Predicting the NCAA Basketball Tournament

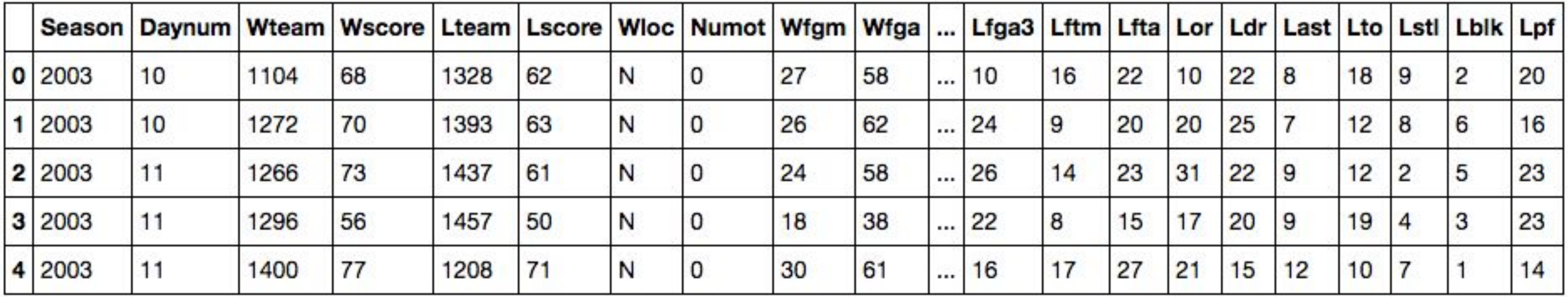
Billy Hines

Project Problem and Hypothesis

- Use historical data and team statistics to predict the outcome of the 2016 NCAA Men’s Basketball Tournament (without peeking)
- This is a classification problem: win or lose for a given matchup
- Win probabilities calculated for every potential matchup then used to create bracket
- Goal is to predict more games than by using seed alone

Data Transformation

- Download detailed game results from Kaggle:



	Season	Daynum	Wteam	Wscore	Lteam	Lscore	Wloc	Numot	Wfgm	Wfga	...	Lfga3	Lftm	Lfta	Lor	Ldr	Last	Lto	Lstl	Lblk	Lpf
0	2003	10	1104	68	1328	62	N	0	27	58	...	10	16	22	10	22	8	18	9	2	20
1	2003	10	1272	70	1393	63	N	0	26	62	...	24	9	20	20	25	7	12	8	6	16
2	2003	11	1266	73	1437	61	N	0	24	58	...	26	14	23	31	22	9	12	2	5	23
3	2003	11	1296	56	1457	50	N	0	18	38	...	22	8	15	17	20	9	19	4	3	23
4	2003	11	1400	77	1208	71	N	0	30	61	...	16	17	27	21	15	12	10	7	1	14

- Wrangle game level stats into season level stats per team

Feature Creation

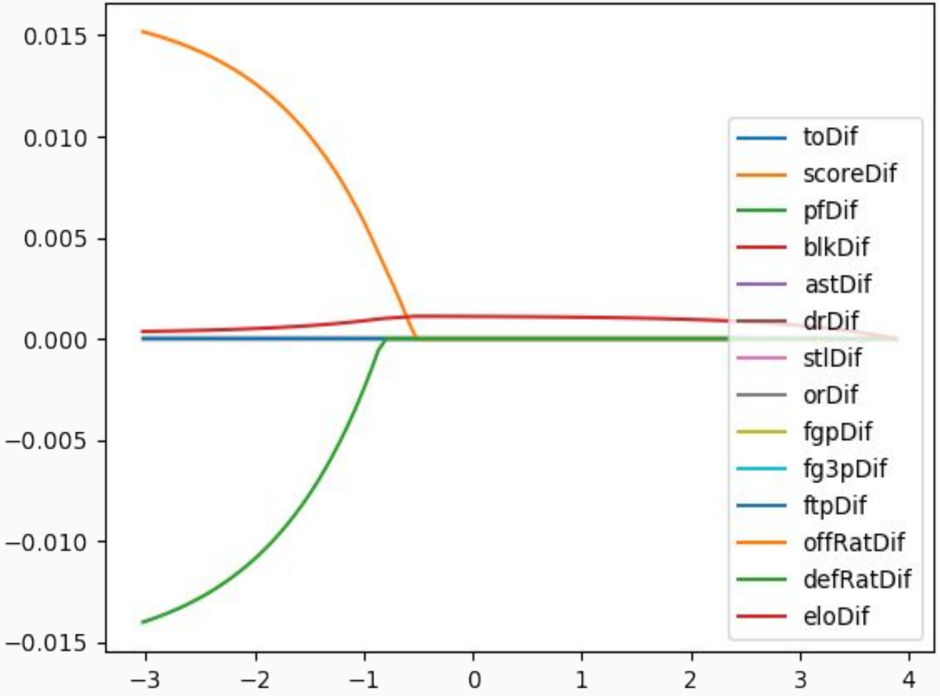
- Using win and loss records, iterate through entire data set to create ELO scores for each team
- Using season level team stats, derive offensive and defensive efficiency stats
- Merge data, take differences between both teams to create feature set for modeling

Data Left Behind

- Scraped 2017 data
- Historical point spreads
- Professional power rankings
- Home / Away and distance between school and arena

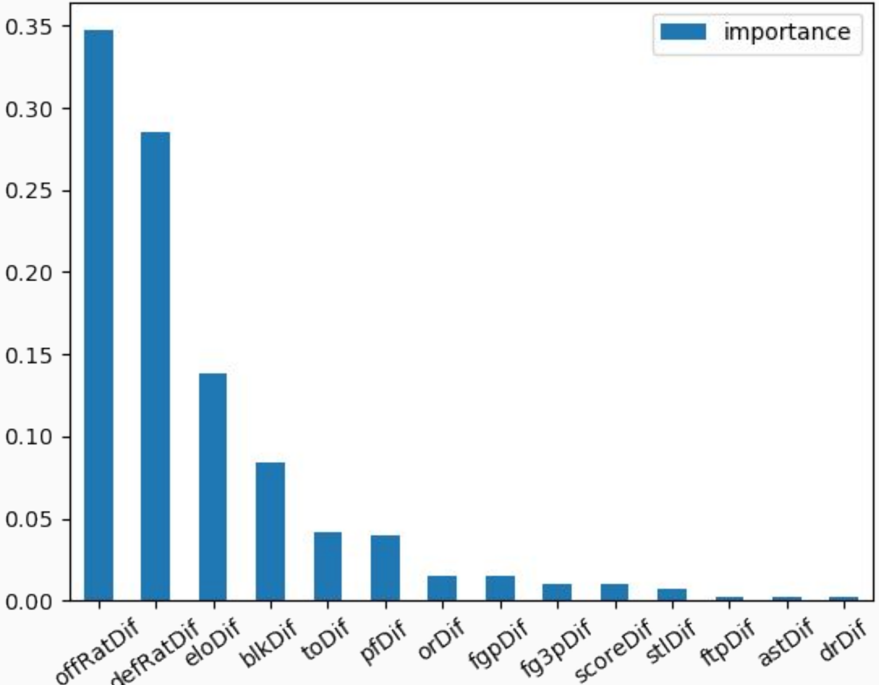
Initial Modeling Insights

Lasso Path:



Initial Modeling Insights

XGBoost Feature Importance:

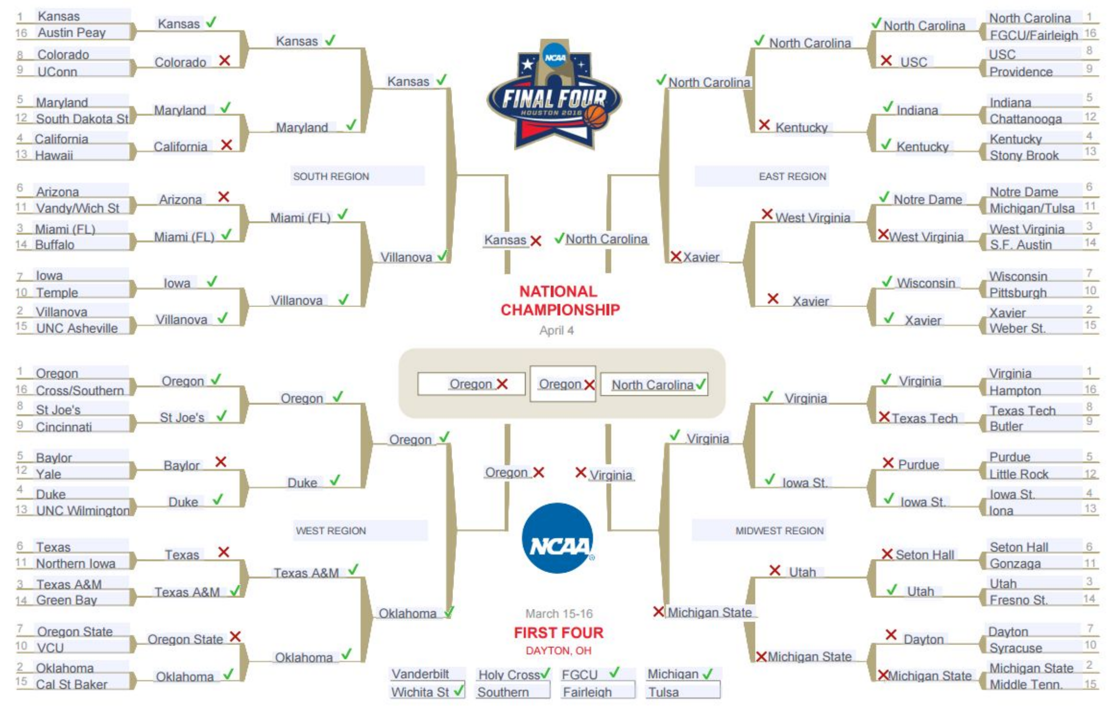


Model Results

Model	best_score_	test_score
LogisticRegression	-0.5234	-0.5172
XGBClassifier	-0.5249	-0.5186
DecisionTreeClassifier	-0.5344	-0.5279
Random Forest	-0.5668	-0.5789

best_score_ from CV over 80% of the data
test_score from evaluation on remaining 20% test set
Scores in neg_log_loss

Chalk Bracket



SOUTH REGION

1 Kansas
16 Austin Peay
8 Colorado
9 UConn
5 Maryland
12 South Dakota St
4 California
13 Hawaii
6 Arizona
11 Vandy/Wich St
3 Miami (FL)
14 Buffalo
7 Iowa
10 Temple
2 Villanova
15 UNC Asheville

Kansas ✓
Colorado ✗
Maryland ✓
California ✗
Arizona ✗
Miami (FL) ✓
Iowa ✓
Villanova ✓

Kansas ✓
Maryland ✓
Miami (FL) ✓
Villanova ✓

Kansas ✓
Villanova ✓

Kansas ✗

WEST REGION

1 Oregon
16 Cross/Southern
8 St Joe's
9 Cincinnati
5 Baylor
12 Yale
4 Duke
13 UNC Wilmington
6 Texas
11 Northern Iowa
3 Texas A&M
14 Green Bay
7 Oregon State
10 VCU
2 Oklahoma
15 Cal St Baker

Oregon ✓
St Joe's ✓
Baylor ✗
Duke ✓
Texas ✗
Texas A&M ✓
Oregon State ✗
Oklahoma ✓

Oregon ✓
Duke ✓
Texas A&M ✓
Oklahoma ✓

Oregon ✓
Oklahoma ✓

Oregon ✗

EAST REGION

North Carolina 1
FGCU/Fairleigh 16
USC 8
Providence 9
Indiana 5
Chattanooga 12
Kentucky 4
Stony Brook 13
Notre Dame 6
Michigan/Tulsa 11
West Virginia 3
S.F. Austin 14
Wisconsin 7
Pittsburgh 10
Xavier 2
Weber St. 15

✓North Carolina
✗ USC
✓ Indiana
✓ Kentucky
✓ Notre Dame
✗West Virginia
✓ Wisconsin
✓ Xavier

✓ North Carolina
✗ Kentucky
✗West Virginia
✗ Xavier

✓North Carolina
✗Xavier

✓North Carolina

MIDWEST REGION

Virginia 1
Hampton 16
Texas Tech 8
Butler 9
Purdue 5
Little Rock 12
Iowa St. 4
Iona 13
Seton Hall 6
Gonzaga 11
Utah 3
Fresno St. 14
Dayton 7
Syracuse 10
Michigan State 2
Middle Tenn. 15

✓ Virginia
✗Texas Tech
✗ Purdue
✓ Iowa St.
✗Seton Hall
✓ Utah
✗ Dayton
✗Michigan State

✓ Virginia
✓ Iowa St.
✗ Utah
✗Michigan State

✓ Virginia
✗Michigan State

✗Virginia

NATIONAL CHAMPIONSHIP
April 4

Oregon ✗ | Oregon ✗ | North Carolina ✓

March 15-16
FIRST FOUR
DAYTON, OH

Vanderbilt
Wichita St ✓

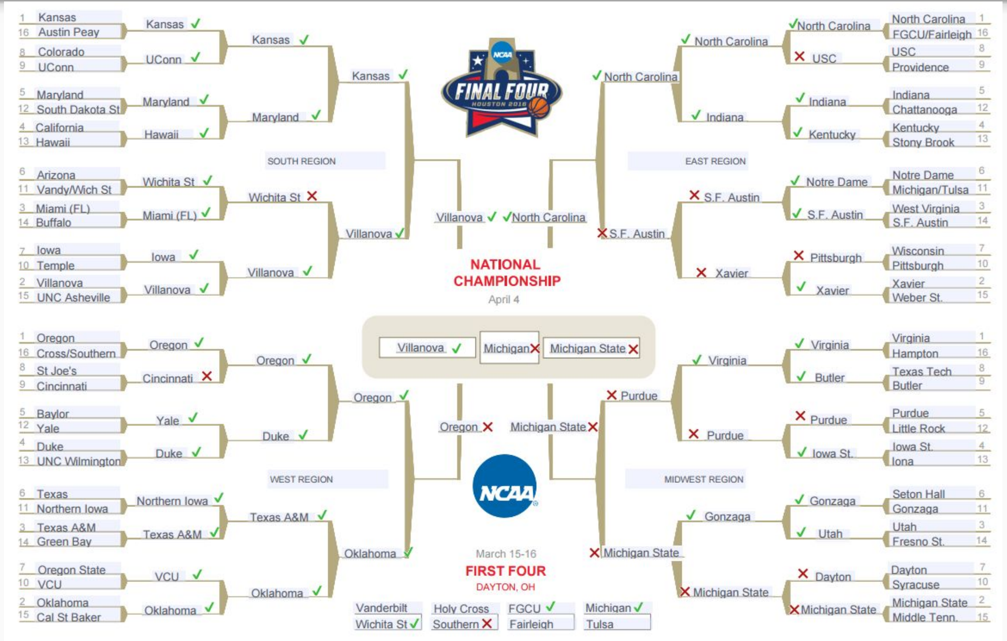
Holy Cross ✓
Southern

FGCU ✓
Fairleigh

Michigan ✓
Tulsa

✓ : 42
✗ : 25

LogisticRegression Bracket



✓ : 48

✗ : 19

Conclusions

- Hit my goal of doing better than a chalk bracket… this time
- Derived stats seem to do a good job, would really like to explore more features
- Evaluation of end results proved much more difficult than anticipated

Next Steps

- Addition of more advanced statistics
- Addition of location data or home/away biases
- End to end evaluation for creating brackets across seasons
- Ensemble modeling