Factors of Dominance in March Madness

Data Overview

What it contains, variables used and more

General Information

What's Inside?

- Our dataset contains information about Division 1 college basketball from 2013 to 2023 but has been updated to this year's tournament.
- Variables range from team information (name, conference, seed and year) to turnover rates, power rankings and many offensive and defensive stats

What was our purpose?

- Our goal is to see which of the variables are more prominent in championship level teams?
- What can we do to predict possible outcomes of a teams postseason position with 10 years of data?

Important Data

What we used

- Postseason: Position a team finished in March Madness
- ADJOE/ADJDE: Adjusted offensive and defensive efficiency (points made/allowed per 100 possession)
- BARTHAG: Power Rating (Chance of beating an average Division I team)
- EFG_O: Effective Field Goal Percentage Shot
- 2P_O/3P_O: Two point and three point shooting percentage
- TOR: Turnover rate

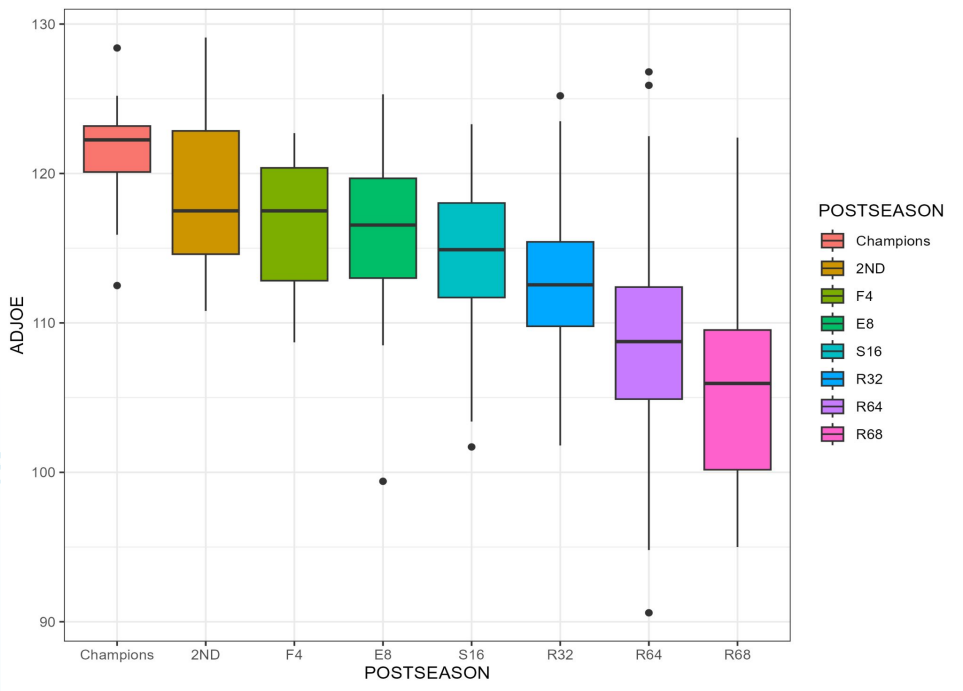
Created Variables

- Power 5: Whether or not a team is in a power five conference
- Win Percentage

Variable Analysis

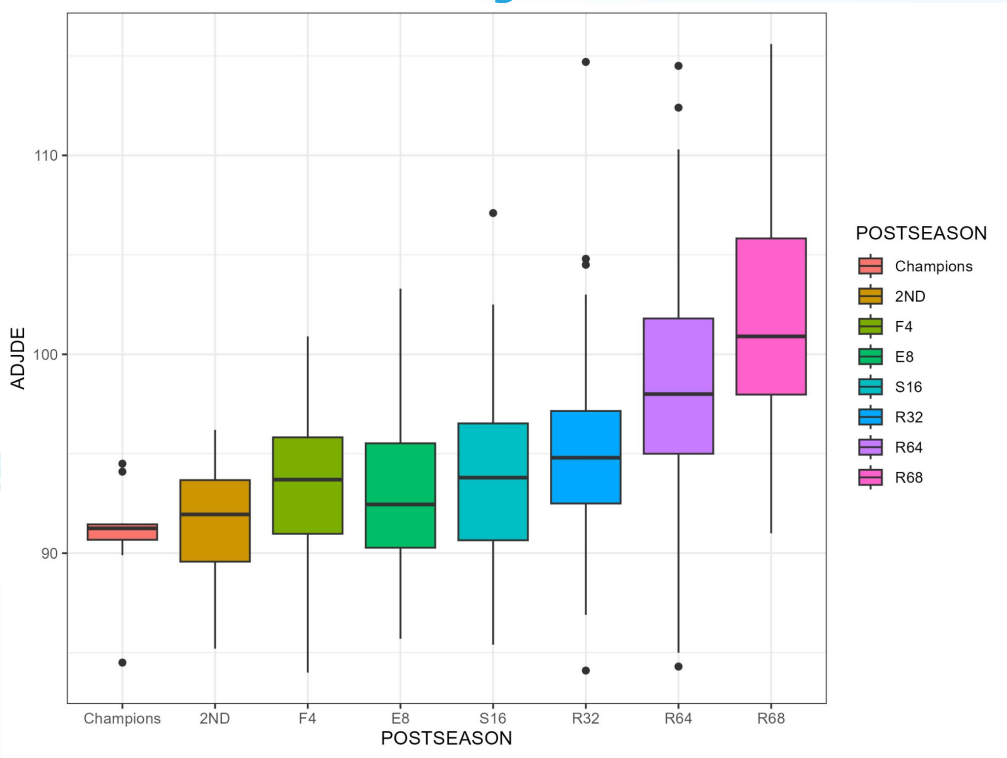
Adjusted Offensive Efficiency

- ADJOE was the statistic with the highest correlation with championship teams.
- These teams had a median ADJOE of around 120
- This median was higher than the 75th percentile of all other teams in all other rounds.



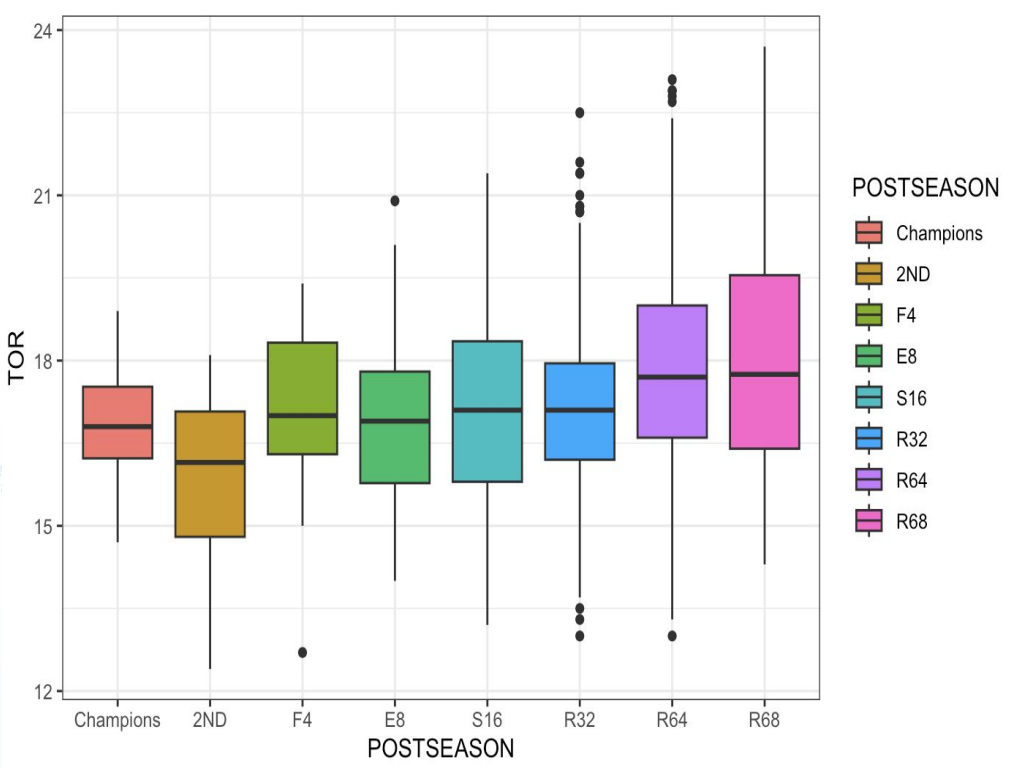
Adjusted Defensive Efficiency

- Since ADJDE is a defensive stat, the lower the better.
- The Defensive Efficiency of championship teams while superior to teams in other rounds, was less significant than Offensive.
- The median of the championship teams is 94.



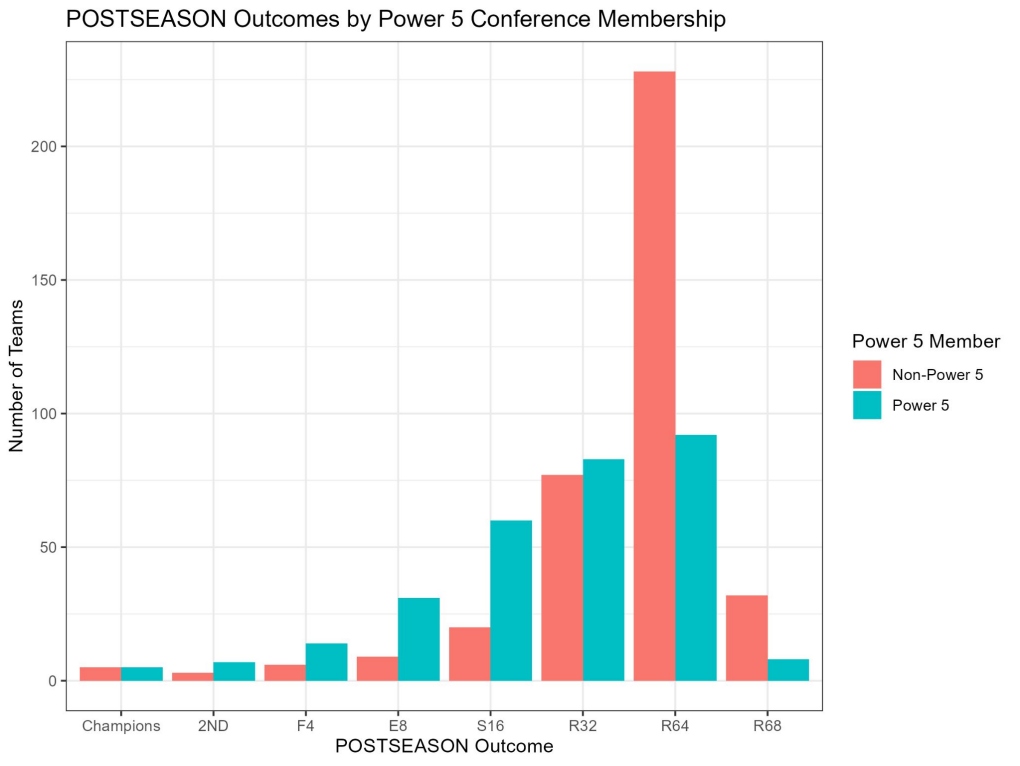
Turnover Rate

- The lower the turnover rate the better
- Turnover rate seems to have somewhat of a correlation with the teams that perform well
- The championship and 2nd place teams have a median of around 16%



Power 5 Conferences

- A power 5 conference is the 5 most influential and known conferences in college sports
- Do they perform better than Non-power 5 teams?
- Power 5 conferences have a higher percentage of teams move onto further phases of the tournament.



Decision Tree

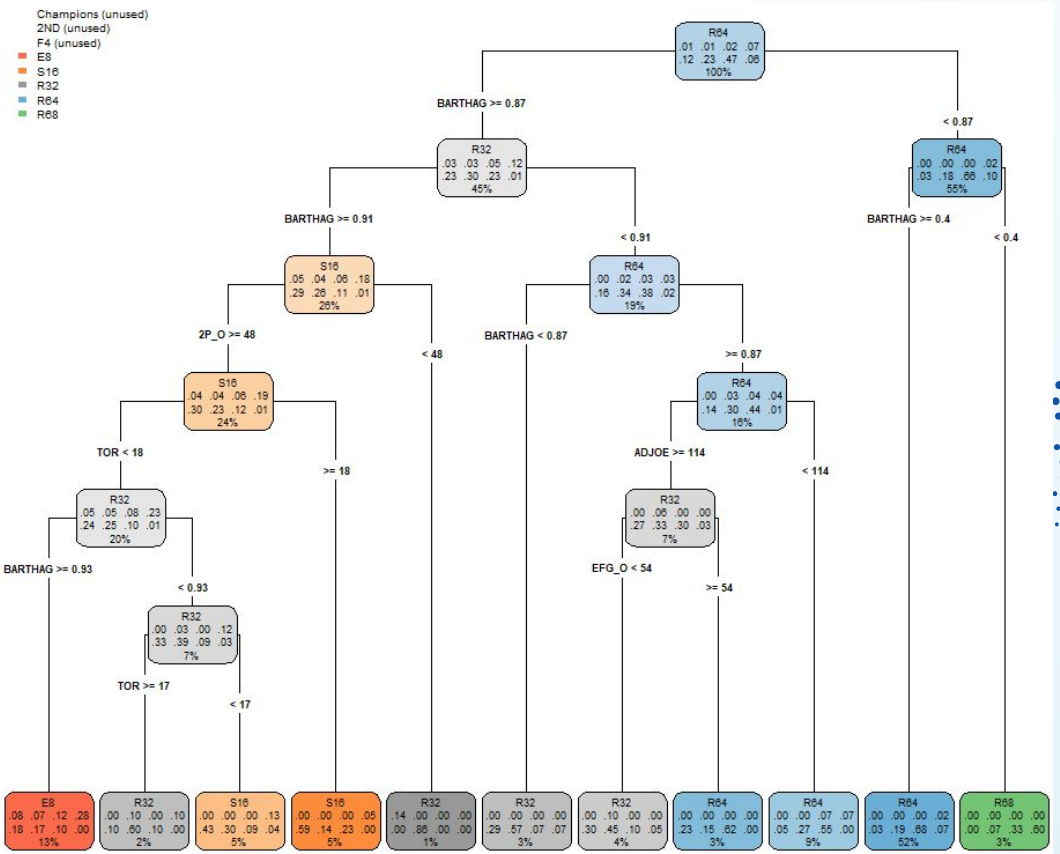
And prediction

Decision Tree

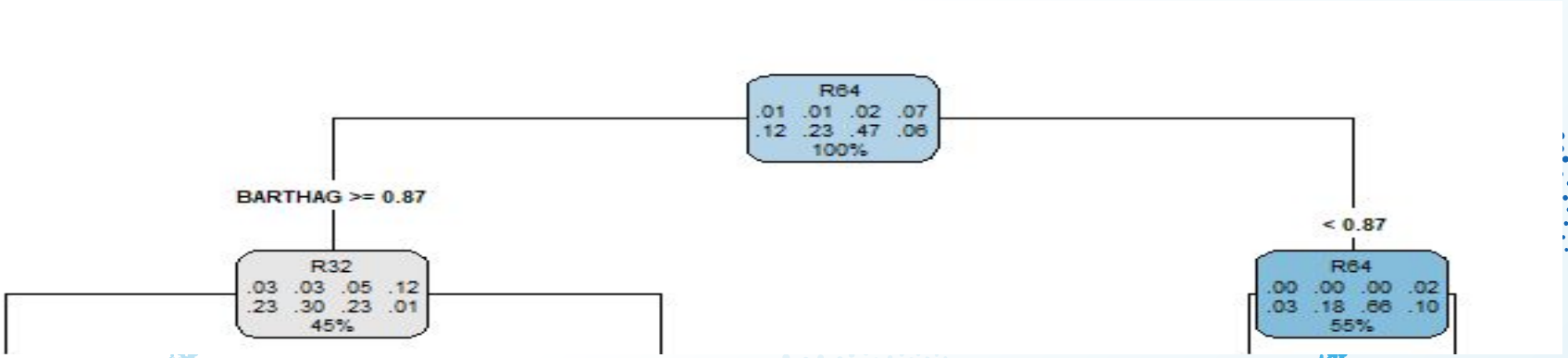
A predictive model, working similar to a flow chart to make decisions

Used to predict a category during classification

Predicting how far a team will go during a postseason tournament, i.e., March Madness

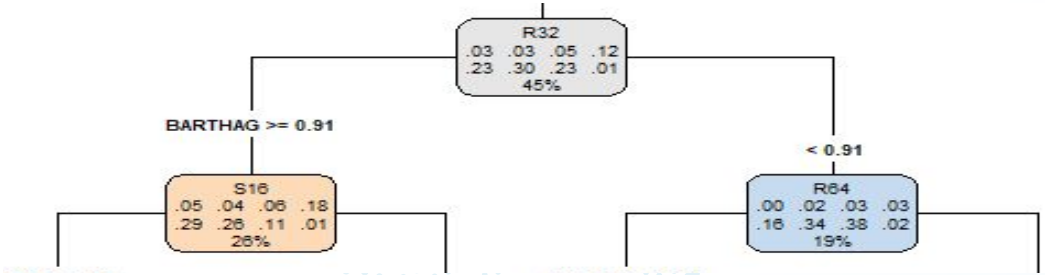


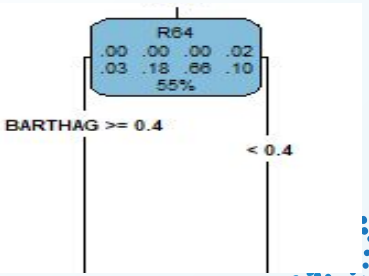
Split 1



Root Split: Overall Team Strength

Split 2

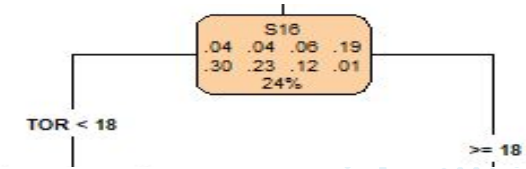




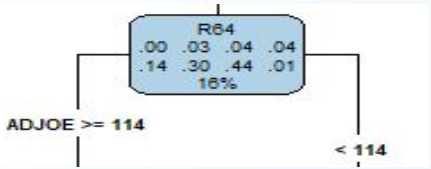
- Split 2: Elite vs. Strong Teams

Final Split: Weakest Teams

Split 3

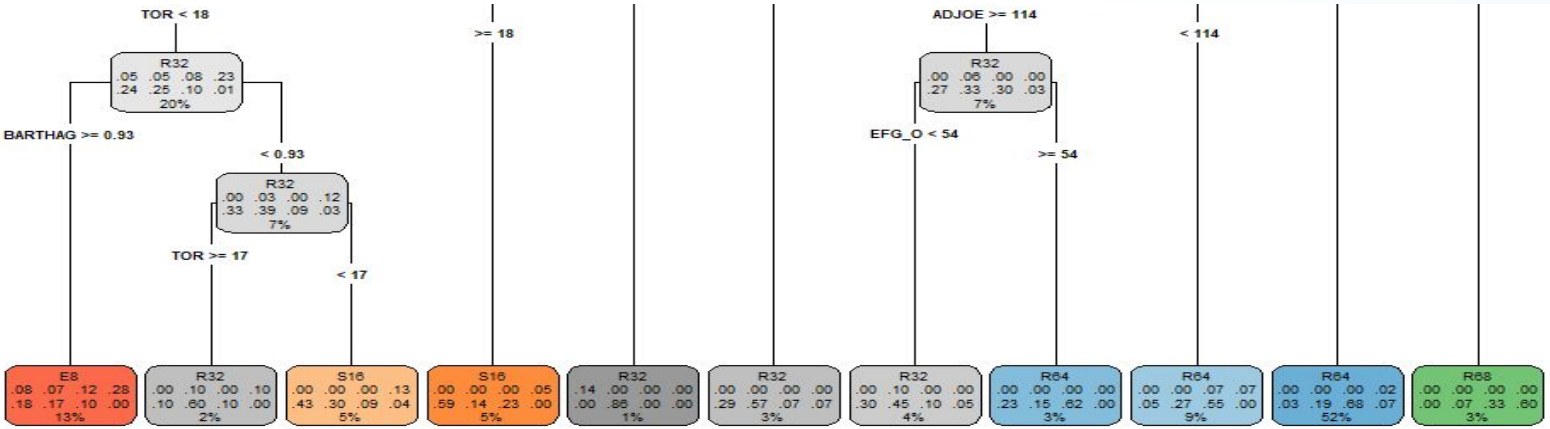


Split 4: Ball Control



Split 5: Offensive Firepower

Last Splits



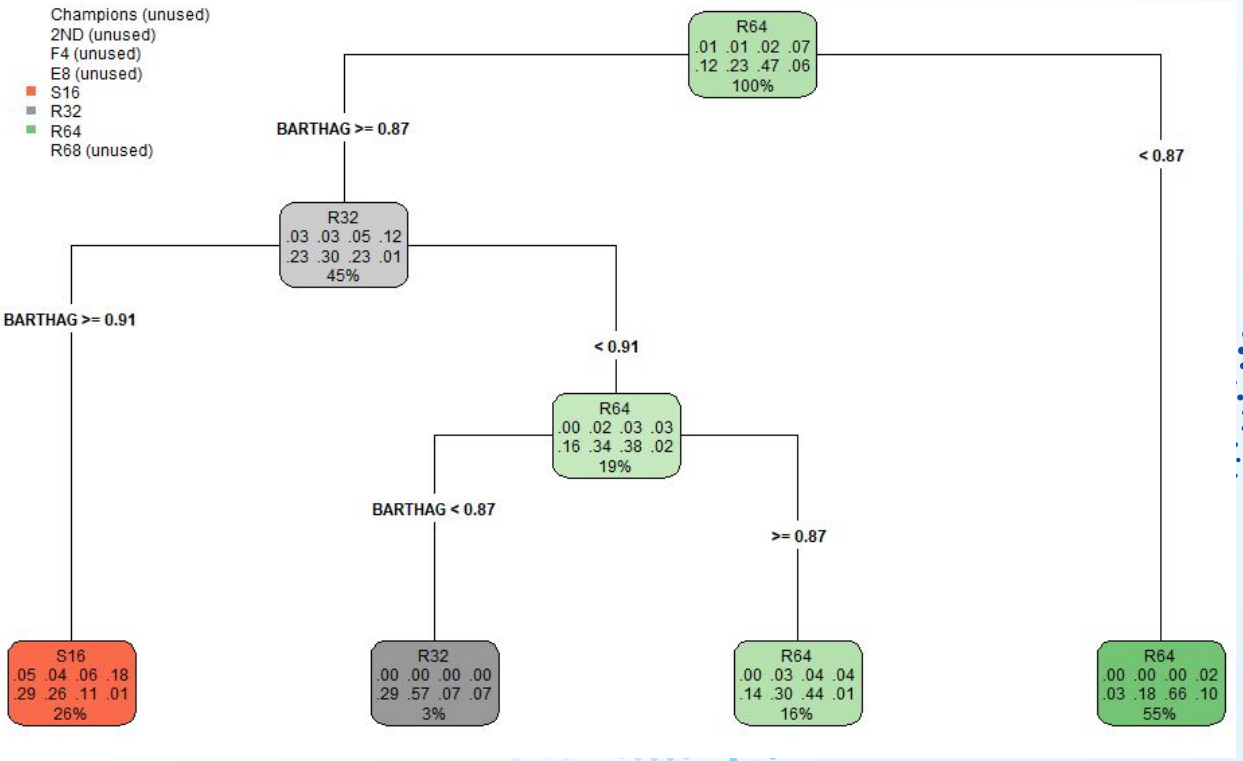
The decision tree uses a mix of team strength, offensive efficiency, and turnover control to predict how far a team will go in the tournament.

It doesn't use all POSTSEASON levels because some are too rare or hard to separate from others based on the data.

Pruned Decision Tree

We created a “pruned” or a simplified tree to reduce overfitting.

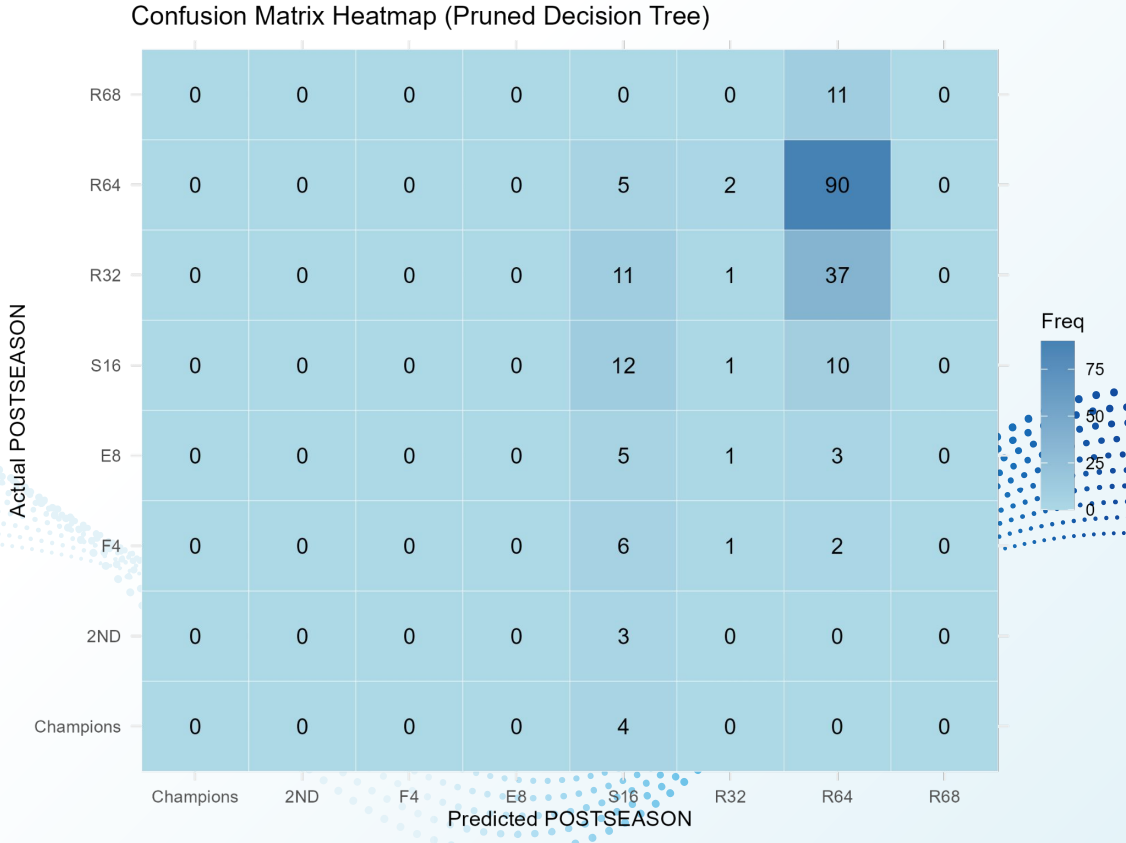
Cuts out branches that are not making significant improvements



Prediction

Using a 70/30 split for training and testing data, our model was able to predict roughly 50% with the values we used.

Based on our pruned tree

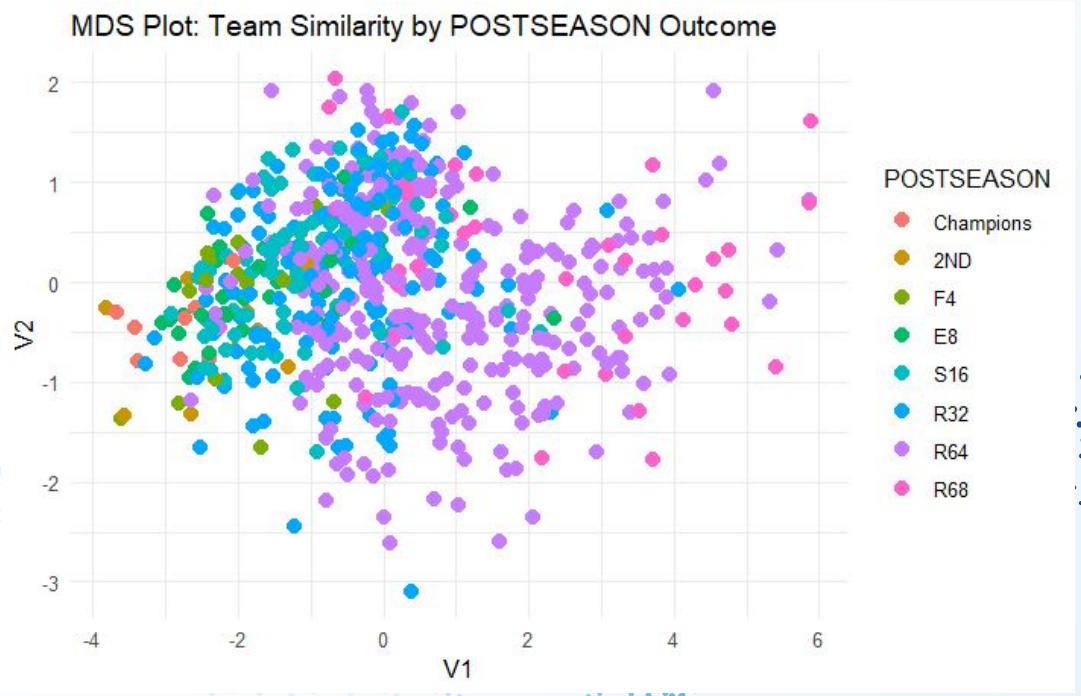


Wrap-Up

Team Similarity

Using offensive efficiency, power rating, win percentage and wins above bubble, this mds plot shows how similar teams are to each other.

Color corresponds to their postseason result



Conclusion

- ADJOE correlates the most towards championship teams
- Early in the tournament outcomes are easier to predict compared to later rounds
- Teams in the power 5 conference tend to make it to the late phases of the tournament.
- With the values we used, we were only able to predict up to 50%

Questions?