Predicting The Perfect March Madness Bracket

Matthew Zaback

Overview

1. Background
2. What Others Have Done
3. My Approach
4. Results
5. Conclusion

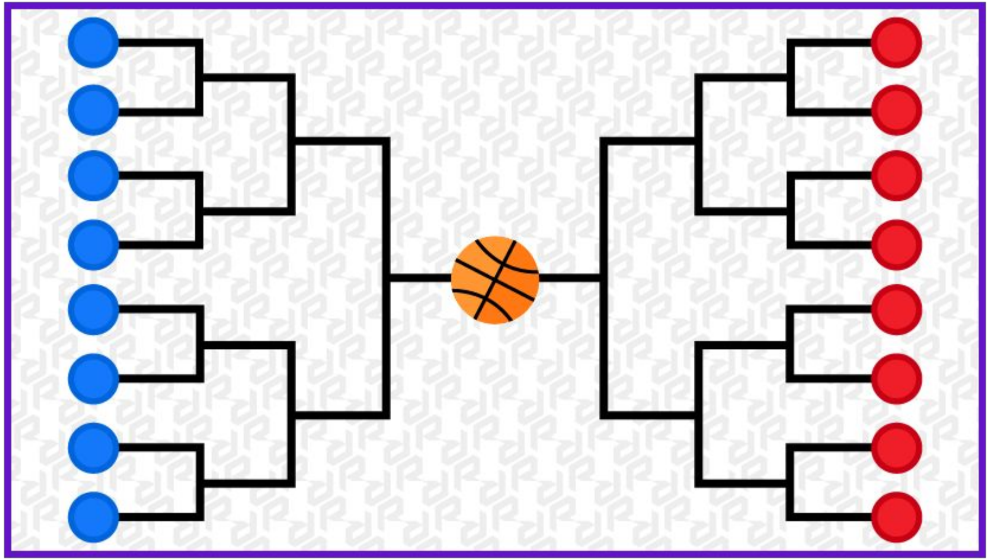
Background

What is Machine Learning?

Subfield of artificial intelligence that gives the computer the ability to learn without explicitly being programmed

Allows the user to feed it immense amount of data and have the computer analyze it and make data driven recommendations

What is March Madness?



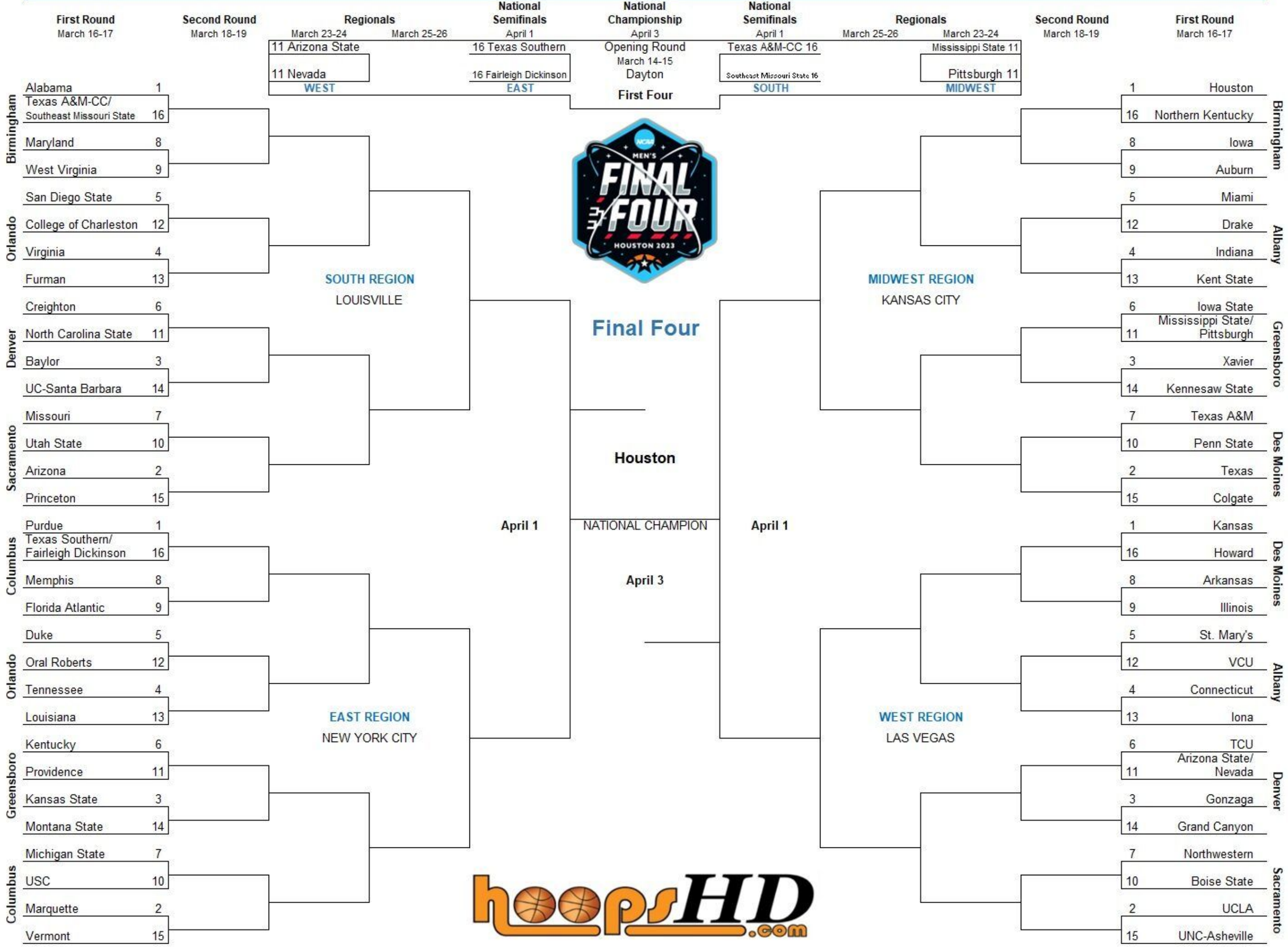
March Madness

Time when NCAA college basketball tournament is held

Single elimination featuring 68 teams

“Upset” = when a higher seeded team beats a lower seeded team

2023 NCAA DIVISION I MEN'S BASKETBALL CHAMPIONSHIP



First Round	Second Round	Regionals		National Semifinals	National Championship	National Semifinals	Regionals		Second Round	First Round
March 16-17	March 18-19	March 23-24	March 25-26	April 1	April 3	April 1	March 25-26	March 23-24	March 18-19	March 16-17

Opening Round
March 14-15
Dayton
First Four

WEST	EAST	SOUTH	MIDWEST
11 Arizona State	16 Texas Southern	Texas A&M-CC 16	Mississippi State 11
11 Nevada	16 Fairleigh Dickinson	Southeast Missouri State 16	Pittsburgh 11

SOUTH REGION
LOUISVILLE

	Team	Seed
Birmingham	Alabama	1
	Texas A&M-CC/ Southeast Missouri State	16
	Maryland	8
	West Virginia	9
Orlando	San Diego State	5
	College of Charleston	12
	Virginia	4
	Furman	13
Denver	Creighton	6
	North Carolina State	11
	Baylor	3
	UC-Santa Barbara	14
Sacramento	Missouri	7
	Utah State	10
	Arizona	2
	Princeton	15

EAST REGION
NEW YORK CITY

	Team	Seed
Columbus	Purdue	1
	Texas Southern/ Fairleigh Dickinson	16
	Memphis	8
	Florida Atlantic	9
Orlando	Duke	5
	Oral Roberts	12
	Tennessee	4
	Louisiana	13
Greensboro	Kentucky	6
	Providence	11
	Kansas State	3
	Montana State	14
Columbus	Michigan State	7
	USC	10
	Marquette	2
	Vermont	15

Final Four

April 1

Houston

NATIONAL CHAMPION

April 3

April 1

MIDWEST REGION
KANSAS CITY

Seed	Team	
1	Houston	Birmingham
16	Northern Kentucky	
8	Iowa	
9	Auburn	
5	Miami	Albany
12	Drake	
4	Indiana	
13	Kent State	
6	Iowa State	Greensboro
11	Mississippi State/ Pittsburgh	
3	Xavier	
14	Kennesaw State	
7	Texas A&M	Des Moines
10	Penn State	
2	Texas	
15	Colgate	

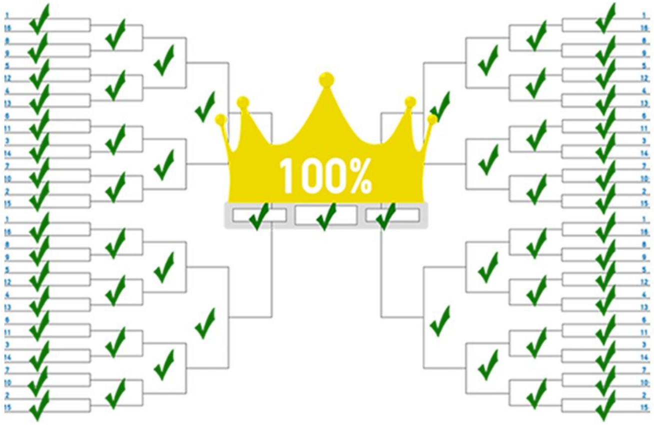
WEST REGION
LAS VEGAS

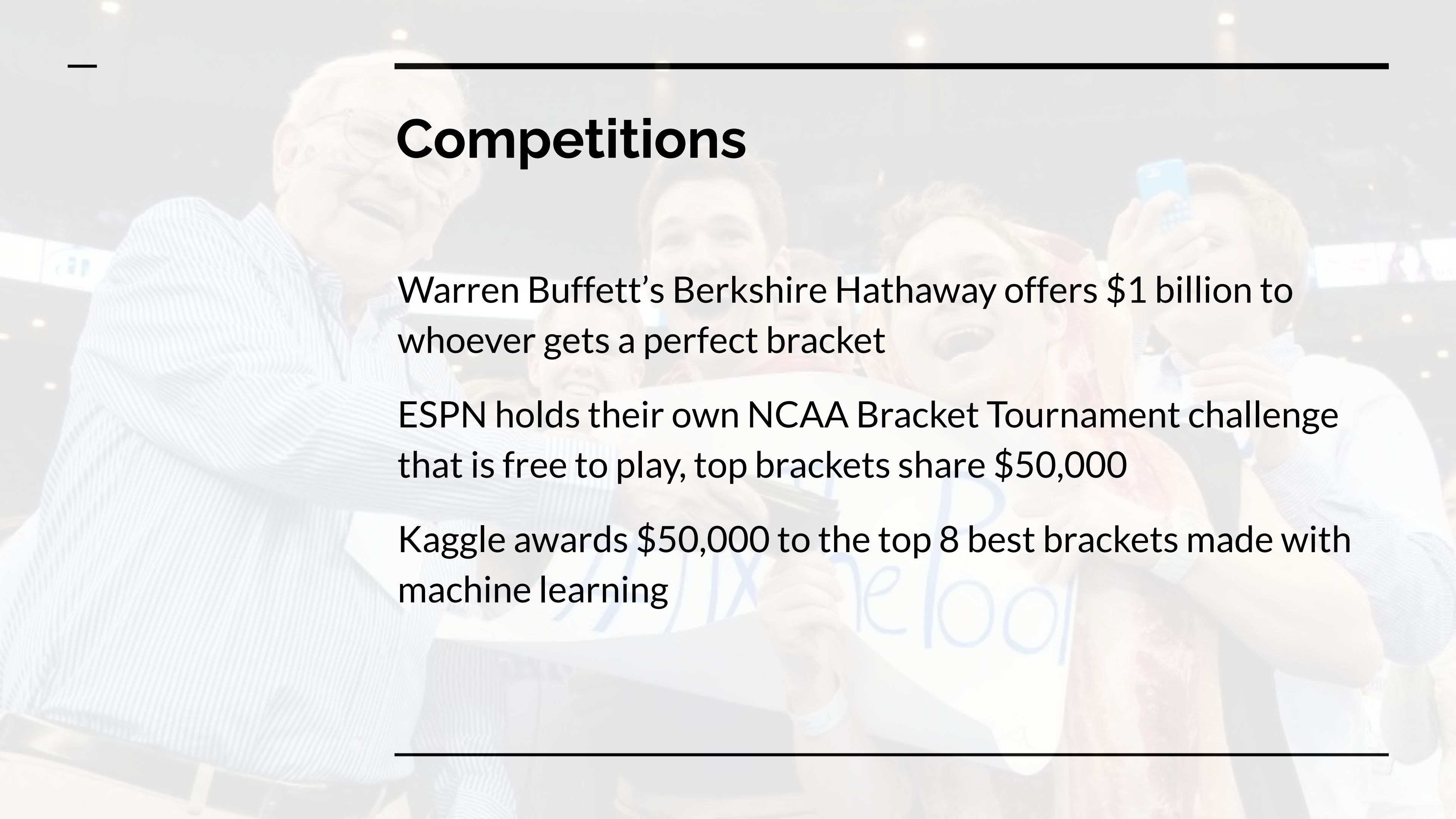
Seed	Team	
1	Kansas	Des Moines
16	Howard	
8	Arkansas	
9	Illinois	
5	St. Mary's	Albany
12	VCU	
4	Connecticut	
13	Iona	
6	TCU	Denver
11	Arizona State/ Nevada	
3	Gonzaga	
14	Grand Canyon	
7	Northwestern	Sacramento
10	Boise State	
2	UCLA	
15	UNC-Asheville	

The Odds Of Guessing Every Winner

1 in 9,223,372,036,854,775,808

2^{63}





Competitions

Warren Buffett’s Berkshire Hathaway offers $1 billion to whoever gets a perfect bracket

ESPN holds their own NCAA Bracket Tournament challenge that is free to play, top brackets share $50,000

Kaggle awards $50,000 to the top 8 best brackets made with machine learning

Why do it?

According to the American Gaming Association, roughly 80 million brackets are filled out each year with 40 million americans participating.

Most do it for a chance to win something. The AGA estimates the average entry fee for an office pool is $29, with $2 billion wagered on pools alone.

80M
Tournament brackets completed each year

VS.

156M
Ballots cast in the 2020 presidential election

78%
Of employees say celebrating March Madness at work boosts morale

29%
Of March Madness fans participate in office pools

39%
Of workers say they became closer with a coworker after participating in an office pool

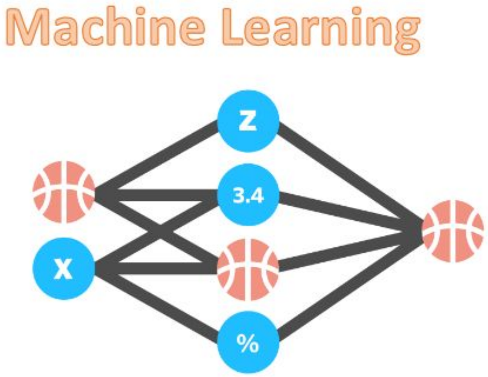
Closest to Perfect

GREGG NIGL
PERFECT NCAA BRACKET
GMA
U.S. EDUCATION DEPARTMENT REPORTEDLY LAUNCHING PRELIMINARY INVESTIGATION INTO 8 UNIV

Year	Games Lasted
2023	24 games
2022	27 games
2021	27 games
2019	49 games
2018	25 games
2017	39 games
2016	25 games
1977-2015	36 games

What Others Have Done

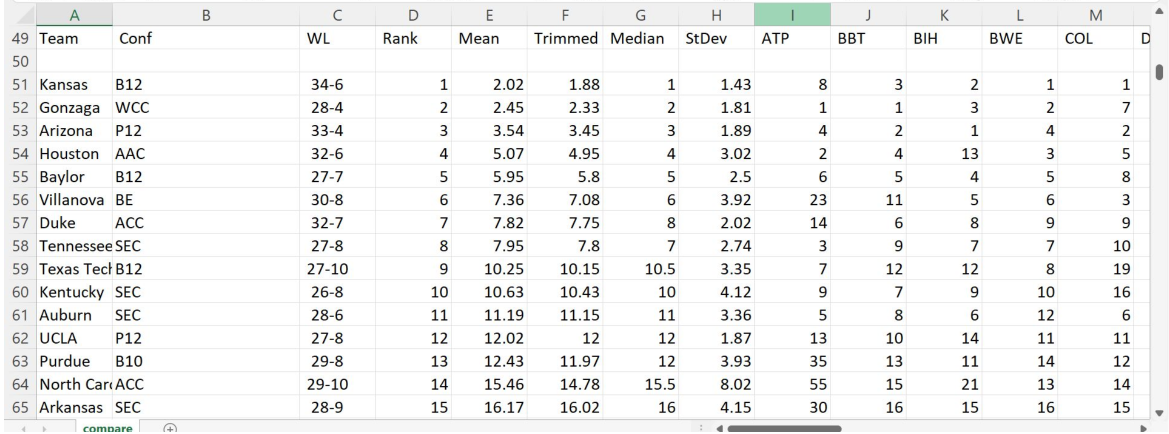
What Others Have Done



Machine Learning Algorithms

Neural Net
XGBoost
Random Forest
Naive Bayes
Logistic Regression
KNN
Support Vector Machine
AdaBoost

Will Geoghegan (2021)



	A	B	C	D	E	F	G	H	I	J	K	L	M	
49	Team	Conf	WL	Rank	Mean	Trimmed	Median	StDev	ATP	BBT	BIH	BWE	COL	D
50														
51	Kansas	B12	34-6	1	2.02	1.88	1	1.43	8	3	2	1	1	
52	Gonzaga	WCC	28-4	2	2.45	2.33	2	1.81	1	1	3	2	7	
53	Arizona	P12	33-4	3	3.54	3.45	3	1.89	4	2	1	4	2	
54	Houston	AAC	32-6	4	5.07	4.95	4	3.02	2	4	13	3	5	
55	Baylor	B12	27-7	5	5.95	5.8	5	2.5	6	5	4	5	8	
56	Villanova	BE	30-8	6	7.36	7.08	6	3.92	23	11	5	6	3	
57	Duke	ACC	32-7	7	7.82	7.75	8	2.02	14	6	8	9	9	
58	Tennessee	SEC	27-8	8	7.95	7.8	7	2.74	3	9	7	7	10	
59	Texas Tech	B12	27-10	9	10.25	10.15	10.5	3.35	7	12	12	8	19	
60	Kentucky	SEC	26-8	10	10.63	10.43	10	4.12	9	7	9	10	16	
61	Auburn	SEC	28-6	11	11.19	11.15	11	3.36	5	8	6	12	6	
62	UCLA	P12	27-8	12	12.02	12	12	1.87	13	10	14	11	11	
63	Purdue	B10	29-8	13	12.43	11.97	12	3.93	35	13	11	14	12	
64	North Car	ACC	29-10	14	15.46	14.78	15.5	8.02	55	15	21	13	14	
65	Arkansas	SEC	28-9	15	16.17	16.02	16	4.15	30	16	15	16	15	

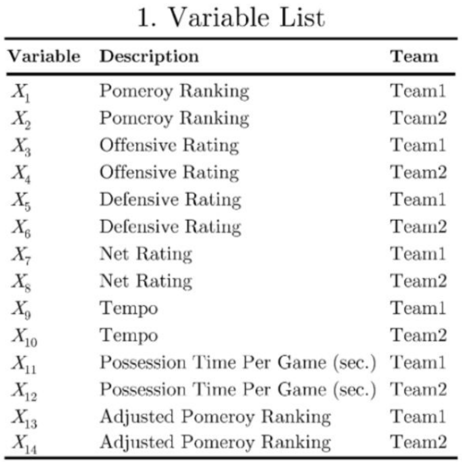
Screenshot of Massey Ratings Data

Top 0.2% of ESPN brackets in 2021

He used rankings from many different websites and sports analysts

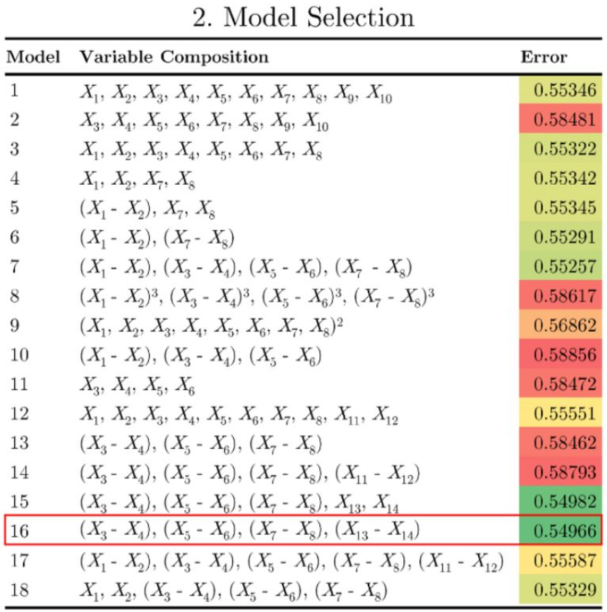
AdaBoost machine learning algorithm

Lotan Weininger (2019)



1. Variable List

Variable	Description	Team
X_1	Pomeroy Ranking	Team1
X_2	Pomeroy Ranking	Team2
X_3	Offensive Rating	Team1
X_4	Offensive Rating	Team2
X_5	Defensive Rating	Team1
X_6	Defensive Rating	Team2
X_7	Net Rating	Team1
X_8	Net Rating	Team2
X_9	Tempo	Team1
X_{10}	Tempo	Team2
X_{11}	Possession Time Per Game (sec.)	Team1
X_{12}	Possession Time Per Game (sec.)	Team2
X_{13}	Adjusted Pomeroy Ranking	Team1
X_{14}	Adjusted Pomeroy Ranking	Team2



2. Model Selection

Model	Variable Composition	Error
1	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$	0.55346
2	$X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$	0.58481
3	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$	0.55322
4	X_1, X_2, X_7, X_8	0.55342
5	$(X_1 - X_2), X_7, X_8$	0.55345
6	$(X_1 - X_2), (X_7 - X_8)$	0.55291
7	$(X_1 - X_2), (X_3 - X_4), (X_5 - X_6), (X_7 - X_8)$	0.55257
8	$(X_1 - X_2)^3, (X_3 - X_4)^3, (X_5 - X_6)^3, (X_7 - X_8)^3$	0.58617
9	$(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)^2$	0.56862
10	$(X_1 - X_2), (X_3 - X_4), (X_5 - X_6)$	0.58856
11	X_3, X_4, X_5, X_6	0.58472
12	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_{11}, X_{12}$	0.55551
13	$(X_3 - X_4), (X_5 - X_6), (X_7 - X_8)$	0.58462
14	$(X_3 - X_4), (X_5 - X_6), (X_7 - X_8), (X_{11} - X_{12})$	0.58793
15	$(X_3 - X_4), (X_5 - X_6), (X_7 - X_8), X_{13}, X_{14}$	0.54982
16	$(X_3 - X_4), (X_5 - X_6), (X_7 - X_8), (X_{13} - X_{14})$	0.54966
17	$(X_1 - X_2), (X_3 - X_4), (X_5 - X_6), (X_7 - X_8), (X_{11} - X_{12})$	0.55587
18	$X_1, X_2, (X_3 - X_4), (X_5 - X_6), (X_7 - X_8)$	0.55329

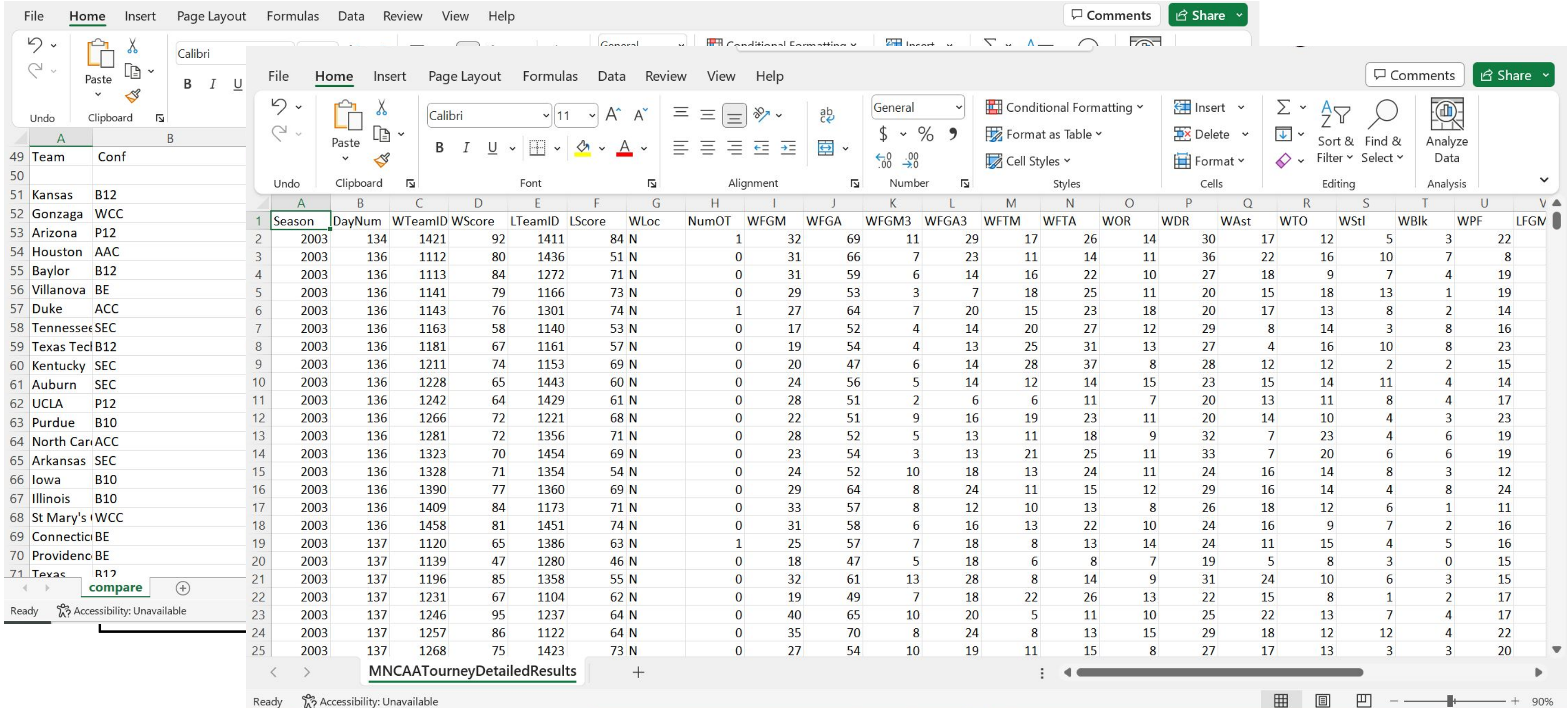
Logistic regression model

Simulated the performance of the variables by testing the predictions on previous years data

Top 10% in Kaggle competition

My Approach

Data Available (Since 2003)



	A	B
49	Team	Conf
50		
51	Kansas	B12
52	Gonzaga	WCC
53	Arizona	P12
54	Houston	AAC
55	Baylor	B12
56	Villanova	BE
57	Duke	ACC
58	Tennesse	SEC
59	Texas Tec	B12
60	Kentucky	SEC
61	Auburn	SEC
62	UCLA	P12
63	Purdue	B10
64	North Car	ACC
65	Arkansas	SEC
66	Iowa	B10
67	Illinois	B10
68	St Mary's (	WCC
69	Connectic	BE
70	Providenc	BE
71	Texas	B12

compare

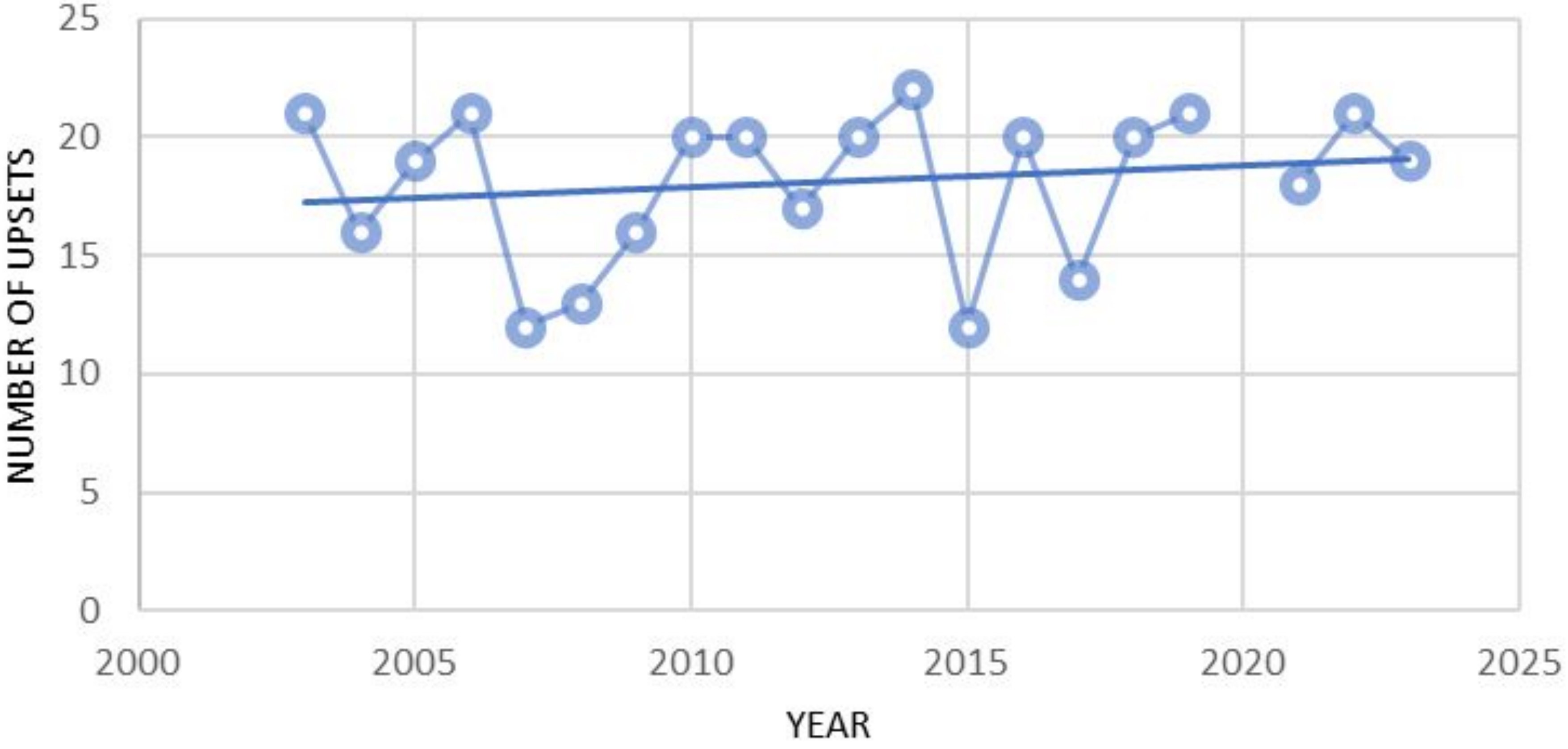
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Season	DayNum	WTeamID	WScore	LTeamID	LScore	WLoc	NumOT	WFGM	WFGA	WFGM3	WFGA3	WFTM	WFTA	WOR	WDR	WAst	WTO	WStl	WBlk	WPF	LFGM
2	2003	134	1421	92	1411	84	N	1	32	69	11	29	17	26	14	30	17	12	5	3	22	
3	2003	136	1112	80	1436	51	N	0	31	66	7	23	11	14	11	36	22	16	10	7	8	
4	2003	136	1113	84	1272	71	N	0	31	59	6	14	16	22	10	27	18	9	7	4	19	
5	2003	136	1141	79	1166	73	N	0	29	53	3	7	18	25	11	20	15	18	13	1	19	
6	2003	136	1143	76	1301	74	N	1	27	64	7	20	15	23	18	20	17	13	8	2	14	
7	2003	136	1163	58	1140	53	N	0	17	52	4	14	20	27	12	29	8	14	3	8	16	
8	2003	136	1181	67	1161	57	N	0	19	54	4	13	25	31	13	27	4	16	10	8	23	
9	2003	136	1211	74	1153	69	N	0	20	47	6	14	28	37	8	28	12	12	2	2	15	
10	2003	136	1228	65	1443	60	N	0	24	56	5	14	12	14	15	23	15	14	11	4	14	
11	2003	136	1242	64	1429	61	N	0	28	51	2	6	6	11	7	20	13	11	8	4	17	
12	2003	136	1266	72	1221	68	N	0	22	51	9	16	19	23	11	20	14	10	4	3	23	
13	2003	136	1281	72	1356	71	N	0	28	52	5	13	11	18	9	32	7	23	4	6	19	
14	2003	136	1323	70	1454	69	N	0	23	54	3	13	21	25	11	33	7	20	6	6	19	
15	2003	136	1328	71	1354	54	N	0	24	52	10	18	13	24	11	24	16	14	8	3	12	
16	2003	136	1390	77	1360	69	N	0	29	64	8	24	11	15	12	29	16	14	4	8	24	
17	2003	136	1409	84	1173	71	N	0	33	57	8	12	10	13	8	26	18	12	6	1	11	
18	2003	136	1458	81	1451	74	N	0	31	58	6	16	13	22	10	24	16	9	7	2	16	
19	2003	137	1120	65	1386	63	N	1	25	57	7	18	8	13	14	24	11	15	4	5	16	
20	2003	137	1139	47	1280	46	N	0	18	47	5	18	6	8	7	19	5	8	3	0	15	
21	2003	137	1196	85	1358	55	N	0	32	61	13	28	8	14	9	31	24	10	6	3	15	
22	2003	137	1231	67	1104	62	N	0	19	49	7	18	22	26	13	22	15	8	1	2	17	
23	2003	137	1246	95	1237	64	N	0	40	65	10	20	5	11	10	25	22	13	7	4	17	
24	2003	137	1257	86	1122	64	N	0	35	70	8	24	8	13	15	29	18	12	12	4	22	
25	2003	137	1268	75	1423	73	N	0	27	54	10	19	11	15	8	27	17	13	3	3	20	

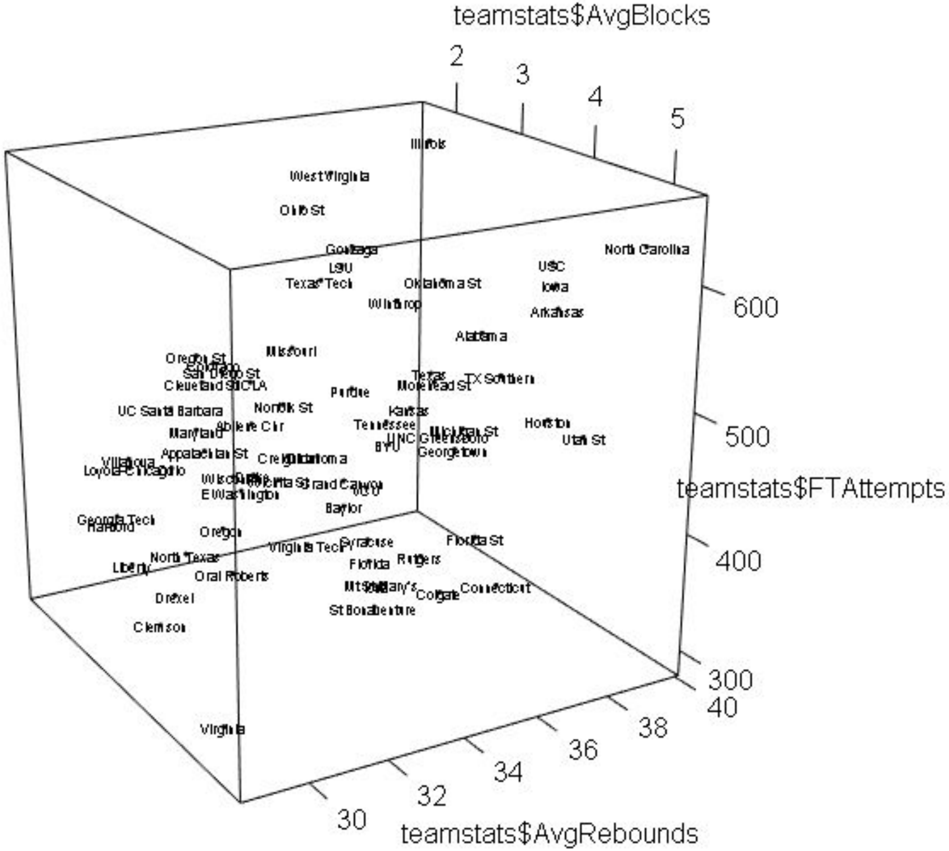
MNCAATourneyDetailedResults

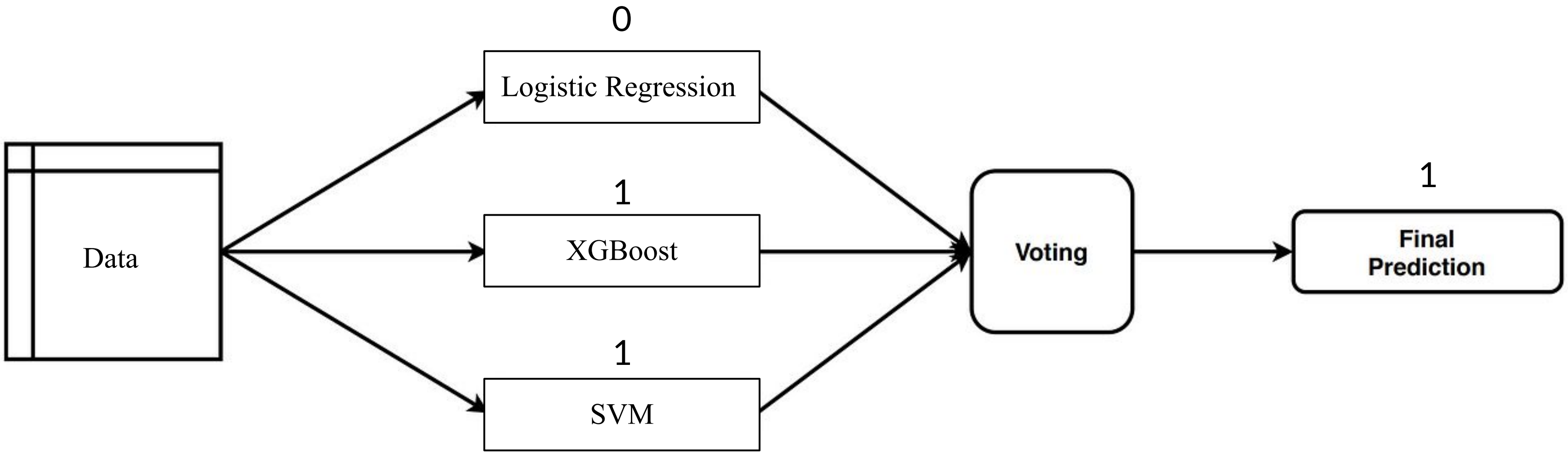
First Round Results Since 2003

Seed	Wins	Losses	Win Percentage
1	78	2	97.5%
2	73	7	91.3%
3	71	9	88.8%
4	63	17	78.8%
5	50	30	62.5%
6	45	35	56.3%
7	50	30	62.5%
8	44	36	55%
9	36	44	45%
10	30	50	37.5%
11	35	45	43.8%
12	30	50	37.5%
13	17	63	21.3%
14	9	71	11.3%
15	7	73	8.8%
16	2	78	2.5%

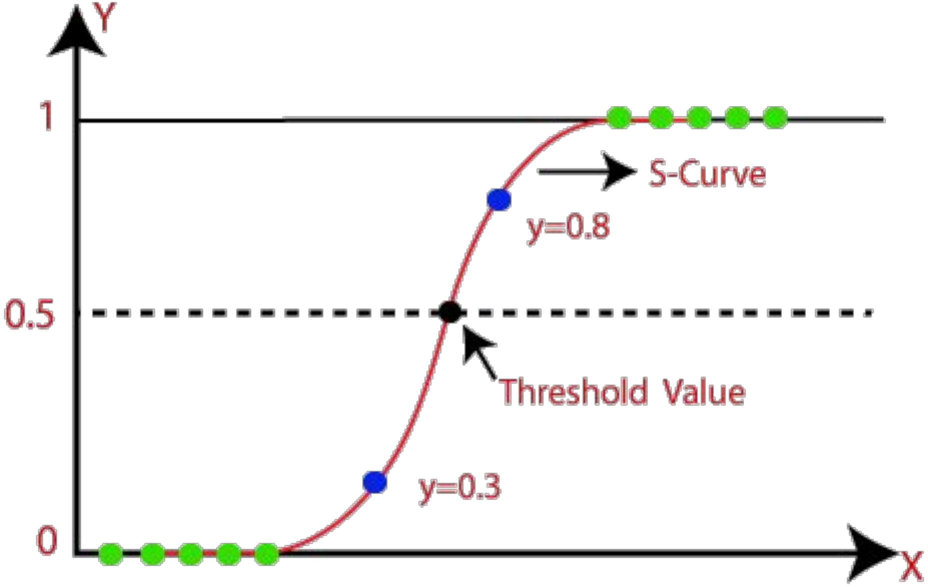
Number of Upsets per Year Since 2003




teamstats$AvgBlocks
2
3
4
5
teamstats$FTAttempts
600
500
400
300
teamstats$AvgRebounds
30
32
34
36
38
40
Illinois
West Virginia
Ohio St
Gonzaga
LSU
Texas Tech
Oklahoma St
Winthrop
USC
Iowa
North Carolina
Arkansas
Alabama
Missouri
Oregon St
Colorado
San Diego St
Cleveland St
UCLA
Purdue
Texas
Morehead St
TX Southern
UC Santa Barbara
Norfolk St
Kansas
Tennessee
Abilene Chr
Maryland
Michigan St
Houston
UNC Greensboro
Utah St
BYU
Georgetown
Appalachian St
Creighton
Oklahoma
Villanova
Loyola-Chicago
Ohio
Wisconsin
Wichita St
Grand Canyon
VCU
E Washington
Baylor
Georgia Tech
Hartford
Oregon
Virginia Tech
Syracuse
Florida St
North Texas
Liberty
Florida
Rutgers
Oral Roberts
Mt St Mary's
Iona
Colgate
Connecticut
Drexel
St Bonaventure
Clemson
Virginia


0
Logistic Regression
1
XGBoost
1
SVM
Data
Voting
1
Final Prediction

Logistic Regression

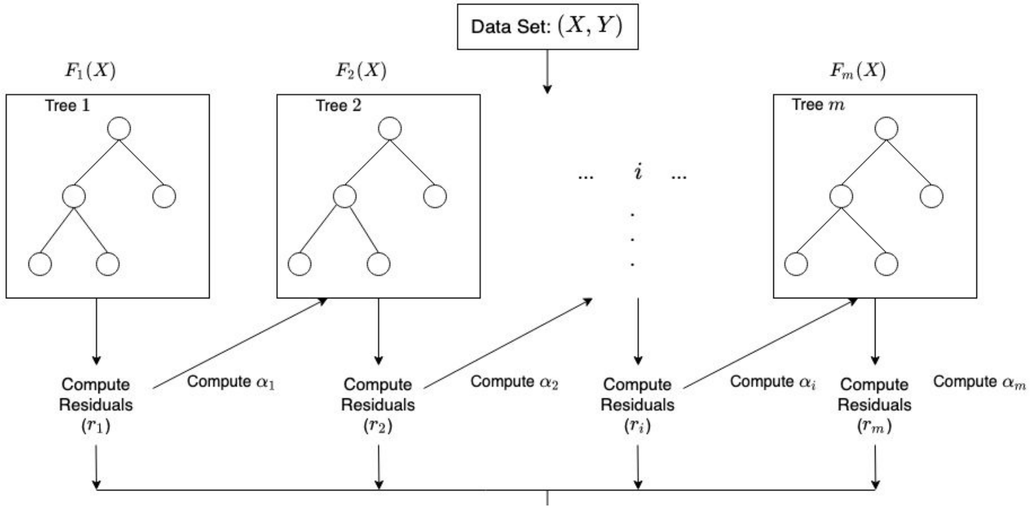


Binary classification model (upset or not)

Gives probability, then classifies it based on the threshold value

Unlike linear regression, it cannot predict an actual value, just the probability

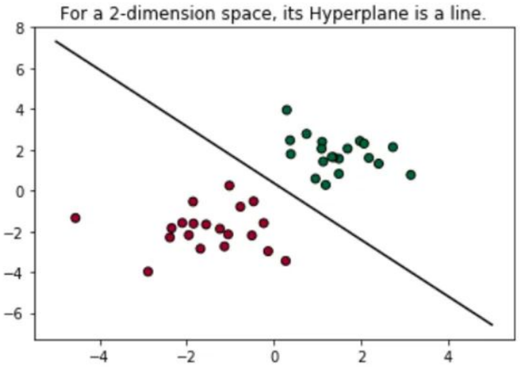
XGBoost (Extreme Gradient Boosting)

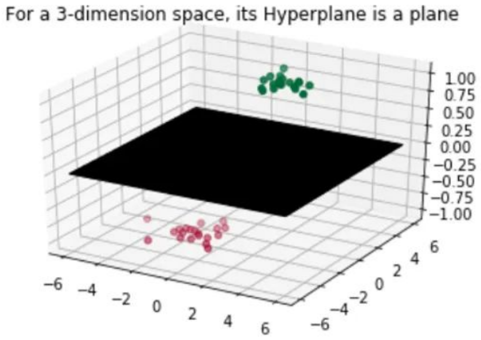


Trains a number of decision trees on subsets of the data, then combines the prediction from each tree into the final prediction

Has become popular because it is new and outperforms other ML models

Support Vector Machine (SVM)



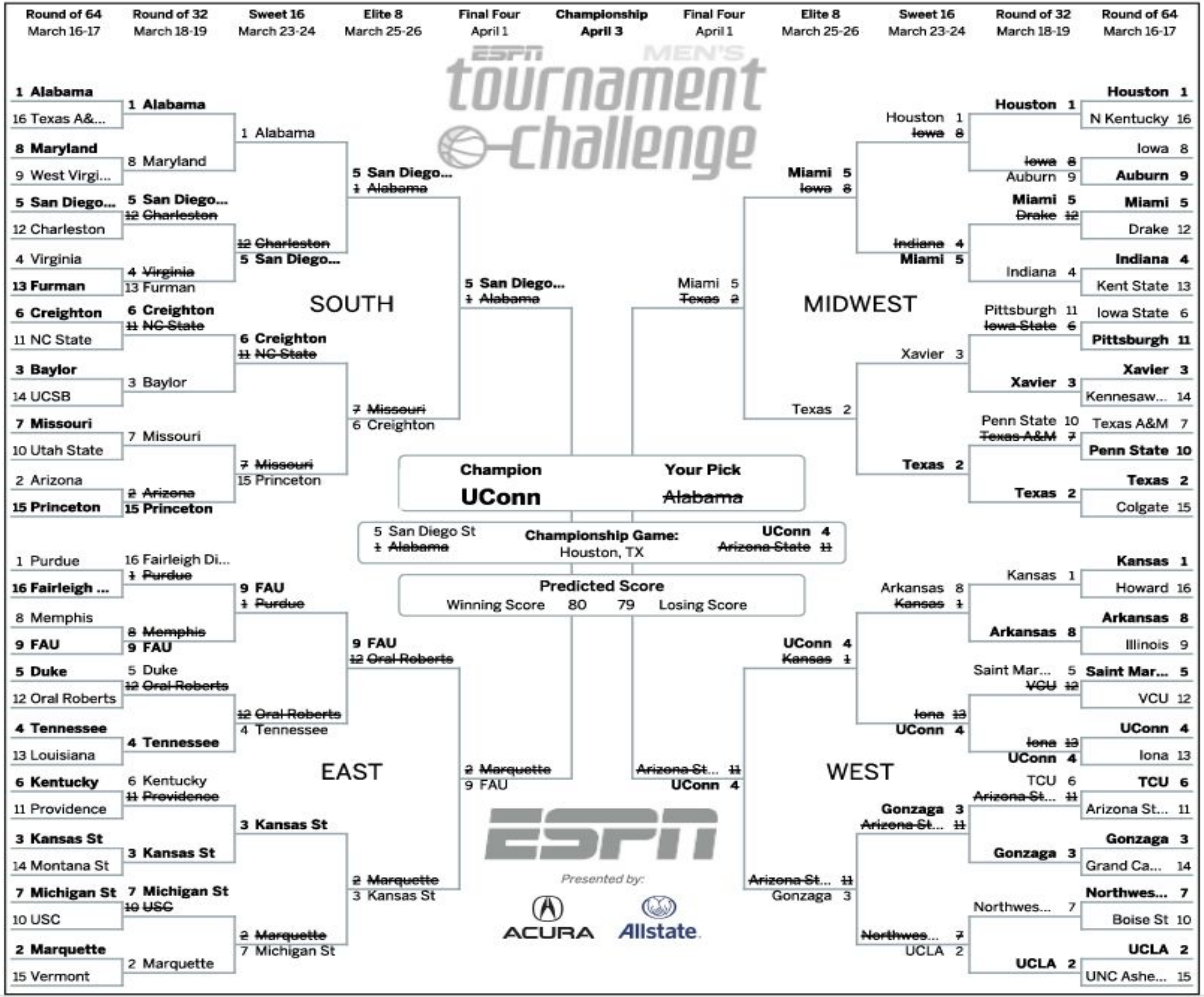


Finds a plane that most accurately separates the data into classes (upset or not)

Picks plane that maximizes the margin

Used 19 features to create a plane

Results



Round of 64 March 16-17 | Round of 32 March 18-19 | Sweet 16 March 23-24 | Elite 8 March 25-26 | Final Four April 1 | Championship April 3 | Final Four April 1 | Elite 8 March 25-26 | Sweet 16 March 23-24 | Round of 32 March 18-19 | Round of 64 March 16-17

ESPN MEN'S tournament challenge

SOUTH

1 Alabama
16 Texas A&...
8 Maryland
9 West Virgi...
5 San Diego...
12 Charleston
4 Virginia
13 Furman
6 Creighton
11 NC State
3 Baylor
14 UCSB
7 Missouri
10 Utah State
2 Arizona
15 Princeton

1 Alabama
8 Maryland
5 San Diego...
~~12 Charleston~~
4 ~~Virginia~~
13 Furman
6 Creighton
~~11 NC State~~
3 Baylor
7 Missouri
~~2 Arizona~~
15 Princeton

1 Alabama
~~12 Charleston~~
5 San Diego...
6 Creighton
~~11 NC State~~
~~7 Missouri~~
15 Princeton

5 San Diego...
~~1 Alabama~~
~~7 Missouri~~
6 Creighton

5 San Diego...
~~1 Alabama~~

EAST

1 Purdue
16 Fairleigh ...
8 Memphis
9 FAU
5 Duke
12 Oral Roberts
4 Tennessee
13 Louisiana
6 Kentucky
11 Providence
3 Kansas St
14 Montana St
7 Michigan St
10 USC
2 Marquette
15 Vermont

16 Fairleigh Di...
~~1 Purdue~~
~~8 Memphis~~
9 FAU
5 Duke
~~12 Oral Roberts~~
4 Tennessee
6 Kentucky
~~11 Providence~~
3 Kansas St
7 Michigan St
~~10 USC~~
2 Marquette

9 FAU
~~1 Purdue~~
~~12 Oral Roberts~~
4 Tennessee
3 Kansas St
~~2 Marquette~~
7 Michigan St

9 FAU
~~12 Oral Roberts~~
~~2 Marquette~~
3 Kansas St

~~2 Marquette~~
9 FAU

Champion
UConn

Your Pick
~~Alabama~~

5 San Diego St
~~1 Alabama~~

Championship Game:
Houston, TX

UConn 4
~~Arizona State 11~~

Predicted Score
Winning Score 80 79 Losing Score

ESPN
Presented by:
ACURA Allstate

MIDWEST

Miami 5
~~Texas 2~~

Miami 5
~~Iowa 8~~
Xavier 3
Texas 2

Houston 1
~~Iowa 8~~
~~Indiana 4~~
Miami 5
Xavier 3
Texas 2

Houston 1
~~Iowa 8~~
Auburn 9
Miami 5
~~Drake 12~~
Indiana 4
Pittsburgh 11
~~Iowa State 6~~
Xavier 3
Penn State 10
~~Texas A&M 7~~
Texas 2

Houston 1
N Kentucky 16
Iowa 8
Auburn 9
Miami 5
Drake 12
Indiana 4
Kent State 13
Iowa State 6
Pittsburgh 11
Xavier 3
Kennesaw... 14
Texas A&M 7
Penn State 10
Texas 2
Colgate 15

WEST

~~Arizona St... 11~~
UConn 4

UConn 4
~~Kansas 1~~
~~Arizona St... 11~~
Gonzaga 3

Arkansas 8
~~Kansas 1~~
~~Iona 13~~
UConn 4
Gonzaga 3
~~Arizona St... 11~~
~~Northwes... 7~~
UCLA 2

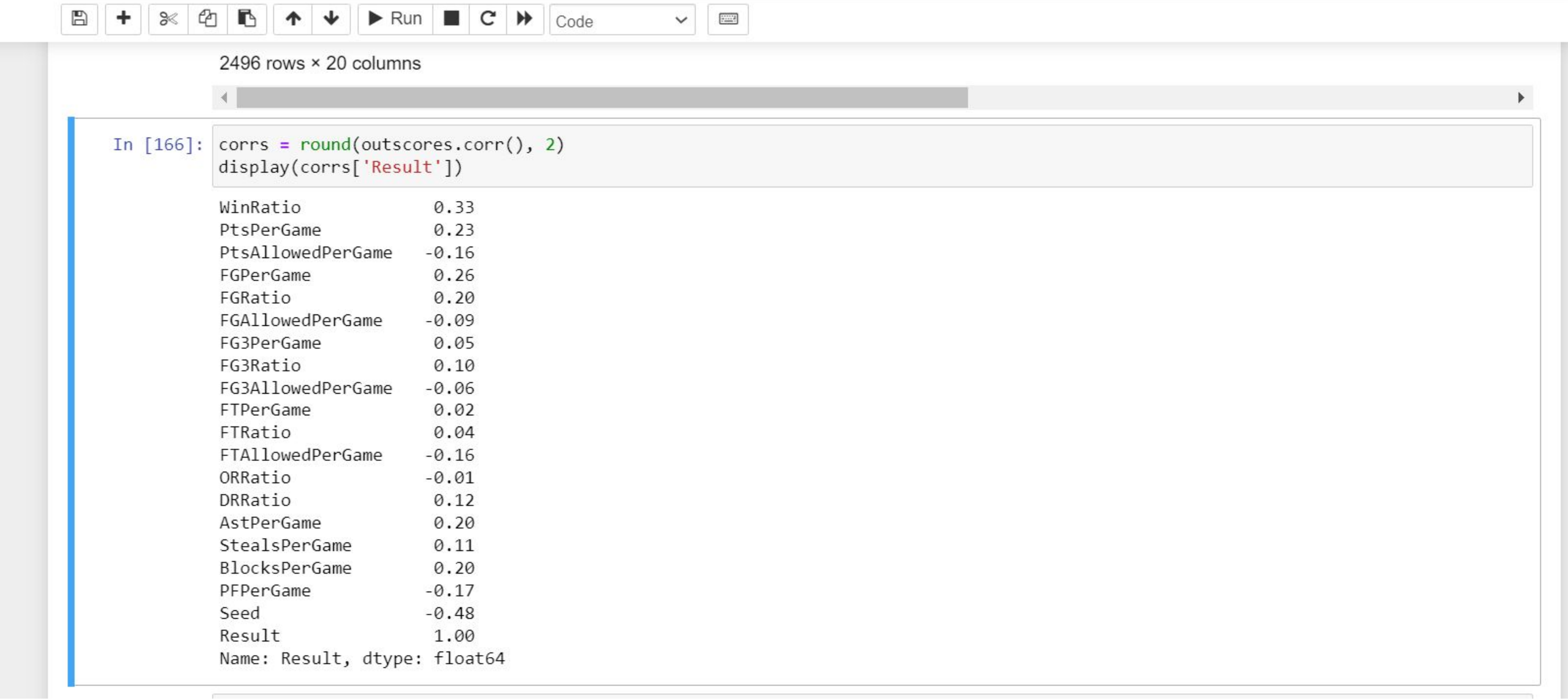
Kansas 1
Arkansas 8
Saint Mar... 5
~~VCU 12~~
~~Iona 13~~
UConn 4
TCU 6
~~Arizona St... 11~~
Gonzaga 3
Northwes... 7
UCLA 2

Kansas 1
Howard 16
Arkansas 8
Illinois 9
Saint Mar... 5
VCU 12
UConn 4
Iona 13
TCU 6
Arizona St... 11
Gonzaga 3
Grand Ca... 14
Northwes... 7
Boise St 10
UCLA 2
UNC Ashe... 15

21/63 games correctly classified (33%)

19/63 games were upsets (30.1%)

0/19 upsets predicted



2496 rows × 20 columns

```
In [166]: corrs = round(outscores.corr(), 2)
          display(corrs['Result'])
```

```
WinRatio             0.33
PtsPerGame           0.23
PtsAllowedPerGame   -0.16
FGPerGame            0.26
FGRatio              0.20
FGAllowedPerGame    -0.09
FG3PerGame           0.05
FG3Ratio             0.10
FG3AllowedPerGame   -0.06
FTPerGame            0.02
FTRatio              0.04
FTAllowedPerGame    -0.16
ORRatio             -0.01
DRRatio              0.12
AstPerGame           0.20
StealsPerGame        0.11
BlocksPerGame        0.20
PFPerGame           -0.17
Seed                -0.48
Result               1.00
Name: Result, dtype: float64
```

Conclusion

Conclusion

The bracket did not do as well as I hoped

It's important to make multiple brackets because going out on a limb for one upset can destroy the rest

It seems like there is a uniquely human element

It will be interesting to see if AI in the future will be able to predict a perfect bracket

Thank you
Any questions?