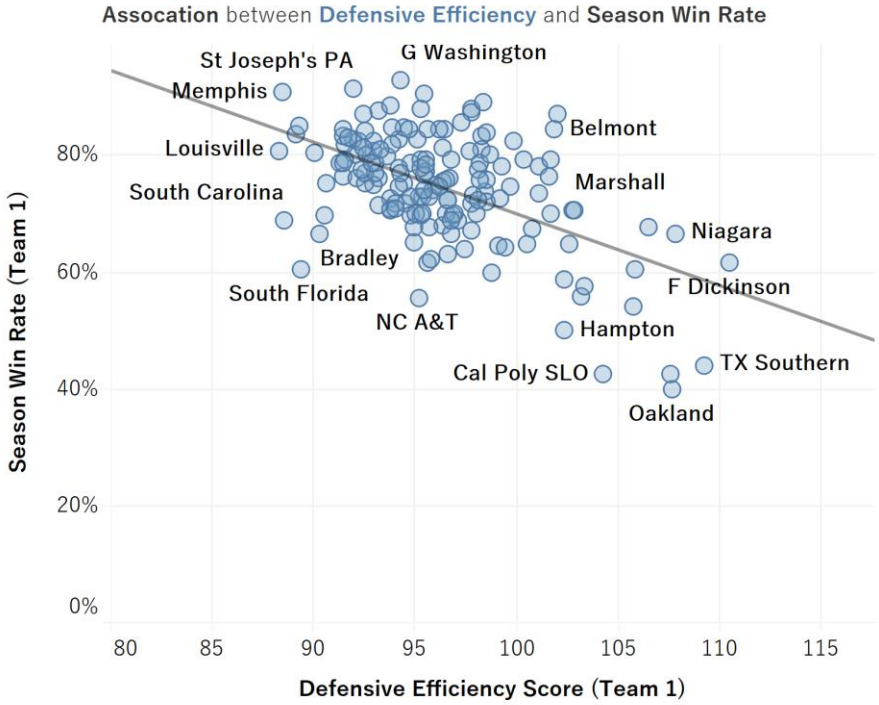
Career PT WinRate

46% 77%

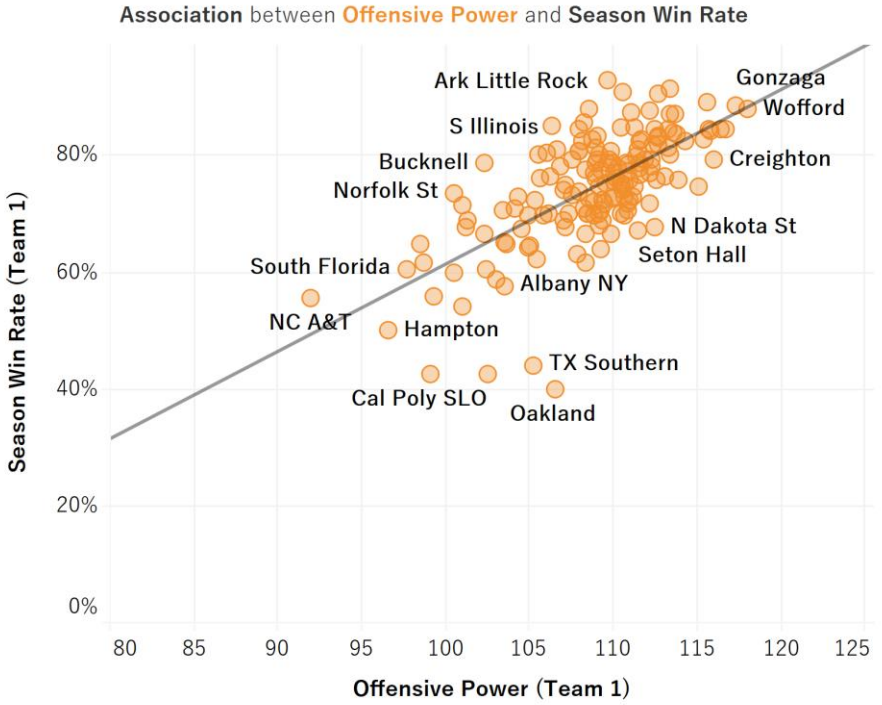
- **Teamwork ability** has a **positive correlation** with **the coach career win rate**. The better teamwork trainning a team received, the higher chance a team can win.

- **Teamwork ability** is calculated as the sum of **80%** of number of **assists** and **20% defensive effeciency** score (teamwork ability = 80% x arate + 20% x adjde)

Can a team have a **higher probability** of **winning** with effective **defense** and **offense** strategy ?



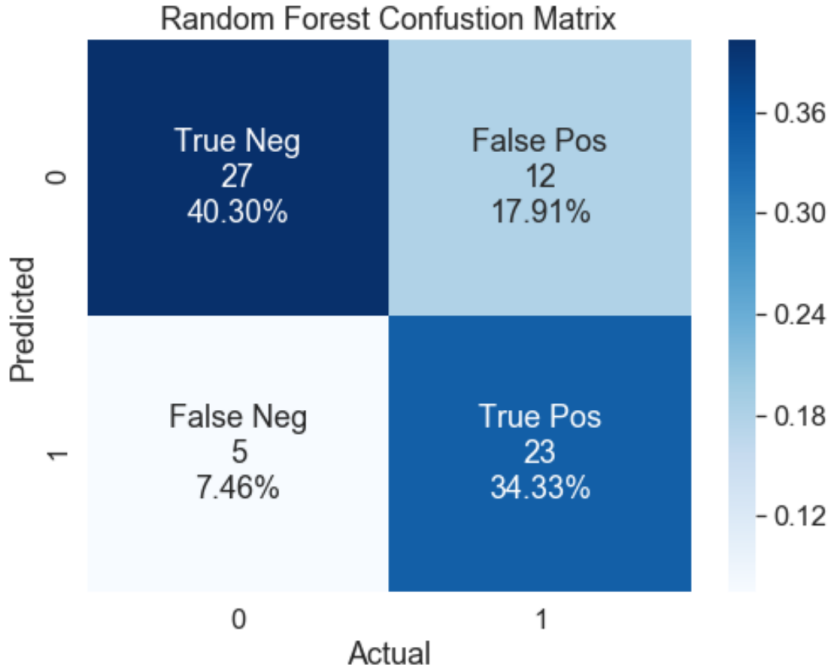
P-value: < 0.0001
Equation: Season Win Rate = 0.0149458*Avg. Oe + -0.881355



P-value: < 0.0001
Equation: Season Win Rate = -0.0122756*Avg. Adjde + 1.9276

Model Selection & Evaluation

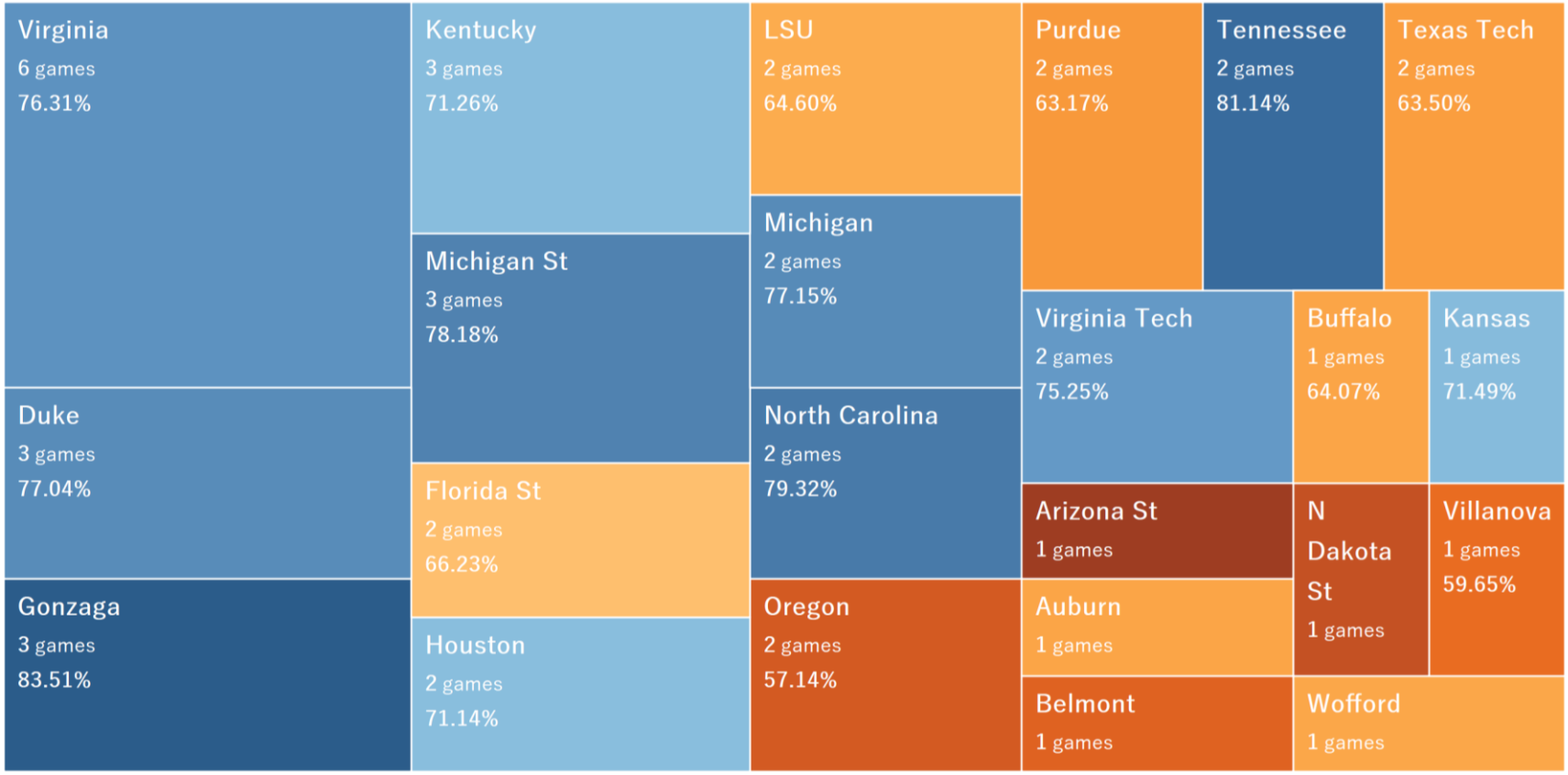
Model	Accuracy Score	Log Loss
Logistic Regression	0.761194	0.563166
Gradient Boosting	0.761194	0.545979
Support Vector Classifier	0.626866	0.680563
Random Forest Classifier	0.776119	0.513814
Linear Discriminant Analysis	0.746269	0.539230



According to the confusion matrix, accurate predictions account for 75% of the outcomes.

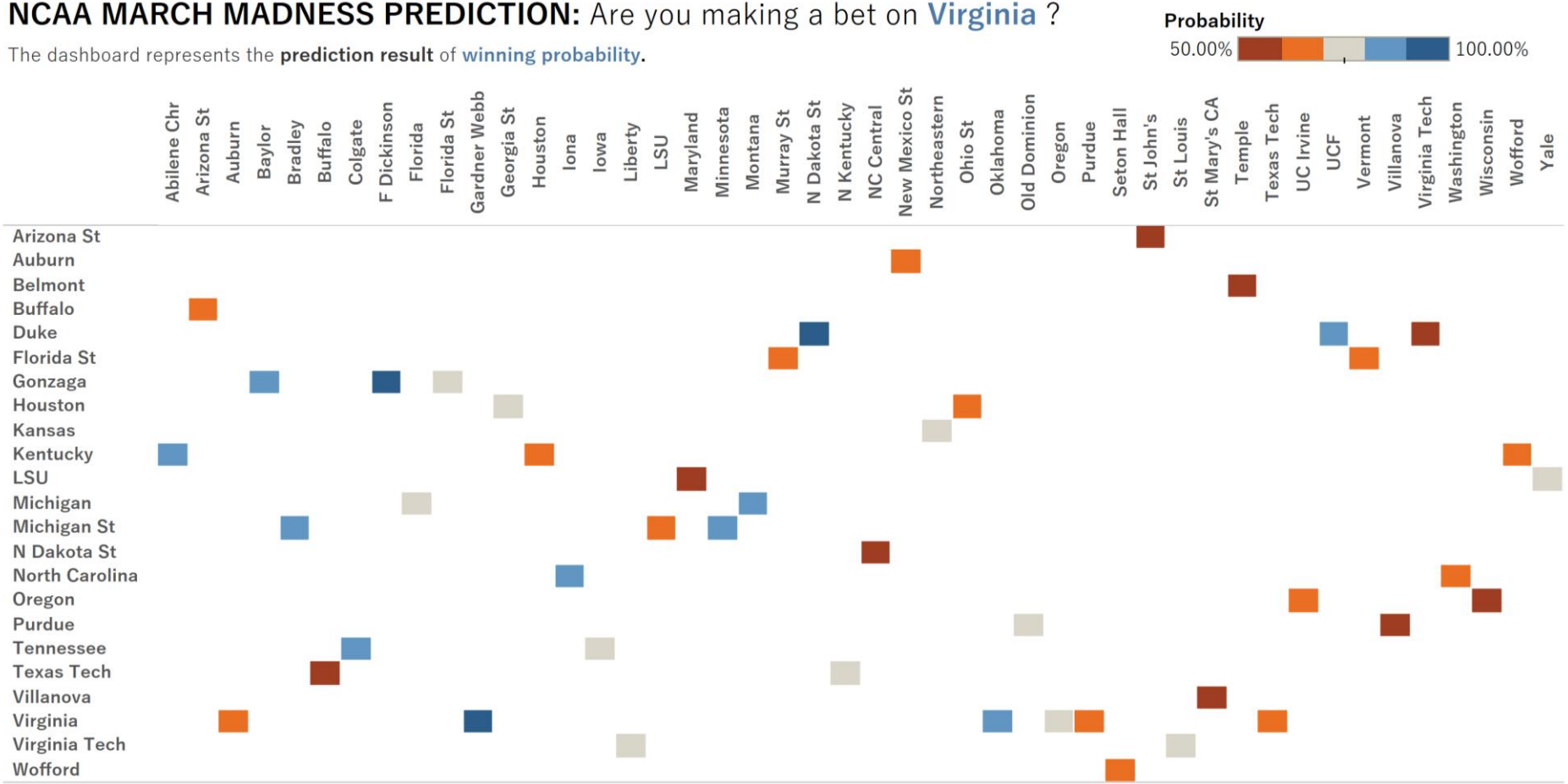
Prediction: Team 1 Wins

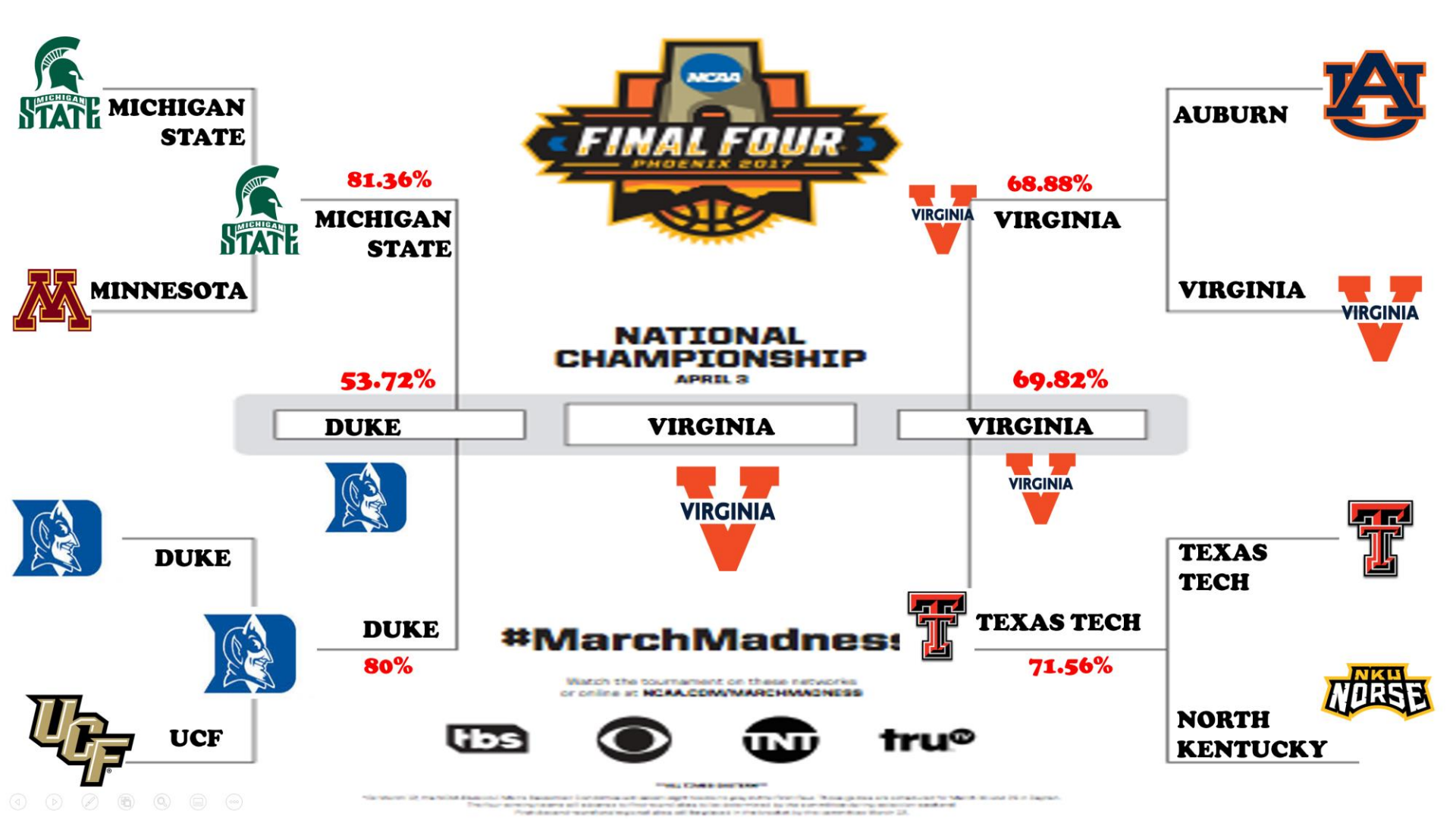
The treemap represents number of **Games** and **Average Probability** for each team.



NCAA MARCH MADNESS PREDICTION: Are you making a bet on Virginia ?

The dashboard represents the **prediction result** of **winning probability**.




MICHIGAN STATE
MINNESOTA
81.36%
MICHIGAN STATE
DUKE
UCF
DUKE
80%
53.72%
DUKE
FINAL FOUR
PHOENIX 2017
NATIONAL CHAMPIONSHIP
APRIL 3
VIRGINIA
#MarchMadness
Watch the tournament on these networks or online at NCAA.COM/MARCHMADNESS
tbs
truTV
AUBURN
VIRGINIA
68.88%
VIRGINIA
VIRGINIA
69.82%
VIRGINIA
TEXAS TECH
TEXAS TECH
71.56%
NORTH KENTUCKY

Reference

- Adit, D. (2017, March 12). Applying Machine Learning To March Madness. Retrieved from https://adeshpande3.github.io/Applying-Machine-Learning-to-March-Madness
- Conor, D. (2018, March 15). Machine Learning Madness: Predicting Every NCAA Tournament Matchup. Retrieved from https://towardsdatascience.com/machine-learning-madness-predicting-every-ncaa-tournament-matchup-7d9ce7d5fc6d
- Kaggle. (2017, April 4). March Machine Learning Mania 2017. Retrieved from https://www.kaggle.com/c/march-machine-learning-mania-2017
- Lotan, W. (2019, April 21). How We Predicted March Madness Using Machine Learning. Retrieved from https://medium.com/@lotanweininger/march-madness-machine-learning-2dbacc948874