Exploring the Causes of Upsets in March Madness Tournament

Cody Greene

DSO579 Final Project

07/27/2023

Agenda

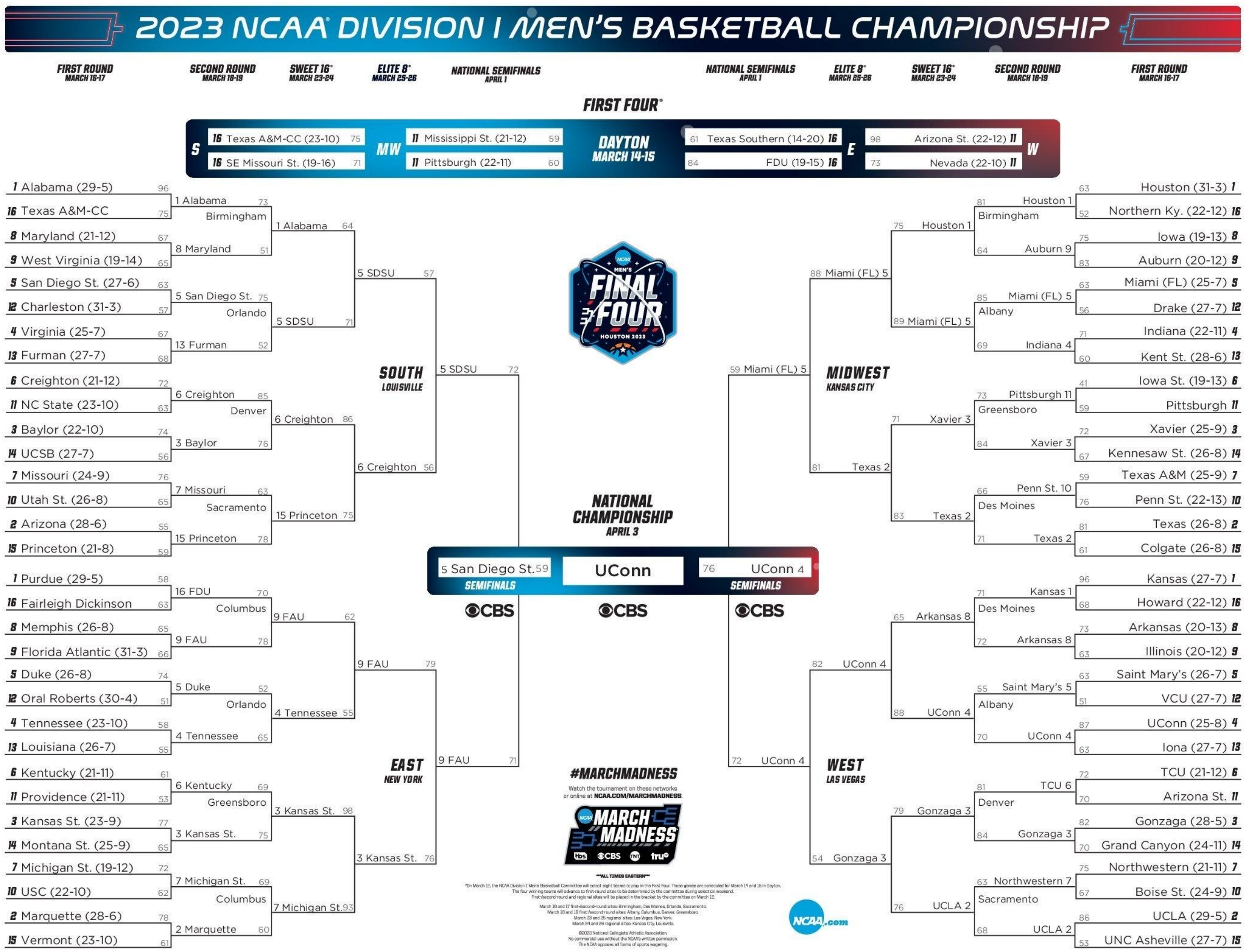
1. What is March Madness?
2. Data Processing and Cleaning
3. Defining “Upset”
4. Upset by Round
5. Feature Selection
6. Prediction Models
    - ❑ Logistic Regression
    - ❑ Random Forest
    - ❑ Gradient Boosting
7. Analysis of Important Features
8. Limitations
9. Conclusion

What is March Madness?

- March Madness is a NCAA Division I Men's Basketball Tournament held every March

- It features 68 college basketball teams competing in a single-elimination format to determine the national champion

- The tournament is known for its unpredictability and intense excitement as underdog teams often pull off upsets, making it one of the country's most popular and widely watched sporting events



ALL TIMES EASTERN

*On March 12, the NCAA Division I Men's Basketball Committee will select eight teams to play in the First Four. Those games are scheduled for March 14 and 15 in Dayton. The four winning teams will advance to first-round sites to be determined by the committee during selection weekend. First-/second-round and regional sites will be placed in the bracket by the committee on March 12.

March 16 and 17 first-/second-round sites: Birmingham, Des Moines, Orlando, Sacramento.
March 18 and 19 first-/second-round sites: Albany, Columbus, Denver, Greensboro.
March 23 and 25 regional sites: Las Vegas, New York.
March 24 and 26 regional sites: Kansas City, Louisville.

©2023 National Collegiate Athletic Association.
No commercial use without the NCAA's written permission.
The NCAA opposes all forms of sports wagering.

Source: https://www.ncaa.com/news/basketball-men/article/2023-04-18/2023-ncaa-bracket-s[illegible]mens-tournament

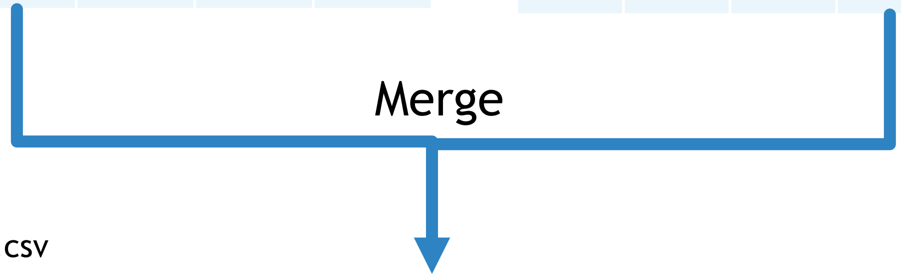
Data Processing and Cleaning

Tournament_Scores_1981-2021.csv

YEAR	ROUND	WSEED	WTEAM	LSEED	LTEAM	…
2021	6	1	Baylor	1	Gonzaga	…
2021	5	1	Gonzaga	11	UCLA	…

Tournament_Team_Data_2008-2022.csv

YEAR	SEED	TEAM	OF EFF	DF EFF	…	WIN %
2022	1	Kansas	1.17	0.85	…	0.760
2022	1	Arizona	1.16	0.86	…	0.729



Merged_Data_2008-2021.csv

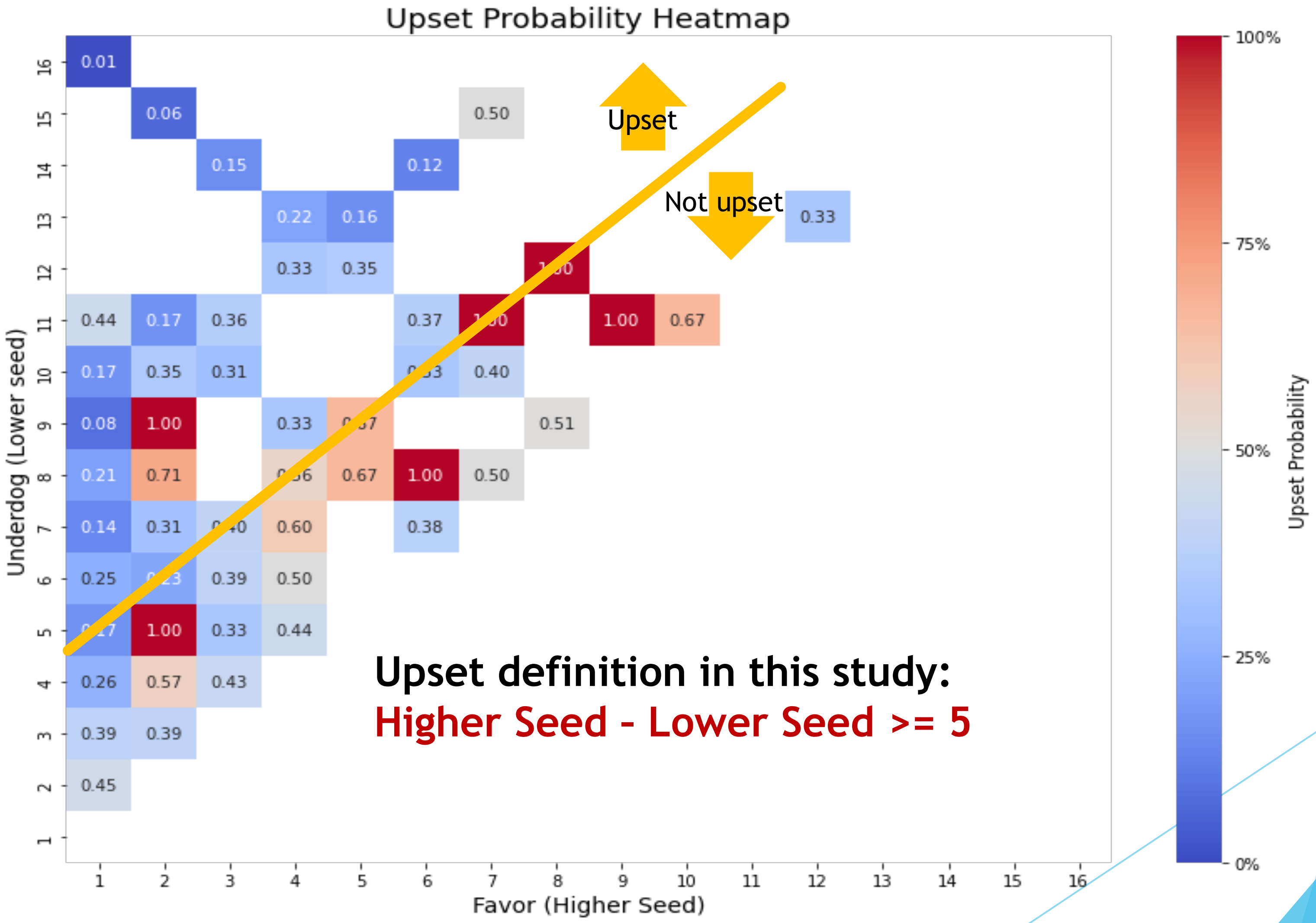
YEAR	ROUND	TEAM_F	SEED_F	OF EFF_F	DF EFF_F	…	TEAM_U	SEED_U	OF EFF_U	DF EFF_U	…	UPSET
2021	6	Baylor	1	1.12	0.771		UCLA	11	1.05	0.910	…	0
2021	5	Gonzaga	1	1.19	0.698		Houston	6	1.09	0.830	…	1

1. Download csv data from Kaggle
2. Upload data into python using pd.read_csv
3. Standardize school names between datasets
   e.x. USC -> Southern California
4. Merge datasets
   Add suffix _F for favorite team data and _U for underdog team
5. Add a binary value column UPSET
6. Drop unnecessary columns

Source: https://www.kaggle.com/datasets/nishaanamin/march-madness-data

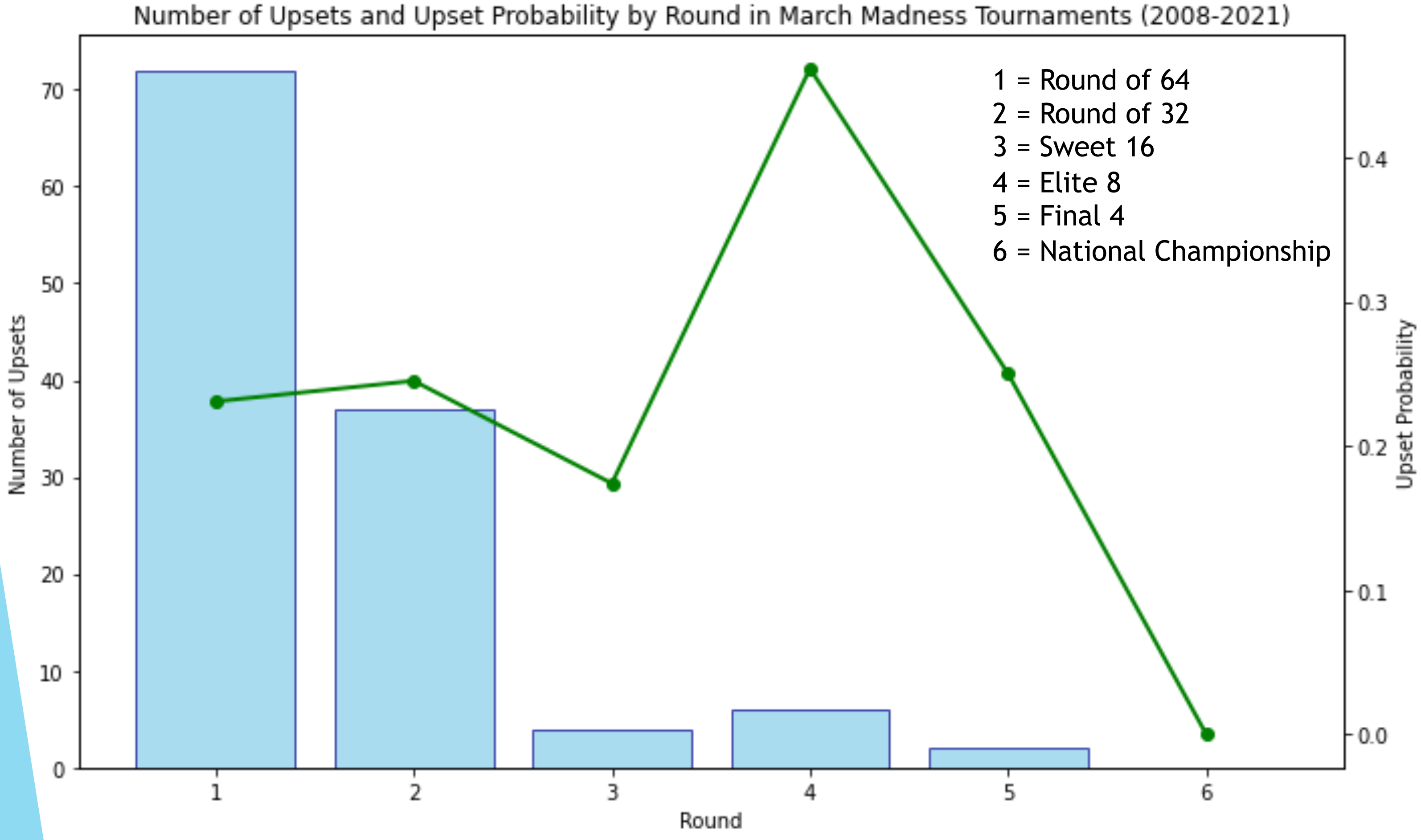
Defining "Upset"

Upset should not be a phenomenon that occurs frequently, and definitions in different studies are inconsistent. Therefore, this study is defining a new definition of upset based on historical probability.



Upset definition in this study:
Higher Seed - Lower Seed >= 5

Upset by Round

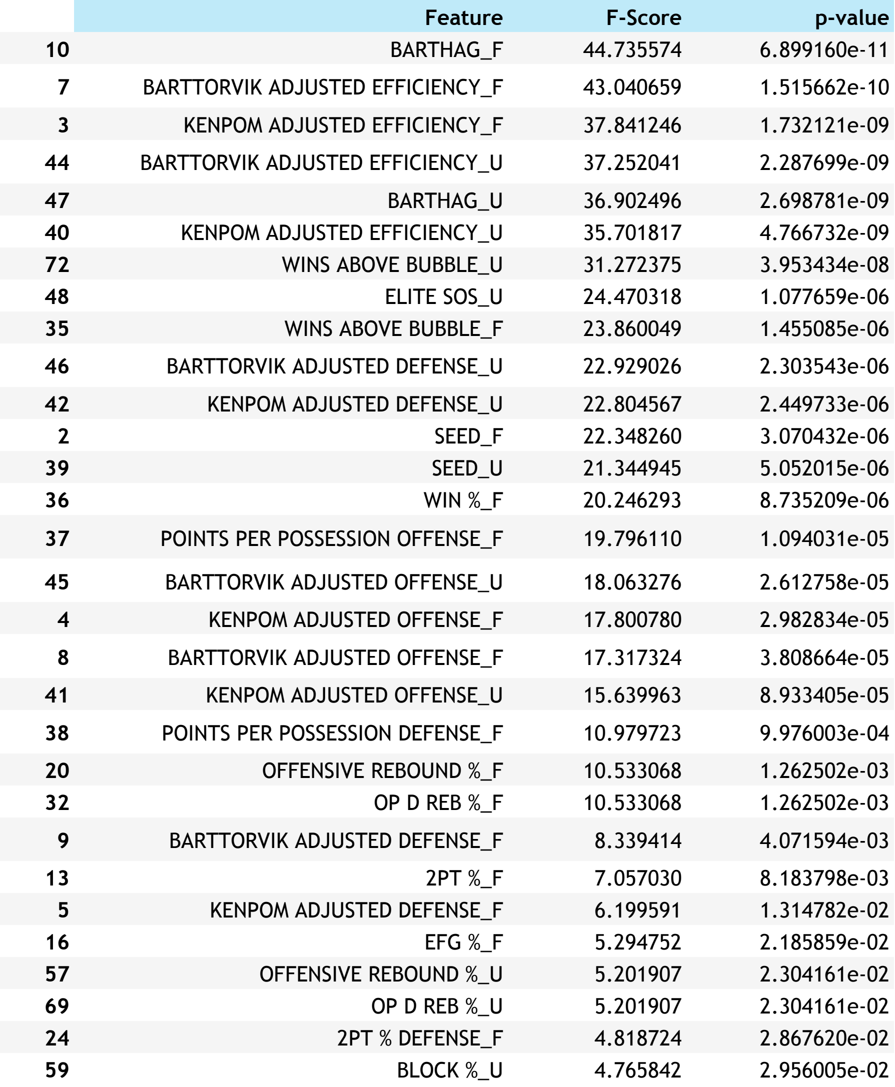


Source: https://championshiphistory.com/ncaahoops.php

- Most upsets occur in 1st round. However, provability is not as high
- Upset probability increases in the 4th round. This indicates that once an underdog makes it to the 3rd round, they have the momentum to advance further
- There are no upsets in the championship within the data period. The last championship upset was in 1988, Kansas (6) vs Oklahoma (1)

Feature Selection

Performed ANOVA and selected top 30 important features to train models.



	Feature	F-Score	p-value
10	BARTHAG_F	44.735574	6.899160e-11
7	BARTTORVIK ADJUSTED EFFICIENCY_F	43.040659	1.515662e-10
3	KENPOM ADJUSTED EFFICIENCY_F	37.841246	1.732121e-09
44	BARTTORVIK ADJUSTED EFFICIENCY_U	37.252041	2.287699e-09
47	BARTHAG_U	36.902496	2.698781e-09
40	KENPOM ADJUSTED EFFICIENCY_U	35.701817	4.766732e-09
72	WINS ABOVE BUBBLE_U	31.272375	3.953434e-08
48	ELITE SOS_U	24.470318	1.077659e-06
35	WINS ABOVE BUBBLE_F	23.860049	1.455085e-06
46	BARTTORVIK ADJUSTED DEFENSE_U	22.929026	2.303543e-06
42	KENPOM ADJUSTED DEFENSE_U	22.804567	2.449733e-06
2	SEED_F	22.348260	3.070432e-06
39	SEED_U	21.344945	5.052015e-06
36	WIN %_F	20.246293	8.735209e-06
37	POINTS PER POSSESSION OFFENSE_F	19.796110	1.094031e-05
45	BARTTORVIK ADJUSTED OFFENSE_U	18.063276	2.612758e-05
4	KENPOM ADJUSTED OFFENSE_F	17.800780	2.982834e-05
8	BARTTORVIK ADJUSTED OFFENSE_F	17.317324	3.808664e-05
41	KENPOM ADJUSTED OFFENSE_U	15.639963	8.933405e-05
38	POINTS PER POSSESSION DEFENSE_F	10.979723	9.976003e-04
20	OFFENSIVE REBOUND %_F	10.533068	1.262502e-03
32	OP D REB %_F	10.533068	1.262502e-03
9	BARTTORVIK ADJUSTED DEFENSE_F	8.339414	4.071594e-03
13	2PT %_F	7.057030	8.183798e-03
5	KENPOM ADJUSTED DEFENSE_F	6.199591	1.314782e-02
16	EFG %_F	5.294752	2.185859e-02
57	OFFENSIVE REBOUND %_U	5.201907	2.304161e-02
69	OP D REB %_U	5.201907	2.304161e-02
24	2PT % DEFENSE_F	4.818724	2.867620e-02
59	BLOCK %_U	4.765842	2.956005e-02

- KemPom Rankings
  - Developed by Ken Pomeroy in 1999, a meteorologist working for the National Weather Service in Montana
  - The KenPom system is used as a reference in college basketball by bettors, bookmakers and coaches. It’s very accurate and offers precise predictions most of the time
  - Website: https://kenpom.com/

- BartTorvik Rankings
  - Run by Bart Torvik who is a lawyer
  - It was developed by reverse engineering KemPom and it has a similar usage
  - There is a R package called ‘toRvik’
  - Website: https://barttorvik.com/#

Logistic Regression Model

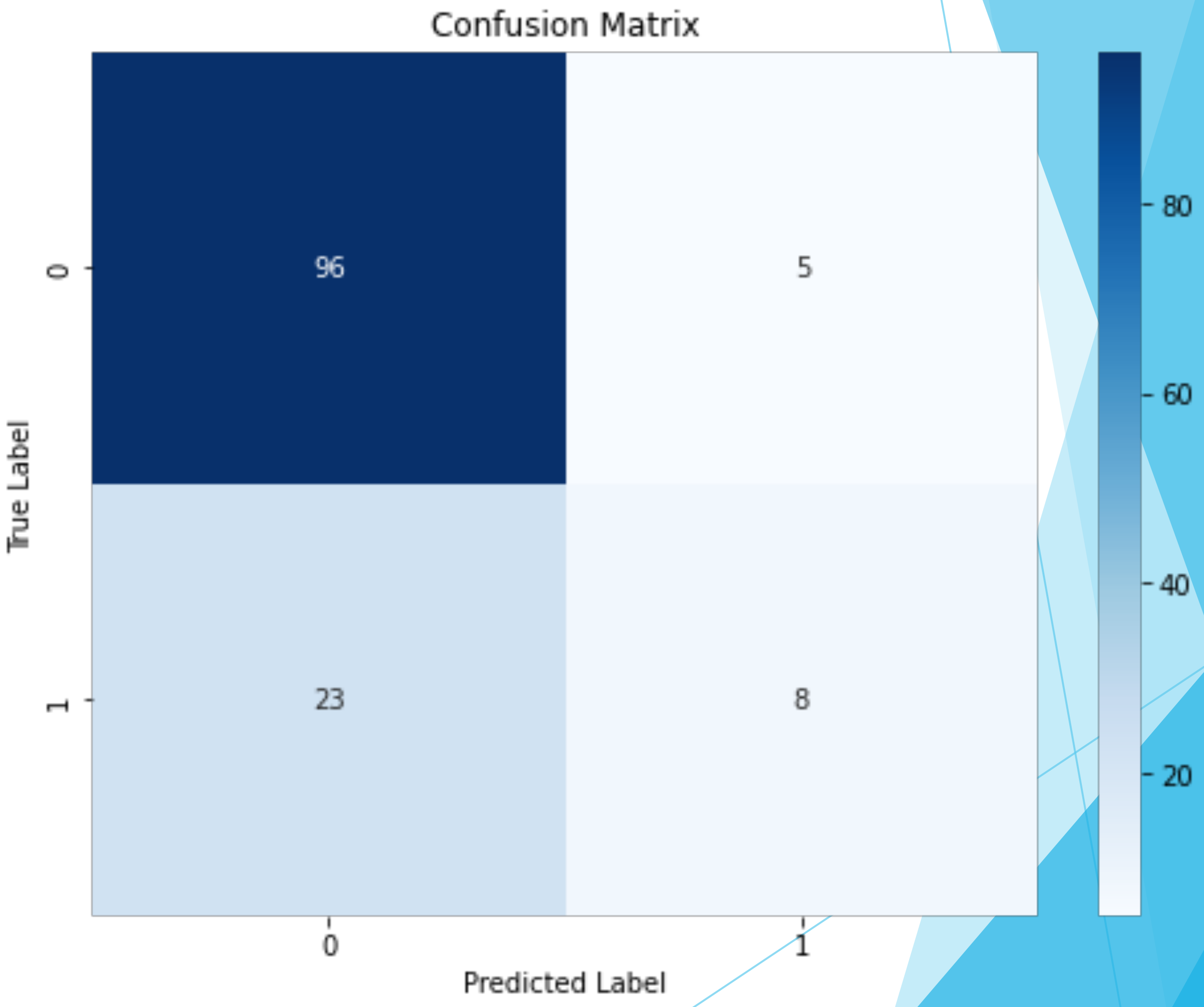
Reason for choosing this model

Logistic regression model is a simple and widely-used linear classification algorithm that works well for binary classification like predicting the outcome of a sports game

Cross-validation results

[0.81818182, 0.79545455, 0.72727273 0.79545455, 0.75, 0.79545455, 0.72727273, 0.68181818, 0.79545455, 0.77272727]

Average cv accuracy: 0.76590



Random Forest Model

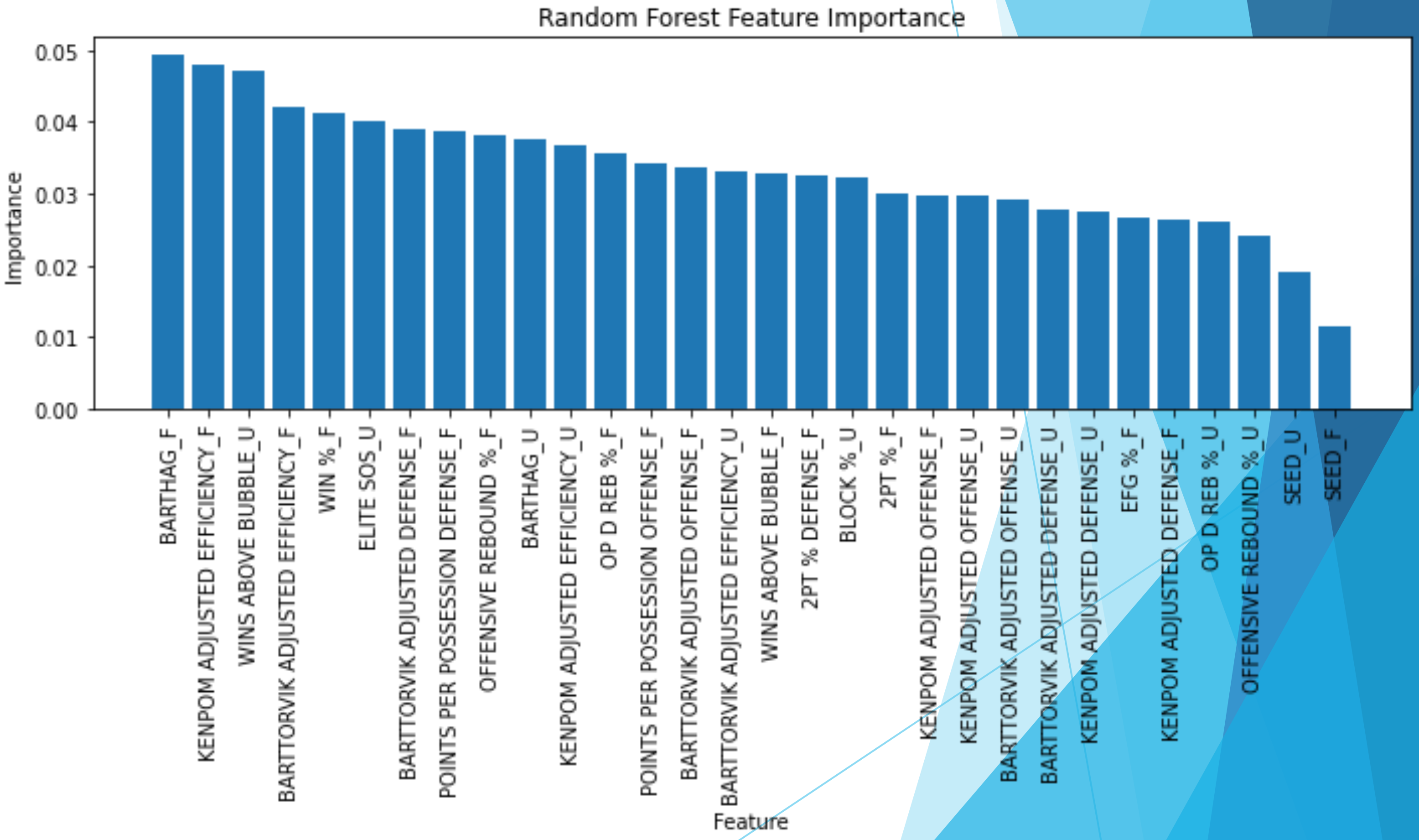
Reason for choosing this model

Random forest model is an ensemble model which can provide high accuracy and reduce overfitting

Cross-validation results

[0.75 0.77272727 0.75 0.77272727 0.77272727 0.75 0.68181818 0.70454545 0.77272727 0.72727273]

Average cv accuracy: 0.74545



Gradient Boosing Model (GBM)

Reason for choosing this model

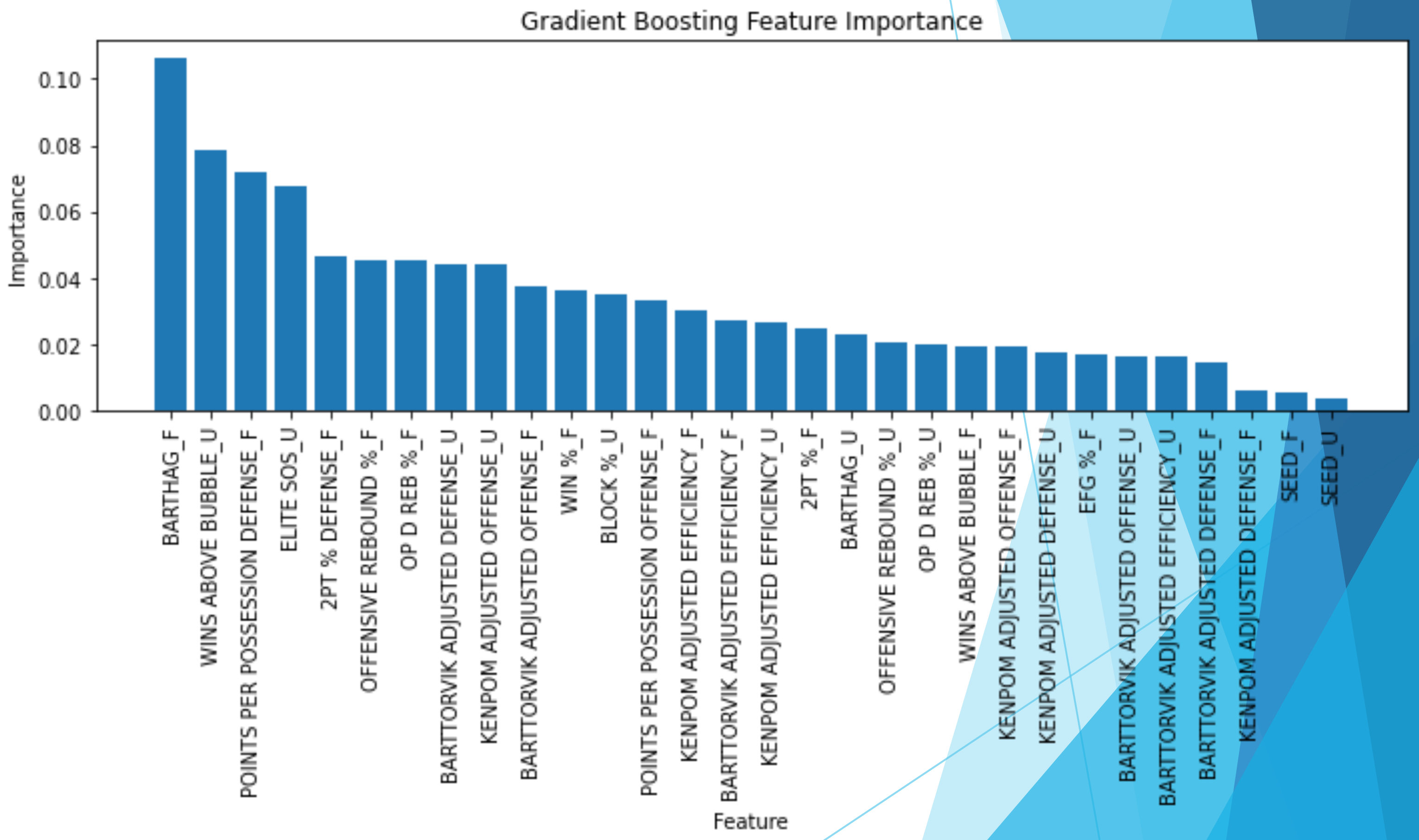
A model using GBM won the *NCAAW March Madness Kaggle Competition* in 2022. Therefore, this model can be applied to predict Men's Tournament

Source: https://towardsdatascience.com/kaggle-march-madness-silver-medal-for-two-consecutive-years-6207ff63b86c

Cross-validation results

[0.81818182, 0.77272727, 0.79545455, 0.72727273, 0.68181818, 0.77272727, 0.65909091, 0.70454545, 0.79545455, 0.70454545]

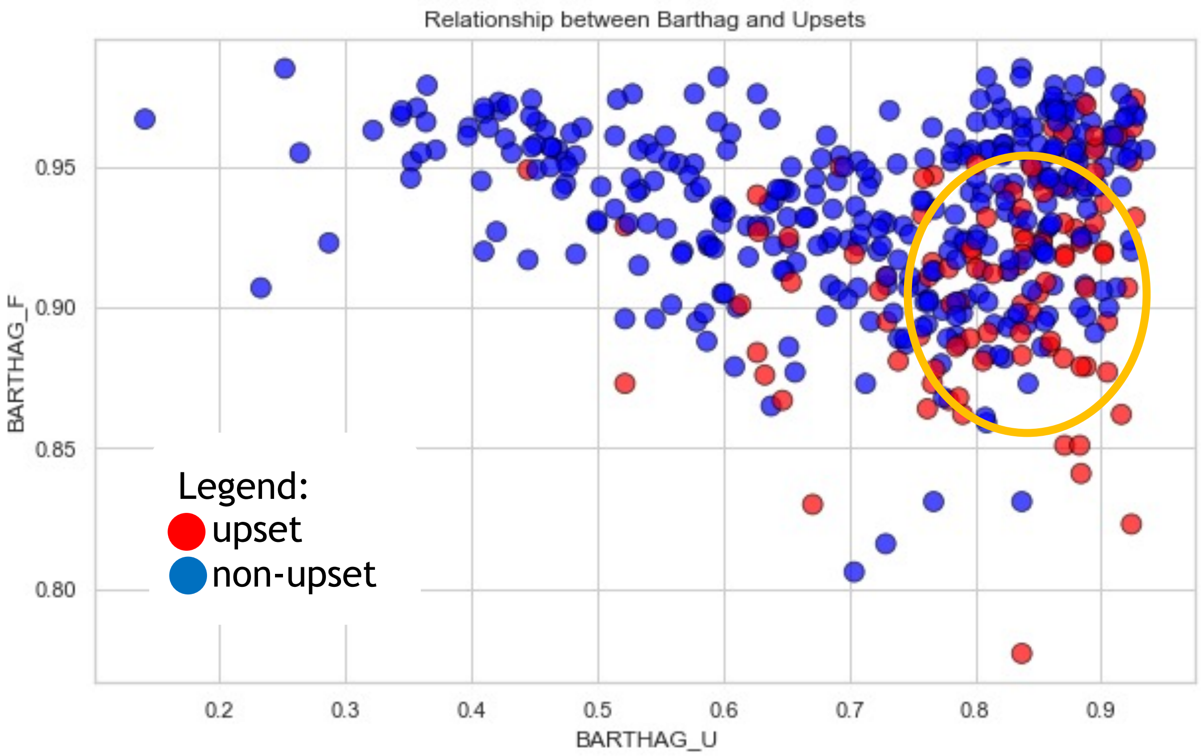
Average cv accuracy: 0.74318

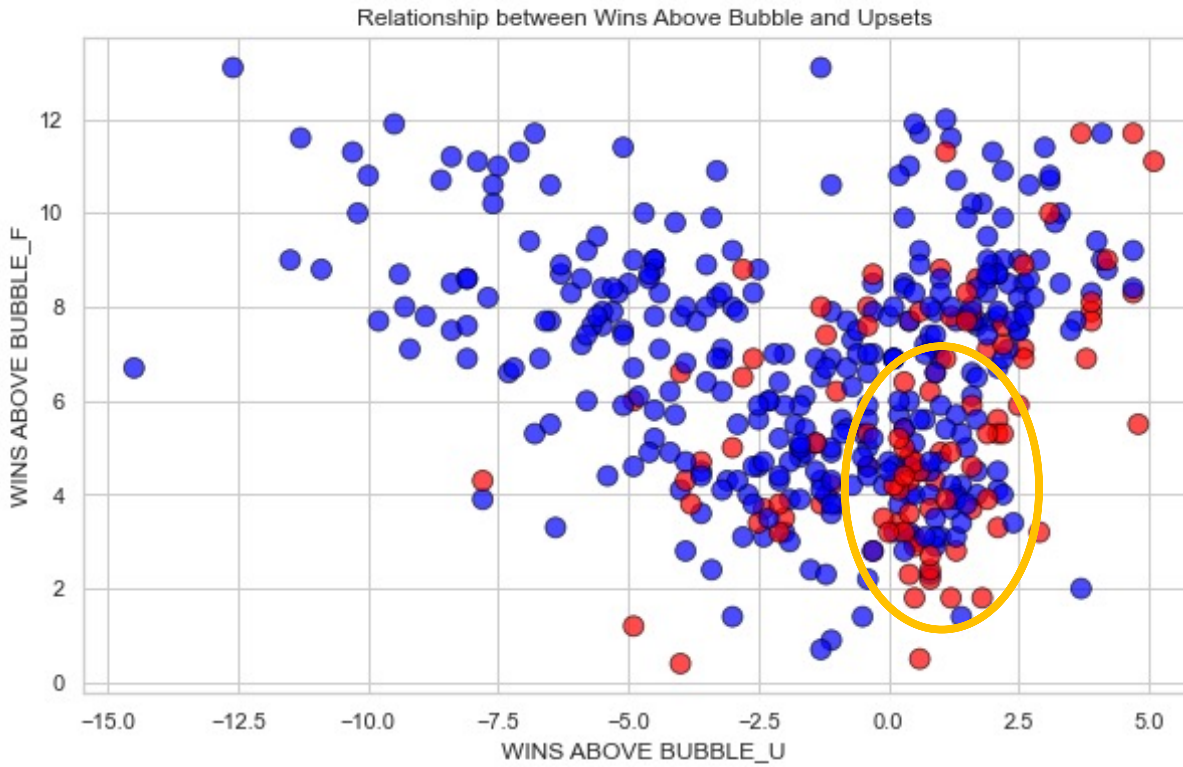


Winner

Logistic Regression Model
Accuracy: 0.76590

Analysis of Barthag and Wins Above Bubble





Barthag: Power Rating (Chance of beating the average Division I basketball team).

A game of favorite team with barthag < 0.95 and underdog with barthag > 0.75 has a higher chance of an upset

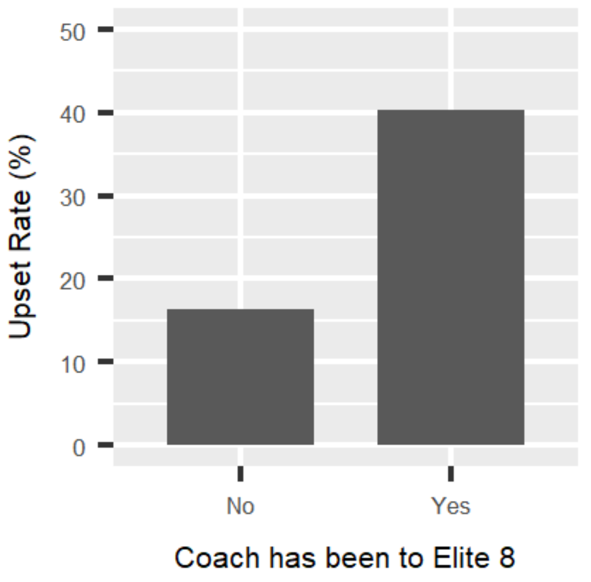
Wins Above Bubble (WAB): The expected winning percentage for an average *bubble team in each game of a team's schedule subtracted from the total of the team's actual wins.

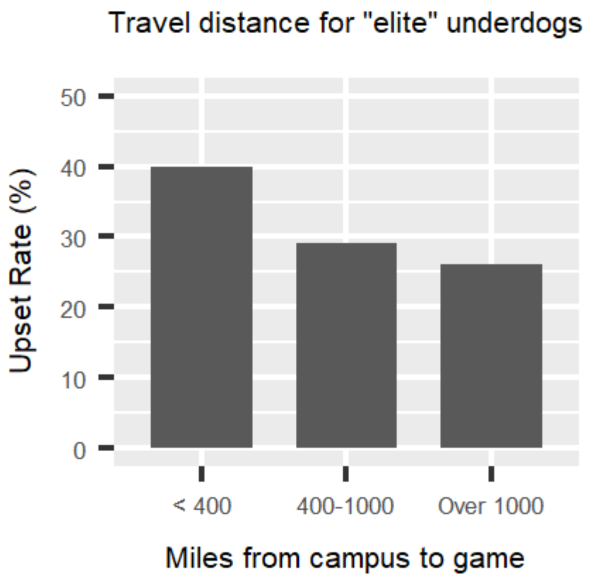
A game of favorite team with WAB < 6 and underdog with WAB > 0 has a higher chance of an upset

*The bubble teams for the NCAA tournament are on the cusp of making the field of 68, but an invitation isn't guaranteed

Limitations

- Data period
  - Used 2008-2021 data for this project. March Madness has a long history since 1939
- Alternative data
  - Did not use data outside of team stats. It could produce a model with higher accuracy if able to collect alternative data e.g. coach experience, travel distance
- Recency data
  - No data that takes recency into account. There is something to be said about peaking at the right time in the season
- Resource and time
  - If there were more time and resources it would be interesting to deep dive into the data more by testing additional prediction models or looking into the trends through different rounds





Source: https://towardsdatascience.com/use-my-upset-prediction-model-to-pick-underdogs-in-your-ncaa-tournament-bracket-87c4aa3935f5

Conclusion

Upsets in March Madness are considered very unusual and unpredictable to the general public but through this study, the created model proved to be able to predict with a fairly high accuracy.

Below are key takeaways:

- Upset is defined as an underdog win in game where **Higher Seed - Lower Seed >= 5**
- Round 4 (Elite 8) has the highest probability of upsets
- Out of the three tested models (logistic regression, random forest and gradient boosting) **logistic regression has the highest accuracy**
- **Barthag** and **Wins Above Bubble** features are important to the random forest and gradient boosting models

Thank You

References

1. Worley, M. (2022, March 3). *Use my upset prediction model to pick underdogs in your NCAA tournament bracket*. Medium. https://towardsdatascience.com/use-my-upset-prediction-model-to-pick-underdogs-in-your-ncaa-tournament-bracket-87c4aa3935f5
2. *T-Rank - Customizable College Basketball Tempo Free Stats - T-Rank*. (n.d.). Www.barttorvik.com. Retrieved July 27, 2023, from https://www.barttorvik.com
3. *2021 Pomeroy College Basketball Ratings*. (n.d.). Kenpom.com. https://kenpom.com/
4. *Predicting Upsets*. (n.d.). Kaggle.com. Retrieved July 27, 2023, from https://www.kaggle.com/code/nishaanamin/predicting-upsets#Upset-Count-Per-Tournament-
5. Worley, M. (2022, March 3). *Predicting Upsets in the NCAA Tournament with Machine Learning*. Medium. https://towardsdatascience.com/predicting-upsets-in-the-ncaa-tournament-with-machine-learning-816fecf41f01
6. *Advanced Metrics in College Basketball and How to use them | Hoops Amino*. (n.d.). Hoops | Aminoapps.com. Retrieved July 27, 2023, from https://aminoapps.com/c/hoops/page/blog/advanced-metrics-in-college-basketball-and-how-to-use-them/RrNH_wuKmZrxjxL2VY8Ylkkwoo5p4z
7. C.J. Moore. (2013, March 7). *The Ken Pomeroy Effect: How Advanced Metrics Have Revolutionized NCAA Basketball*. Bleacher Report; Bleacher Report. https://bleacherreport.com/articles/1556819-the-ken-pomeroy-effect-how-advanced-metrics-have-revolutionized-ncaa-basketball
8. *NCAAM March Madness scores 1985-2021*. (n.d.). Www.kaggle.com. Retrieved July 27, 2023, from https://www.kaggle.com/datasets/woodygilbertson/ncaam-march-madness-scores-19852021