Determining the Most Valuable Predictive Stats for March Madness

Our Team

Kevin Yin

Class of 2024

Harry Li

Class of 2024

Ned Donovan

Class of 2025

Rishi Mandapaka

Class of 2025

Agenda

1. Purpose
2. Data Exploration
3. Models
4. Insights
5. Conclusions

Purpose

Sports fan chase after a perfect March Madness bracket every year. Is there a way to utilize the statistics of the teams to better choose your picks?

Every team in the NCAA has an abundance of stats that are tracked. How can we determine which ones are the best predictors for a win?

Key Questions

Which statistics are most important in predicting a win?

Which teams are underperforming compared to their competitors

How well can we predict a win based off team statistics?

Project Goals

01 Determine whether a team will win or lose depending on the statistics between the matchup

02 Determine which statistics have the largest impact on winning a match

03 Visualize how the statistics of college basketball teams have changed over time

Data Exploration

CBB Dataset Overview

The College Basketball dataset was published on **Kaggle** by Andrew Sundberg in 2021

Includes both **offensive** and **defensive** team statistics and advanced metrics

Data ranges from **2013-2019**

Provides data on all **353** Division 1 collegiate basketball teams each season

MM Dataset Overview

The March Madness dataset was published on **Kaggle** by Woody Gilbertson in 2021

Includes basic game data for both the **winning** and **losing** team in each matchup

Data ranges from **1985-2021**

Provides data on all **67** games from each year's tournament

Merged Dataset Overview

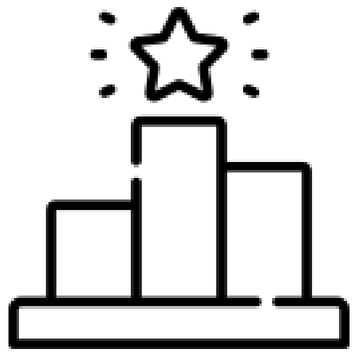
Both datasets were merged using the **year** and **team name** columns

Includes both game results alongside team statistics for both the winning and losing team

Data ranges from **2013-2019**

Provides a more **holistic** view of each matchup between two teams

Data Insights



BARTHAG

MM Team Average: **0.796**

Highest Team Average:

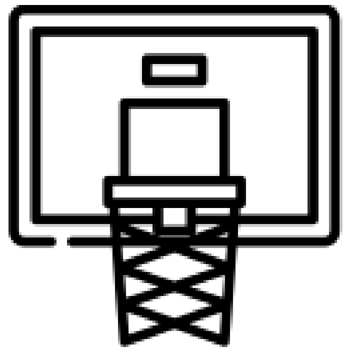
0.959 (Virginia)

Wins

MM Team Average: **24.643**

Highest Team Average:

31.714 (Gonzaga)



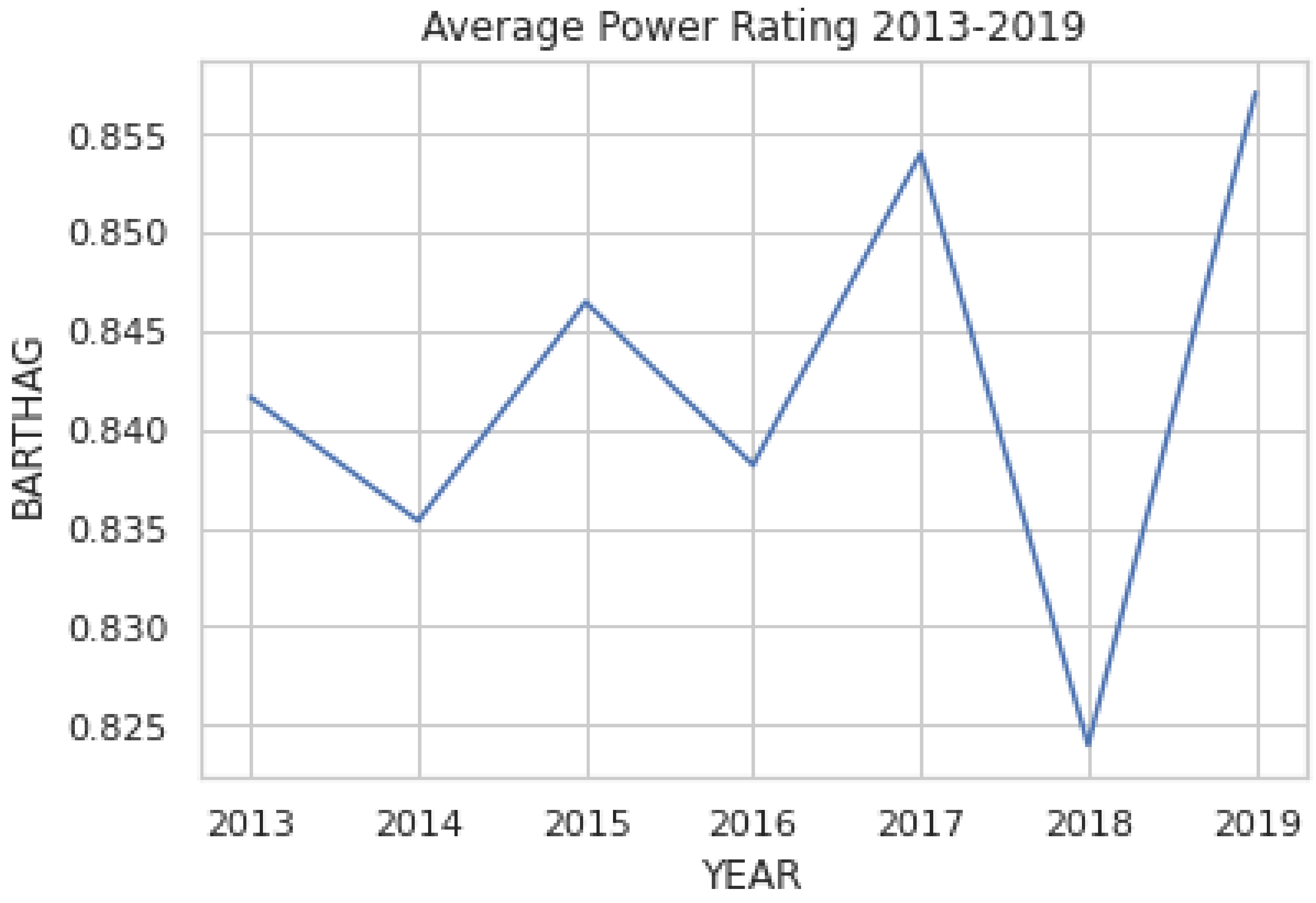
Adjusted Ratio

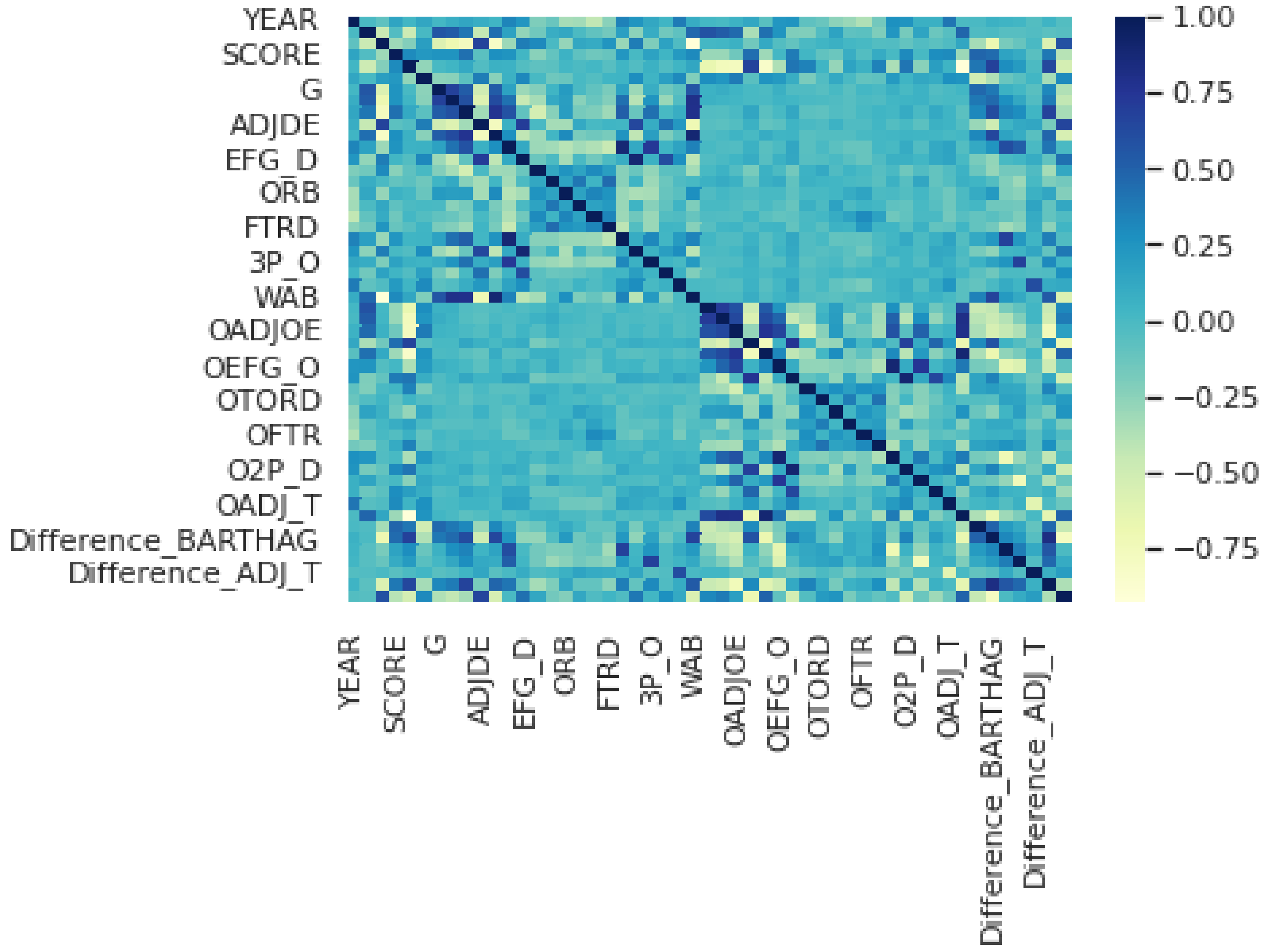
MM Team Average: **1.158**

Highest Team Average:

1.318 (Virginia)

Initial Visualizations





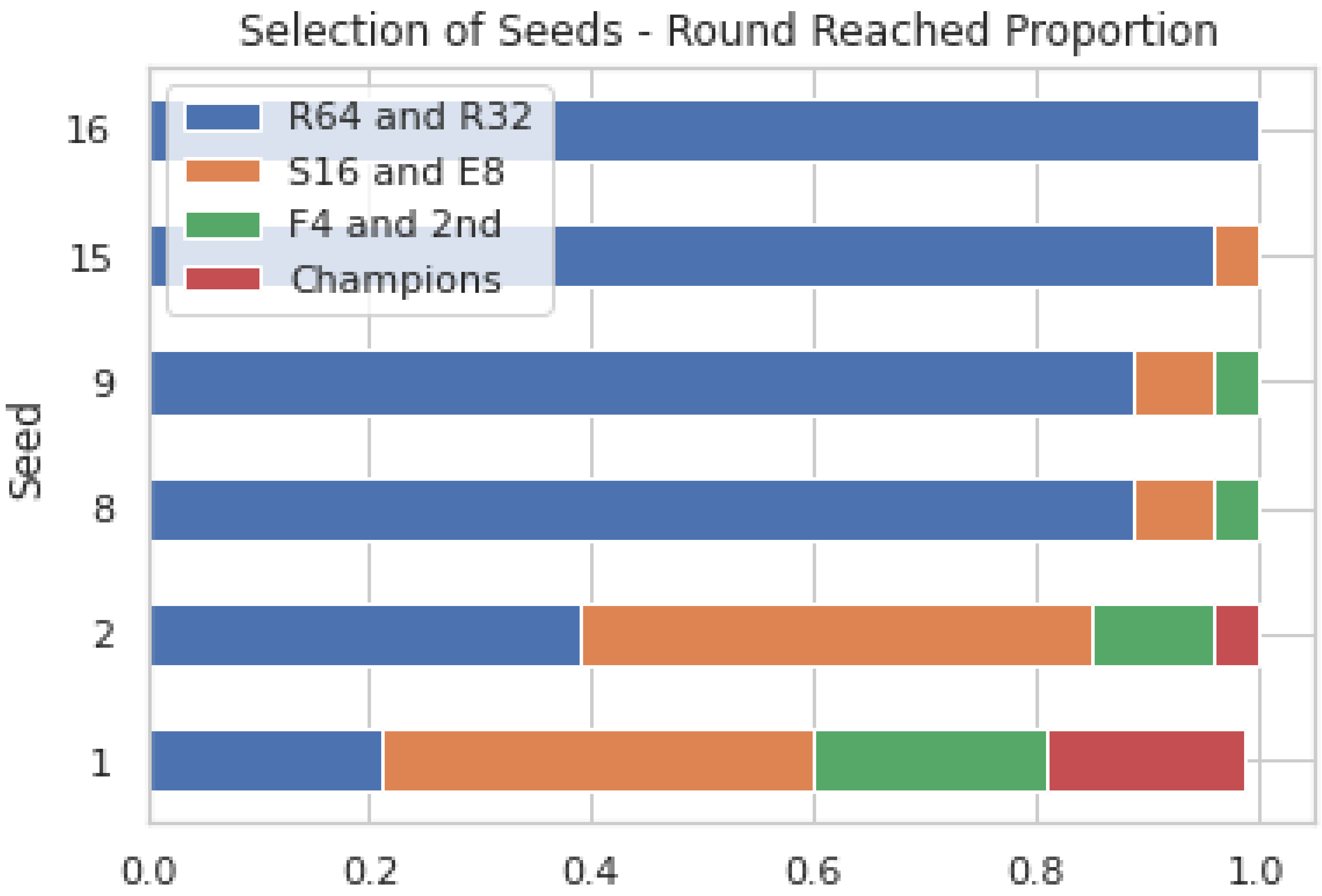
Purpose — Data — Models — Insights — Conclusions

Initial Visualizations

Highest seeds tended to have more variety in their results

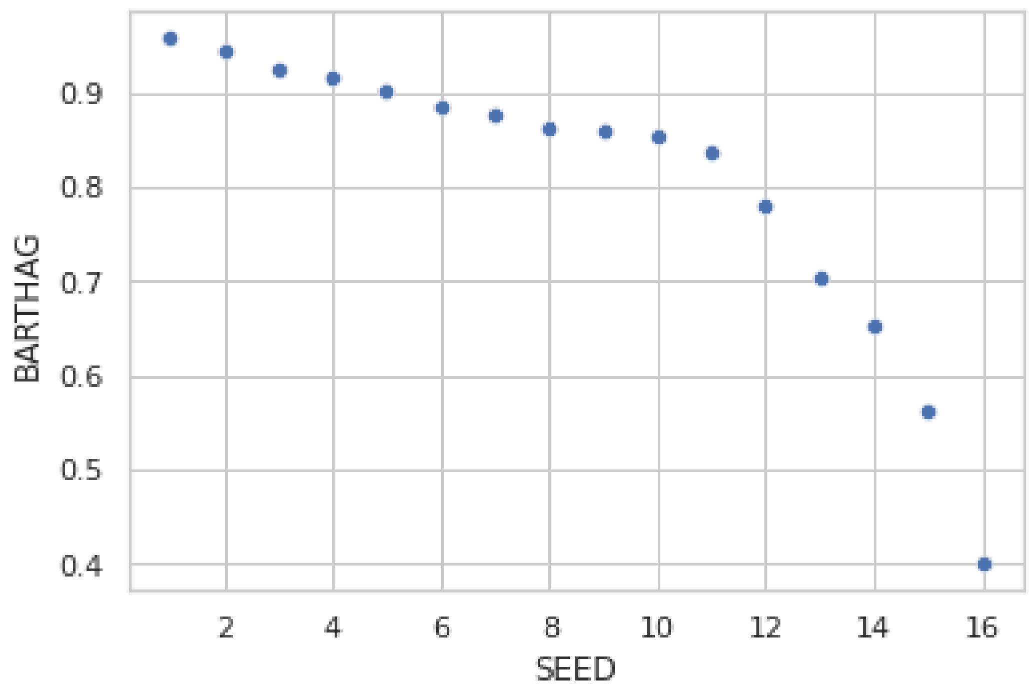
Both middle seeds surpisingly demonstrated the exact same results

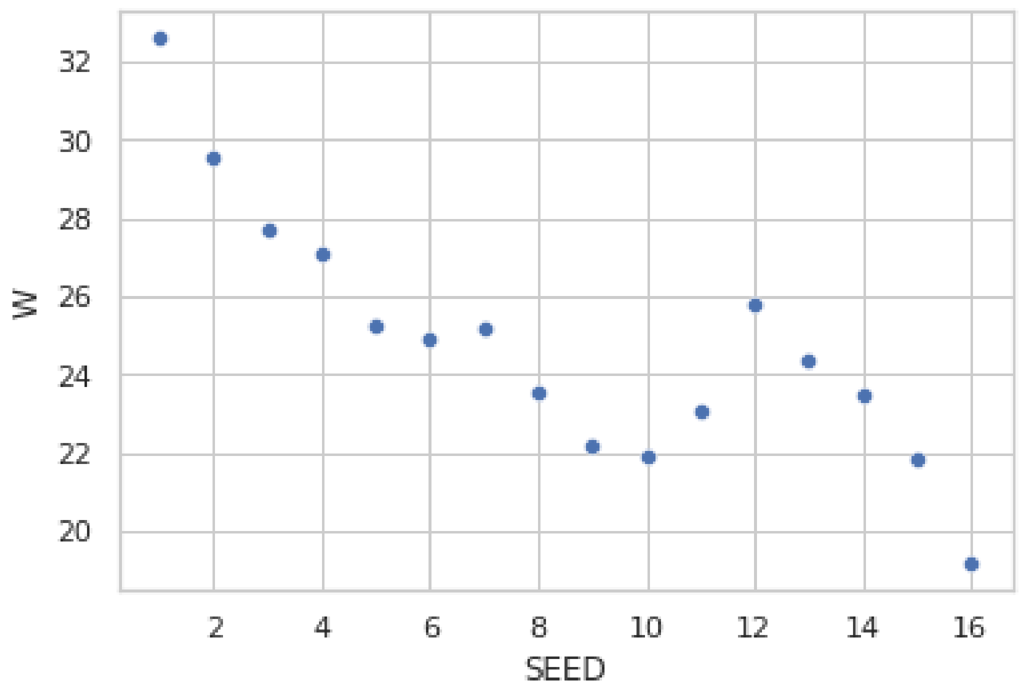
Lowest seeds displayed expected results (98% within R64 and R32)

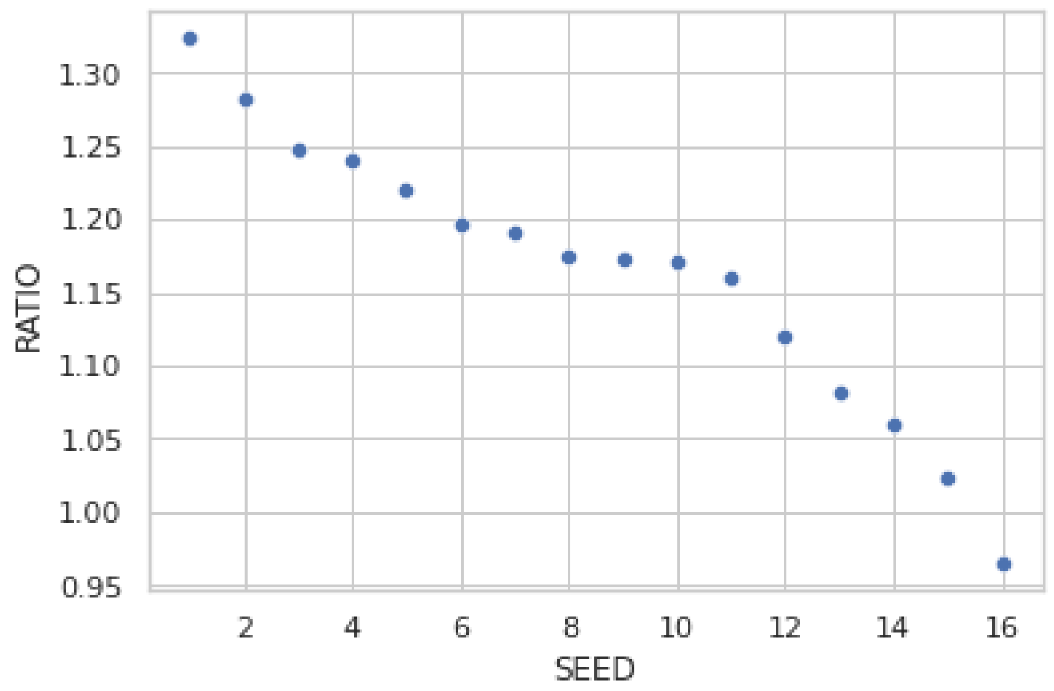


Purpose — Data — Models — Insights — Conclusions

Seed vs Top 3 Success Metrics







Purpose — Data — Models — Insights — Conclusions

Models

Models

Logistic Regression

Decision Tree Classifier

Random Forest Classifier

Why?

We are trying to predict a binary win-loss based on our 18 performance metric inputs

Logistic Regression

Considers linear relationship of a dependent
variable to one or more independent predictor variables

Decision Tree Classifier

Makes decisions that relies on conditional control statements, increasing the homogeneity after each split.

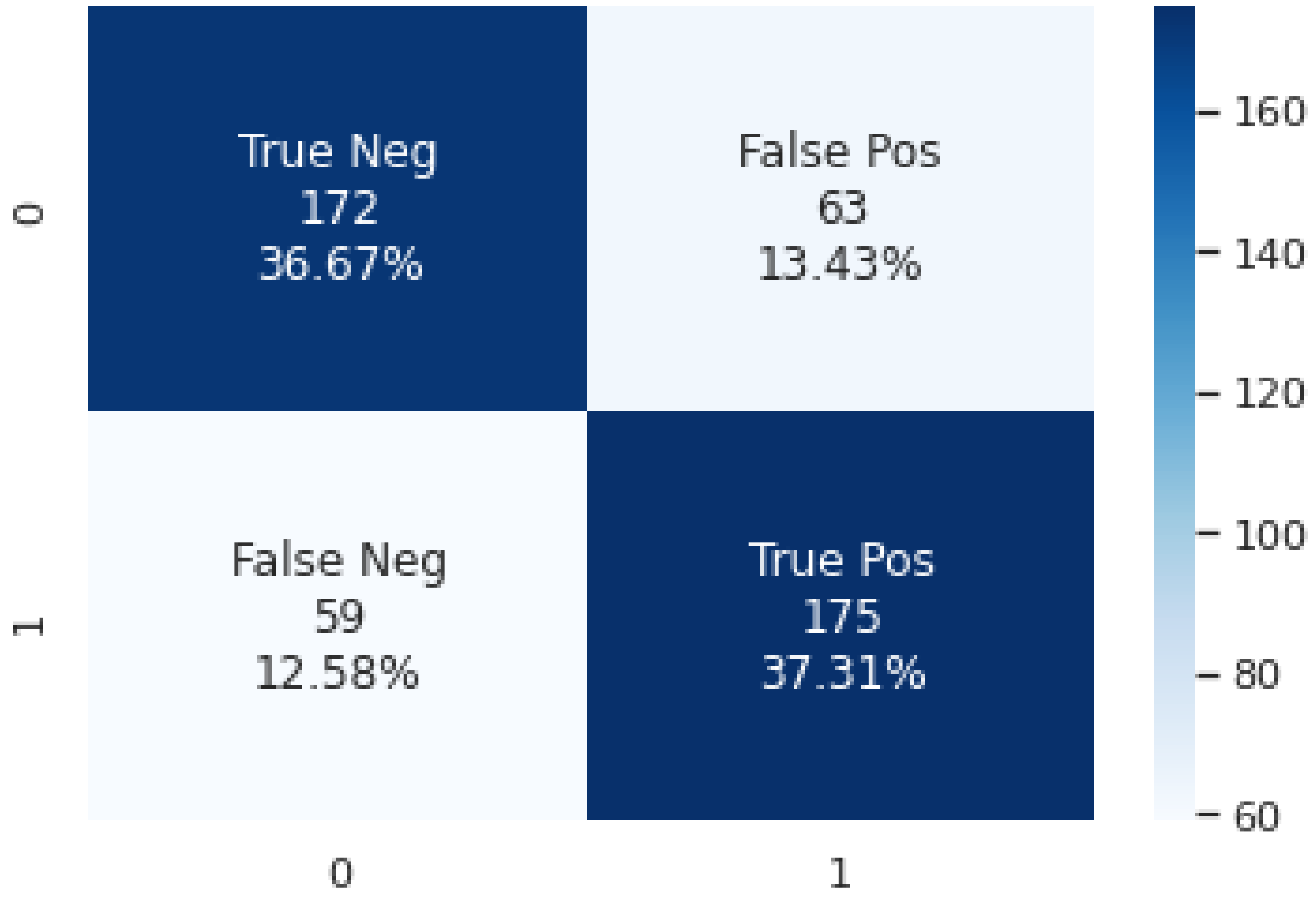
Random Forest Classifier

Constructs several decision trees training the model before outputting the most common prediction.

Logistic Regression

Accuracy: 73.9%
Precision: 73%

Insight:
For every **0.01 power rating difference** between 2 teams, the team with higher power rating's **odds of win increased by 0.08**



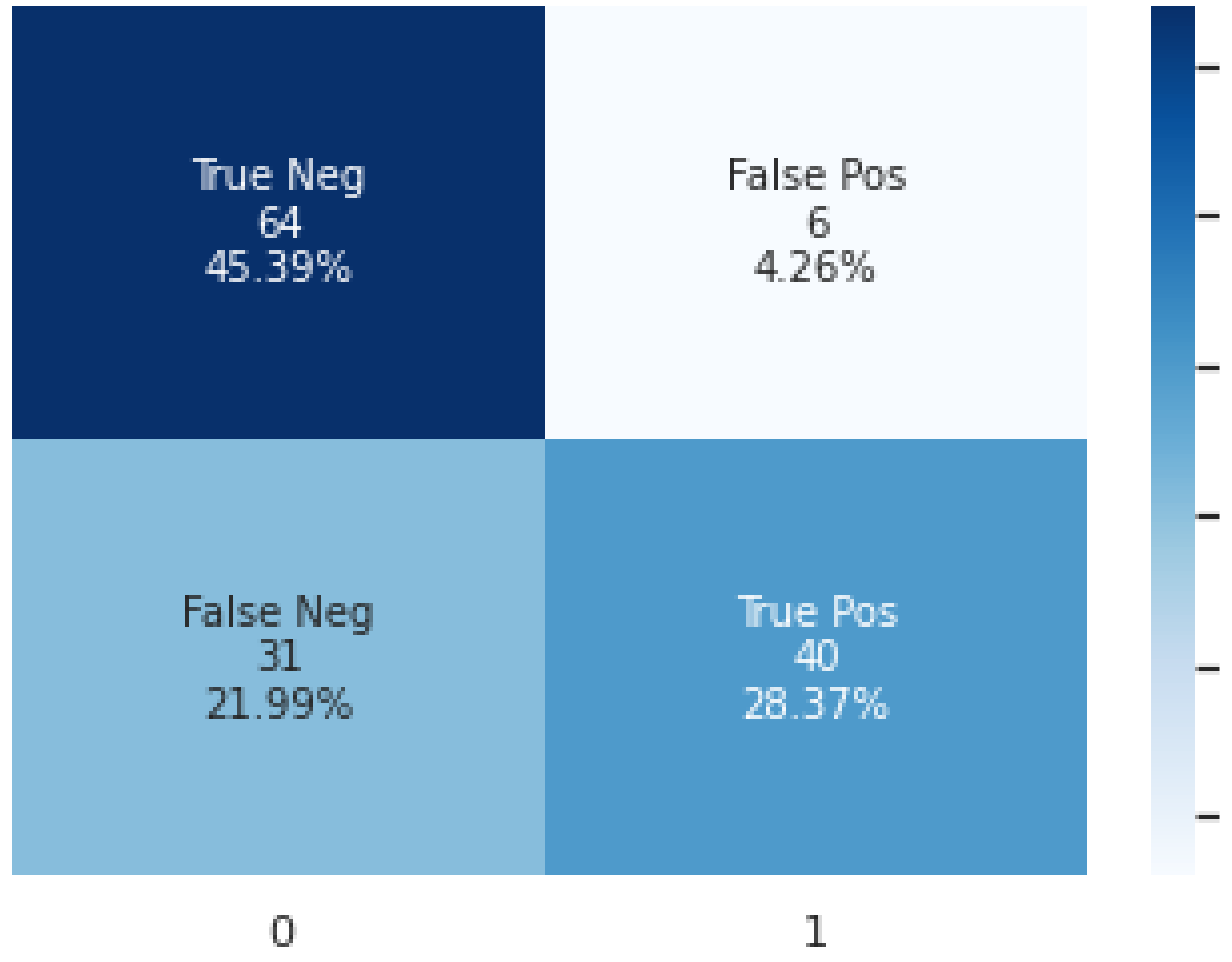
Tuning hyperparameters (Decision Tree)

	param_max_depth	param_min_samples_leaf	mean_test_score
0	5	10	0.726905
1	2	20	0.724731
2	1	20	0.724731
3	2	60	0.724731
4	2	50	0.724731
5	2	40	0.724731
6	2	30	0.724731
7	1	10	0.724731
8	2	10	0.724731
9	1	50	0.724731
10	1	40	0.724731
11	1	30	0.724731
12	1	60	0.724731
13	20	20	0.720407
14	30	20	0.720407
15	4	30	0.714162
16	3	10	0.711828
17	10	20	0.711805
18	15	20	0.709769
19	25	20	0.709769

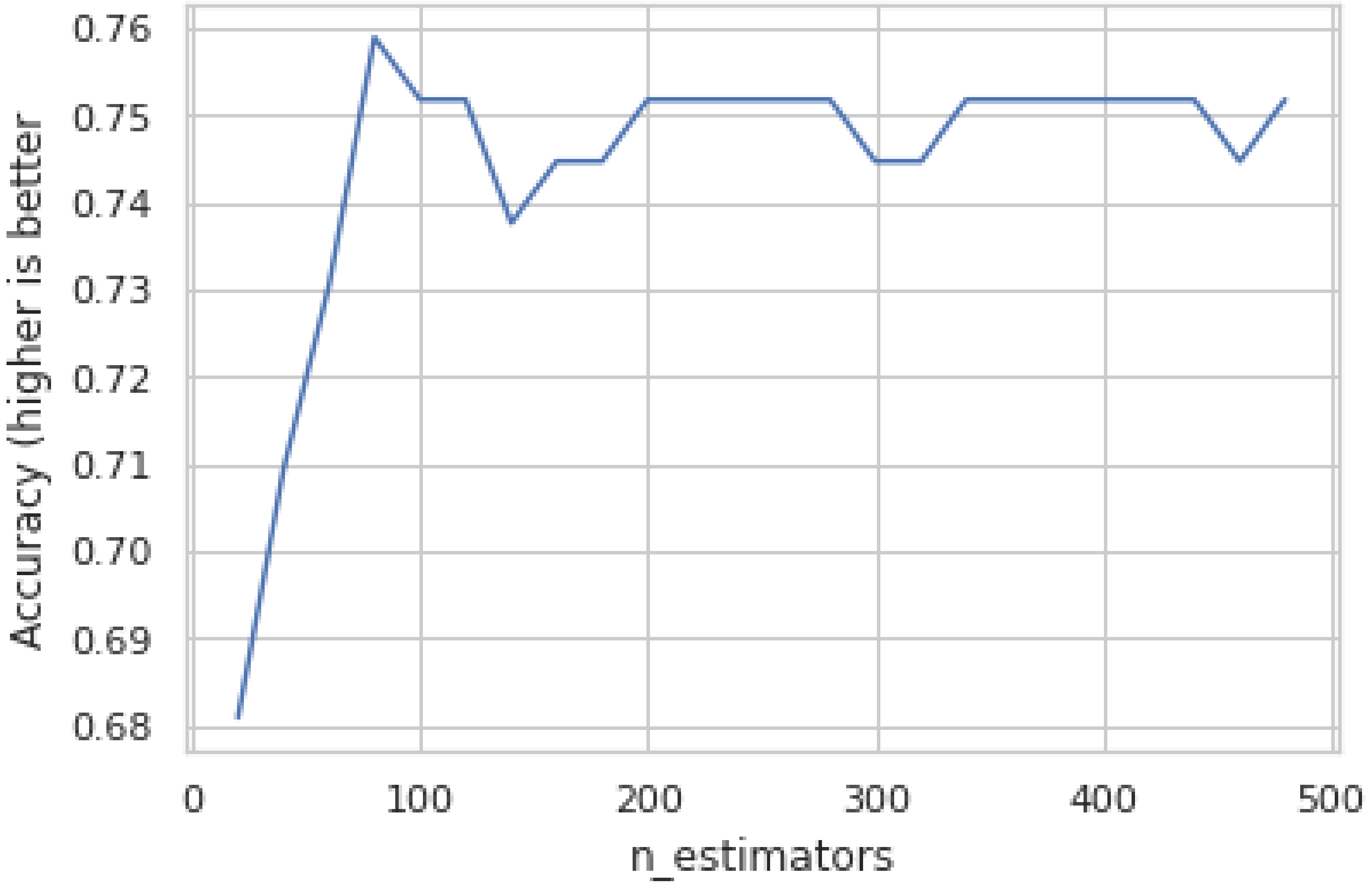
Decision Tree

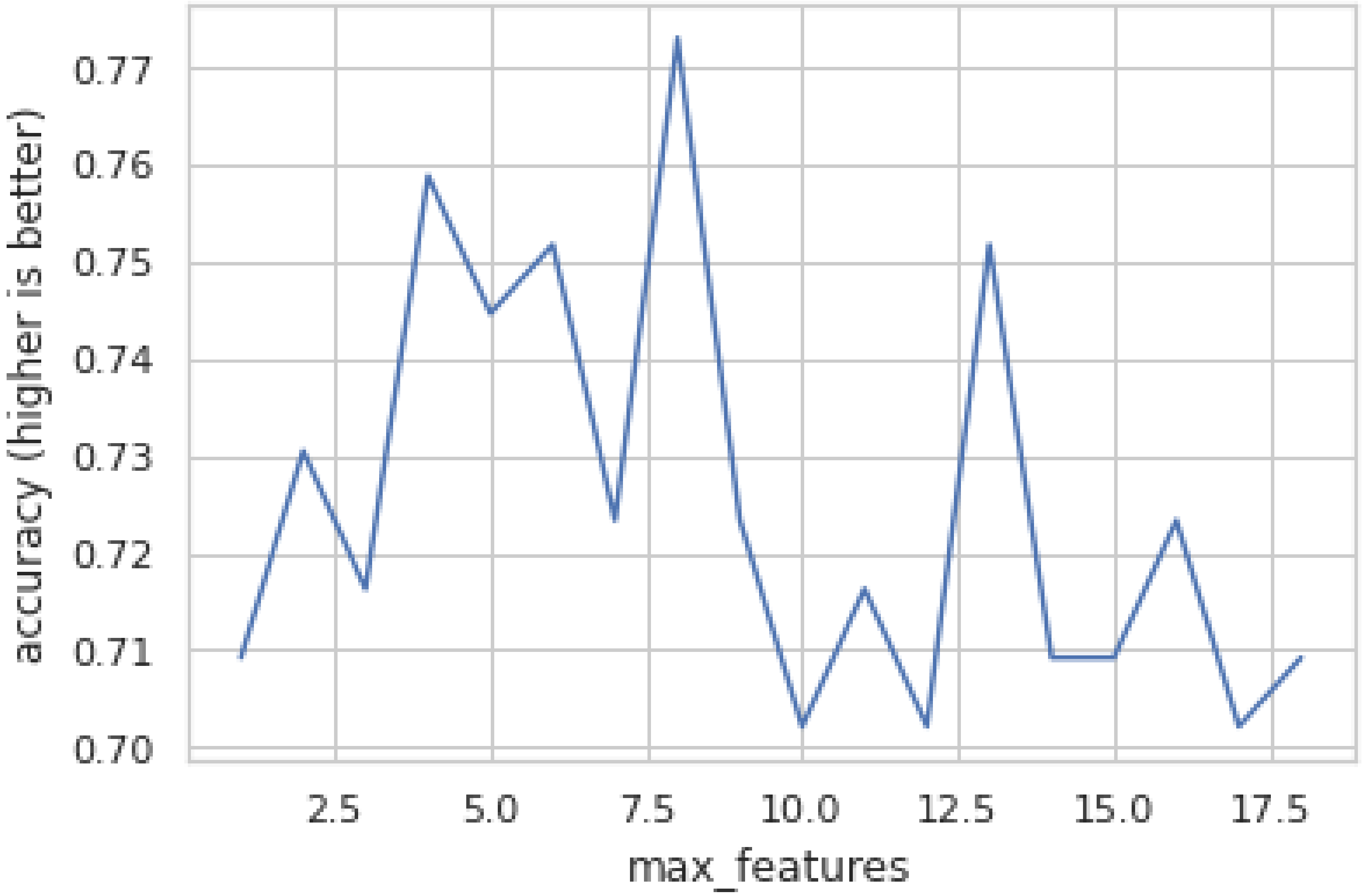
Accuracy		Null Accuracy
70.9%	>	47.5%

	Precision	Recall
0	66%	79%
1	77%	64%



Tuning hyperparameters (Random Forest)



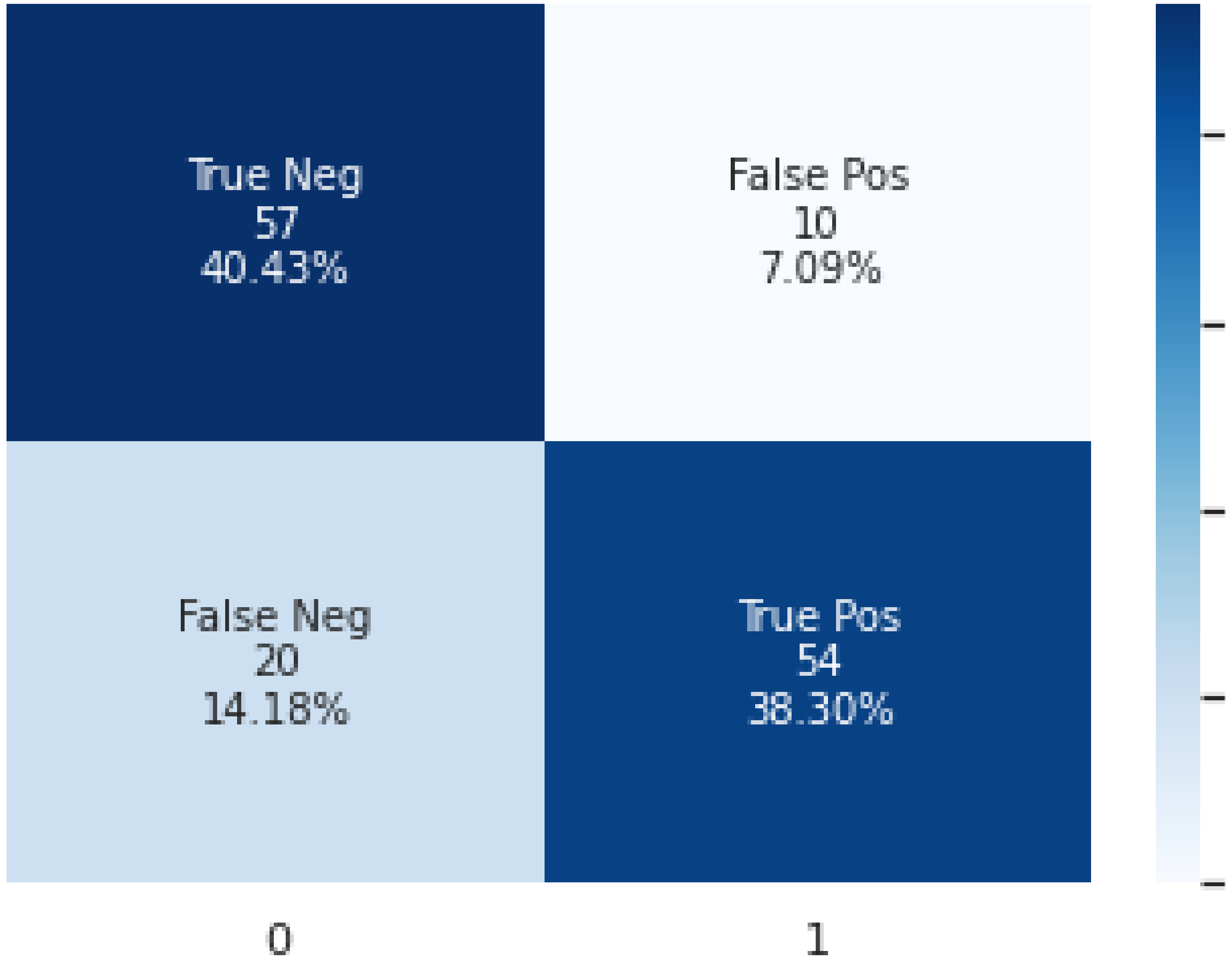


Purpose — Data — Models — Insights — Conclusions

Random Forest

Accuracy		Null Accuracy
78.7%	>	47.5%

	Precision	Recall
0	74%	85%
1	84%	73%



Model Evaluation

	Accuracy	Precision
Logistics Regression	73.9%	73%
Decision Tree Regression	70.9%	65%
Random Forest	78.7%	73.5%

Insights

GINI feature importance

Decision Tree

	feature	importance
2	Difference BARTHAG	0.423009
0	Difference ADJOE	0.187965
16	Difference WAB	0.116003
1	Difference ADJDE	0.067764
6	Difference TORD	0.062043
9	Difference FTR	0.052279
11	Difference 2P_O	0.036528
15	Difference ADJ_T	0.029585
13	Difference 3P_O	0.024824
5	Difference TOR	0.000000
7	Difference ORB	0.000000
8	Difference DRB	0.000000
4	Difference EFG_D	0.000000
10	Difference FTRD	0.000000
12	Difference 2P_D	0.000000
14	Difference 3P_D	0.000000
3	Difference EFG_O	0.000000
17	Difference SEED	0.000000

Random Forest

	feature	importance
2	Difference BARTHAG	0.186312
16	Difference WAB	0.100449
0	Difference ADJOE	0.086989
1	Difference ADJDE	0.074226
17	Difference SEED	0.048965
6	Difference TORD	0.048233
8	Difference DRB	0.044696
13	Difference 3P_O	0.043818
9	Difference FTR	0.040268
11	Difference 2P_O	0.039649
15	Difference ADJ_T	0.038781
4	Difference EFG_D	0.038411
7	Difference ORB	0.037941
10	Difference FTRD	0.036579
5	Difference TOR	0.036381
14	Difference 3P_D	0.035812
12	Difference 2P_D	0.033659
3	Difference EFG_O	0.028832

Calculated by Gini Feature Importance

1. Difference Barthag
2. Difference WAB
3. Difference ADJOE

Model Insights

1. Barthag power rating is the most important feature in both models.

2. Generally, the more advanced metrics (which encompass multiple individual metrics) are more important.

2. The best performing model was the Random Forest Model.

Concluding Insights

1

From our intial visualizations, we learned that:

1. BARTHAG power rating was the best feature to use when it came to probability of winning
2. Seed was not a good measurement of the number of wins a team would achieve, hence why upsets can occur

2

From our models, we learned that though there were obvious performance metrics that influenced a teams chance to win against another, there was still room for error given the complex factors that influence outcomes of games

Conclusions

Why not more accurate?

Shouldn’t a team with better performance metrics overall win?

Our model does not consider:

Potential player injuries or absences

If the teams are playing Home or Away

Play style and strategies favoured by different teams

Next Steps

01 Can we utilize these most useful stats to create a model-generated 2022 March Madness bracket?

02 Can we predict which stats are most important in predicting other sports terms such as “upsets” or “hot streaks”?

03 Can we apply these results and stats to predicting professional basketball games?

Thank you!

Questions?

Appendicies