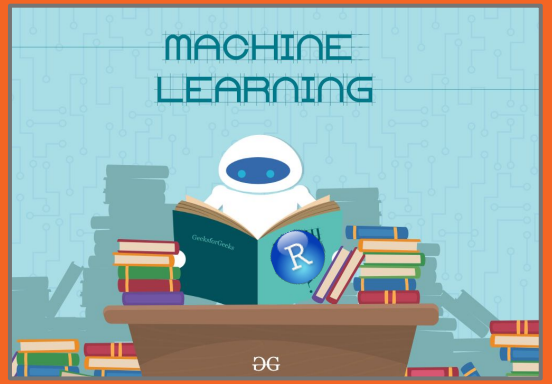
Noam Bendavid
Michael Model
Jessica Needleman
Nathan O'Hara

Bracketology: Man Vs. Machine



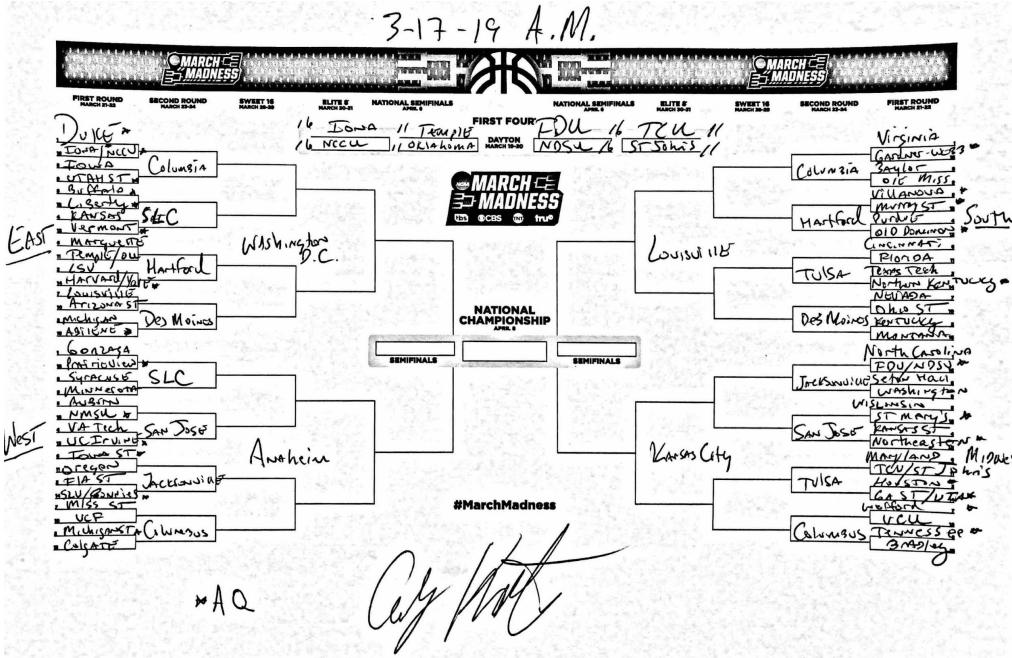
What is bracketology?

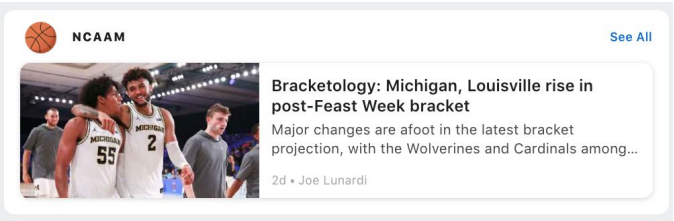
- “The activity of predicting the participants in and outcomes of the games in a sports tournament, especially the NCAA college basketball tournament” (Oxford)
- The NCAA tournament comprises 68 teams. The 32 conference tournament winners and 36 *at-large* teams
- Seeding and at-large bids are chosen by the selection committee. The bracket is announced on Selection Sunday
- There is no clear cut formula the selection committee uses, which is why bracketology has become popular.

DANIEL WILCO | NCAA.COM | MARCH 17, 2019

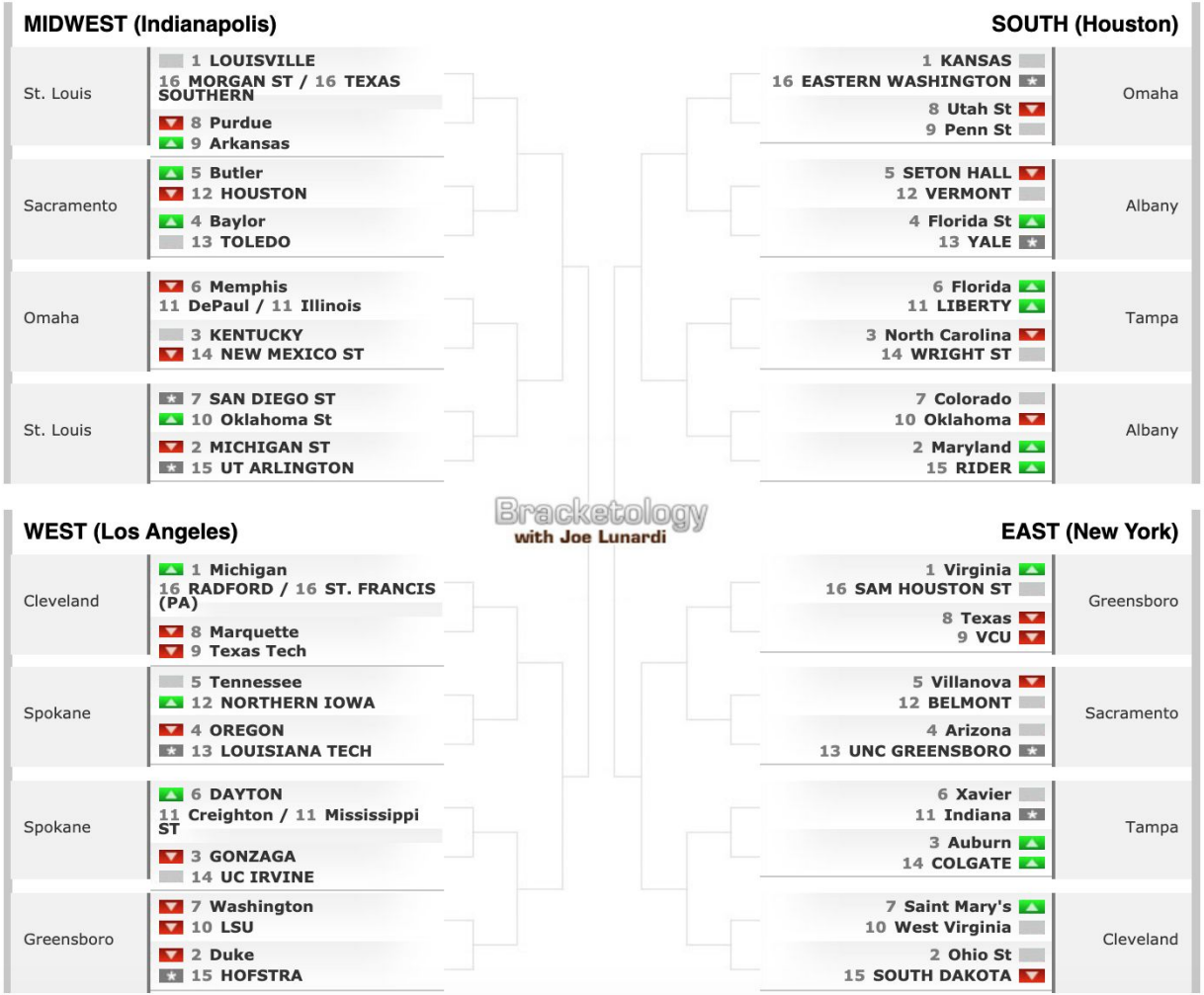
The entire March Madness field of 68 predicted on Selection Sunday

"Our own NCAA.com basketball analyst Andy Katz has traveled the country all season long. He's watched games. He's talked to the best players and coaches. And through it all he's sifted through the mountain of information he's collected to put together full bracket predictions for the March Madness field."





Bracketology does not only occur in March. Joe Lunardi constantly updates his bracket prediction and it's only December….



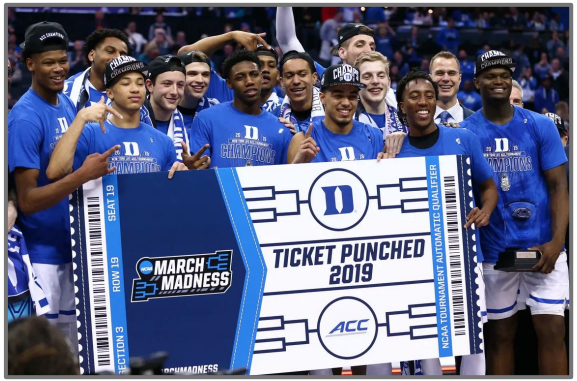
Who makes the tournament?

- *S*electing the 36 *at-large* teams is a challenge. Due to the talent disparity between conferences the committee cannot pick based on record alone

Conference tiers

- *Power 5:* ~5-10 teams per conference (e.g. ACC, Big Ten, Big 12)
- *High-major:* ~2-4 teams per conference (e.g. A-10, AAC)
- *Mid-major:* 1 team per conference, a second bid is extremely rare (e.g. OVC, MVC, A-Sun)

The Problem



Goal: Build a predictive model competitive with bracketologists using just the data rather than inside information and expertise

- **Question 1:** Which 36 teams will earn *at-large* bids?
- **Question 2:** How will the 68 selected teams be seeded?

We hoped to answer both these questions using team, player and schedule data through the end of the conference tournaments.

The Data

The Data

TEAMRANKINGS
Sports Picks, Predictions & Data

Strength of schedule
- Overall and non-conference rating
- Rating against teams of certain ranks

sportradar

Regular season/conference tournament play statistics
- Team averages per game
- Opponent averages per game
- Overall record

BASKETBALL REFERENCE

Conference wins
- Regular season wins
- Tournament performance

The Washington Post
Democracy Dies in Darkness

Historical NCAA Tournament info
- Number of appearances
- Past performance

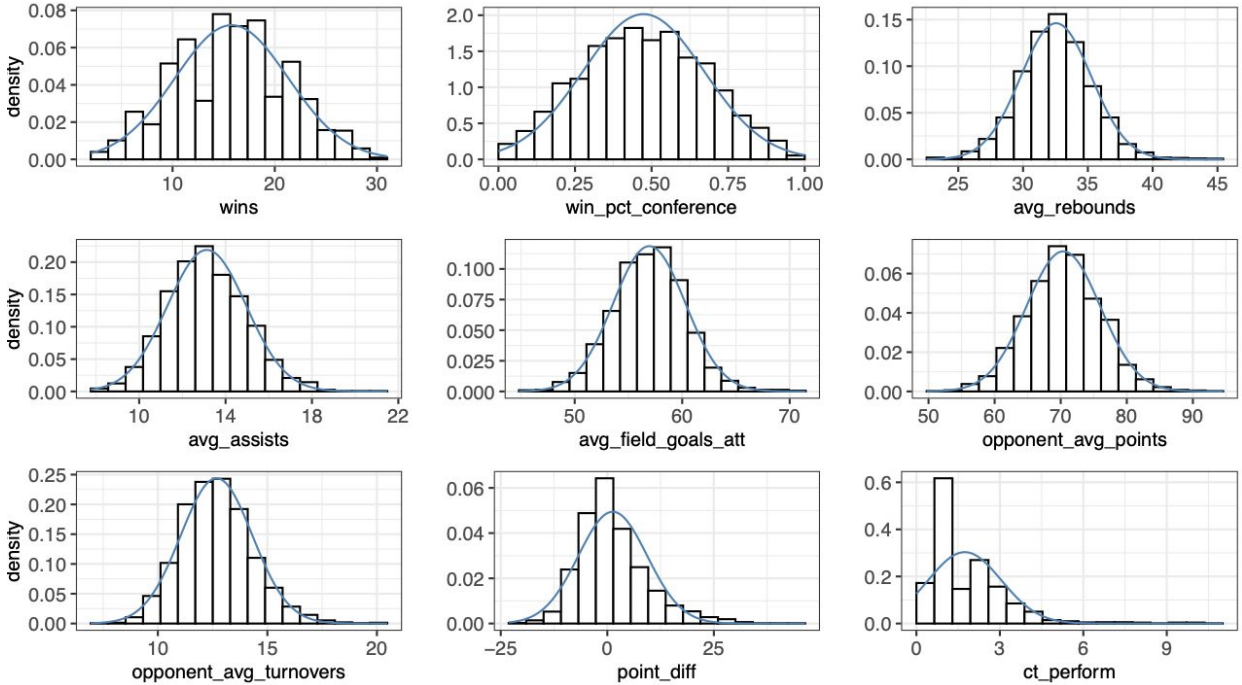
The Data

- **64** total predictors
- **6** years of data
  - (since the conference began including 68 teams)

- **Train models** on data from 2014-2018
- **Evaluate** on data from 2019
- **Compare** to bracketologists' 2019 predictions

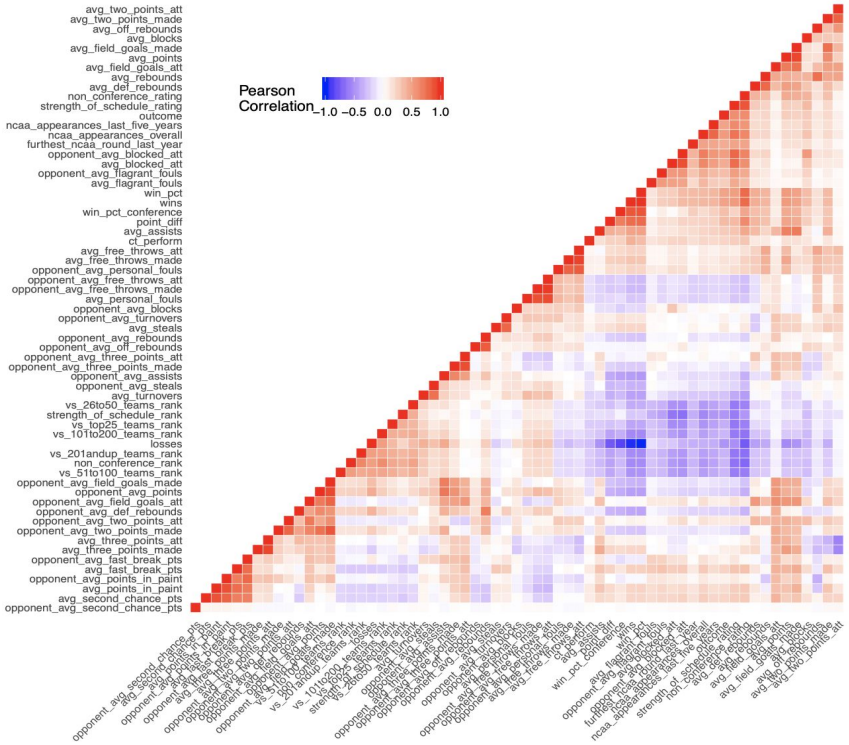
Exploratory Data Analysis

1. **Predictor distributions**
2. Multicollinearity problem
3. Conference interactions
4. Principal Component Analysis



Exploratory Data Analysis

1. Predictor distributions
2. **Multicollinearity problem**
3. Conference interactions
4. Principal Component Analysis



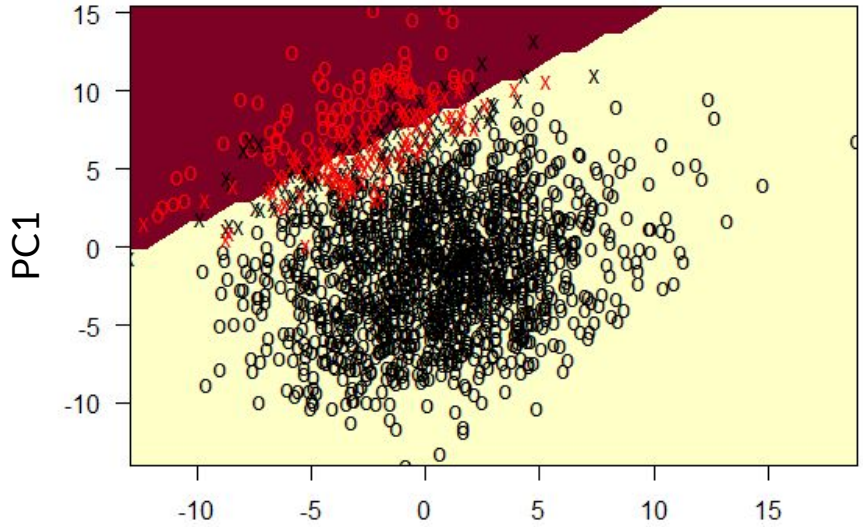
Exploratory Data Analysis

1. Predictor distributions
2. Multicollinearity problem
3. **Conference interactions**
4. Principal Component Analysis

Exploratory Data Analysis

Do the tournament and non-tournament teams seem to separate by the values in our data?

1. Predictor distributions
2. Multicollinearity problem
3. Conference interactions
4. **Principal Component Analysis**



Principal Component Loadings

variable	loading
losses	-0.1340435044
non_conference_rank	-0.1154087305
vs_201andup_teams_rank	-0.1104453423
strength_of_schedule_rank	-0.0929716122
...	...
wins	0.1507909666
non_conference_rating	0.1575496279
total_field_goals_made	0.1681143684
total_points	0.1704113470

PC1

- Teams with a high value score a lot and win a lot.
- Teams with a low value lose a lot, and to weak teams.

- This can be seen as separating **big winners** from **big losers**.

- Along this principal component, the highest percentage of variation in x (feature space) is explained

Principal Component Loadings

variable	loading
win_pct	-0.1168062823
non_conference_rating	-0.1160469526
strength_of_schedule_rating	-0.0961521004
point_diff	-0.0925059645
...	...
opponent_avg_field_goals_att	0.1501689828
opponent_avg_def_rebounds	0.1667115366
opponent_total_assists	0.1755140597
opponent_avg_points	0.2283645527

PC2

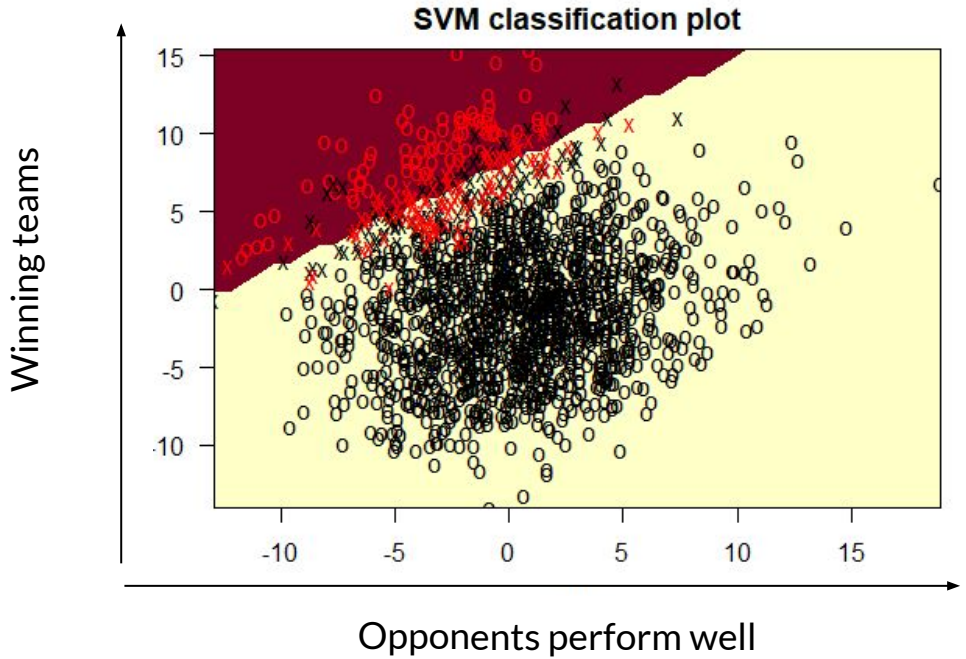
- Teams with a high value get beat down by their opponents.
- Teams with a low value beat their opponents down.

- This can be seen as separating teams primarily based on **opponents' performances**.

- Along this principal component, the second-highest percentage of variation in x (feature space) is explained

Principal Component Analysis

Tournament teams **win more** and their **opponents** post **less-impressive stats**.



Field Predictions

Field Prediction Methodology

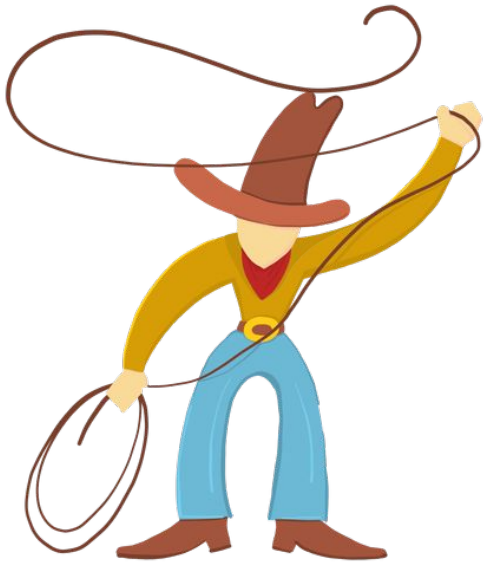
Goal: Correctly predict at-large tournament bids.

- **Assessment**: Classification true positive rate on 2019 tournament data.

Challenges: High dimensionality, potential multicollinearity, not all predictors necessarily active.

- **Solution**: Find justifiable models under these challenges.

Field Prediction Methodology



1. Lasso regression
2. Principal component regression
3. Support vector classifier
4. Trees

Field Prediction Methodology

	True Positive Rate	Bubble Team Accuracy
Lasso Regression, No Interactions	0.9167	0.750
Lasso Regression, Conference Interactions	0.9444	0.875
Lasso Regression, All Interactions	0.9444	0.875
Principle Component Regression	0.8056	0.500
Support Vector Machine	0.8056	0.500
Random Forest Model	0.9167	0.750
Boosting Model	0.9167	0.625

Field Prediction Methodology

	True Positive Rate	Bubble Team Accuracy
Lasso Regression, No Interactions	0.9167	0.750
Lasso Regression, Conference Interactions	0.9444	0.875
Lasso Regression, All Interactions	0.9444	0.875
Principle Component Regression	0.8056	0.500
Support Vector Machine	0.8056	0.500
Random Forest Model	0.9167	0.750
Boosting Model	0.9167	0.625

Field Prediction Significant Predictors

- Wins, Conference Win Percentage, Strength of Schedule
- ConferenceCUSA*WinPercentageConference, ConferenceMVC*WinPercentageConference
- ConferenceBig12*PointDiff, ConferenceBigEast*PointDiff
- ConferenceBigEast*CTPerform, ConferenceACC*CTPerform

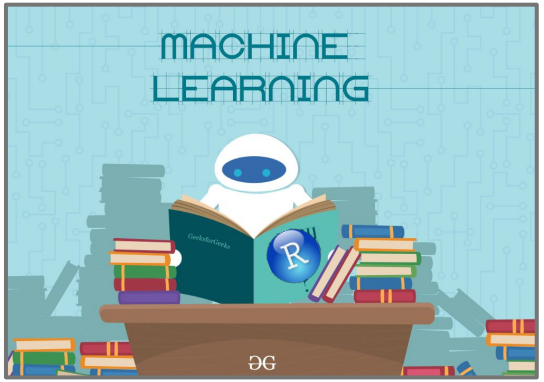
Field Prediction Results

Team	Likelihood	Rank
Virginia	0.9997	1
North Carolina	0.9993	2
Tennessee	0.9986	3
Gonzaga	0.9985	4
Michigan	0.9976	5
Marquette	0.9972	6
Kentucky	0.9967	7
Kansas State	0.9959	8
Temple	0.6438	33
Clemson	0.6131	34
Ohio State	0.5682	35
Oklahoma	0.5515	36
Providence	0.4759	37
Nebraska	0.4741	38
Indiana	0.4537	39
TCU	0.4408	40
UCF	0.4103	41
North Carolina State	0.8479	25
Lipscomb	0.1470	51
Dayton	0.1408	52
UNC Greensboro	0.1252	53
Furman	0.0934	58
Belmont	0.0888	59

Field Prediction Results

Andy Katz
Teams correctly guessed: 35/36
Missed: Belmont in, TCU out

Joe Lunardi
Teams correctly guessed: 35/36
Missed: Belmont in, TCU out



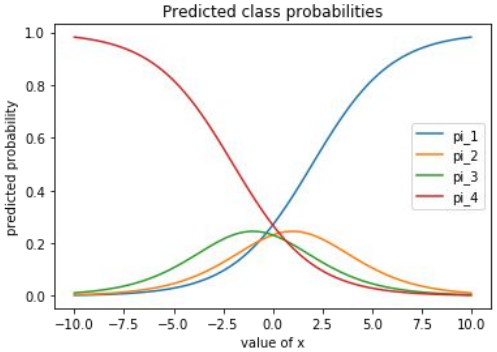
Our Model
Teams correctly guessed: 34/36
Missed: Belmont in, UCF in, NC State out, Clemson out

Seeding Predictions

Seeding Prediction Methodology

1. Cumulative logit model
    a. Principal components
    b. ordinalNet

Model Justification



- High dimensional data with a 16-category, ordinal response
- **Cumulative logit model**: assumes proportional odds to perform ordinal classification
  - Difference between seeds relatively consistent
- **ordinalNet** package allows us to apply lasso regularization - combats multicollinearity and the curse of dimensionality
- Principal component predictors are another possible solution

Additional Model Constraints

The **cumulative logit model** predicts class probabilities.

Typically, we predict the class **with the highest probability**.

In our case there can only be **4 teams per class**. (sometimes 6)

Instead, we take the top four teams by probability for each class **iteratively**, excluding the teams already chosen.

- Results in four teams predicted for each seed.

Seeding Prediction Results

	truth															
predict	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	3	1	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0
7	0	0	0	0	0	2	1	0	1	0	0	0	0	0	0	0
8	0	0	0	0	0	0	1	0	2	1	0	0	0	0	0	0
9	0	0	0	0	0	0	0	3	1	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	1	0	2	1	0	0	0	0	0
11	0	0	0	0	0	0	1	0	0	0	4	1	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	1	1	2	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	1	2	1	0	0
14	0	0	0	0	0	0	0	0	0	0	0	1	0	2	1	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	1
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	5

Best model: ordinalNet (cumulative logit model with elastic net penalty)

Results on our 2019 Field Predictions:

Correct classification rate: 47.1%

- vs. 18% for principal component model

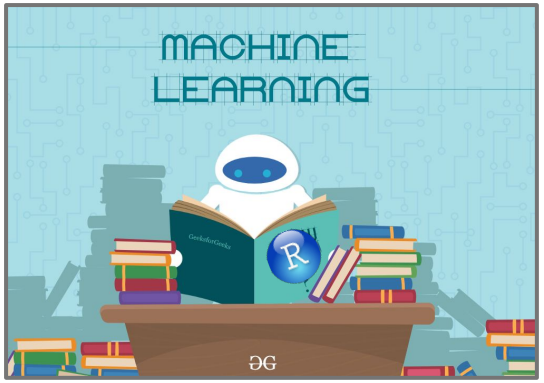
Percent correct within 1 seed: 86.8%

- Most incorrect: in 6-11 range
  - Consistent with human analysts

Seeding Prediction Results

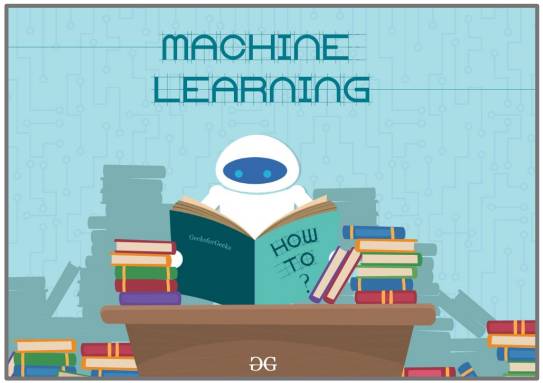
Andy Katz
Correct classification rate: 64.7%
CCR within 1: 91.2%

Joe Lunardi
Correct classification rate: 72.1%
CCR within 1: 95.6%



Our Model
Correct classification rate: 47.1%
CCR within 1: **86.8%**

Seeding Prediction Results (cont'd)



Originally, predicted on our (slightly incorrect) field predictions.

When predicting on the actual field:

Correct classification rate: 50%

CCR within 1 seed: 91.2%

Team.Name	Actual	Predicted
Duke	1	1
North Carolina	1	1
Virginia	1	1
Tennessee	2	1
Gonzaga	1	2
Kentucky	2	2
Michigan	2	2
Michigan State	2	2
LSU	3	3
Texas Tech	3	3
Florida State	4	3
Kansas	4	3
Houston	3	4
Purdue	3	4
Kansas State	4	4
Virginia Tech	4	4

Significant Predictors: Seeding

50 predictors were found significant by the ordinalNet model

- **14** interactions with `midmajor`
  - All negative, except `ct_perform`
- **36** main effects
  - Spanning all categories of data (team performance statistics, win/loss, SOS, NCAA past appearances)

Limitations

- Geographic / contextual factors contribute to seeding
  - Analysts can speculate on this, however it will not show in the data, which measures team performance
- While the above favor the human analysts, our model does comparably well **within one seed** of the truth
- Difficult computationally to include all interactions
  - Only interacted with a bucketed `midmajor` conference indicator variable.

Conference Tournament Performance Analysis

Conference Tournament Performance Analysis

- How much can a good performance in your conference tournament impact your chances of making the NCAA tournament?
- Or affect tournament seeding?

- How can you best **quantify** conference tournament performance? Beyond CT games played?

Conference Tournament Performance Analysis

`ct_perform:` our original metric for conference tournament performance

Conference Tournament Performance Analysis

`ct_perform:` our original metric for conference tournament performance

Strength of Schedule Difference =

(Strength of schedule on Selection Sunday - Strength of schedule the day before conference tournament)

Conference Tournament Performance Analysis

`ct_perform:` our original metric for conference tournament performance

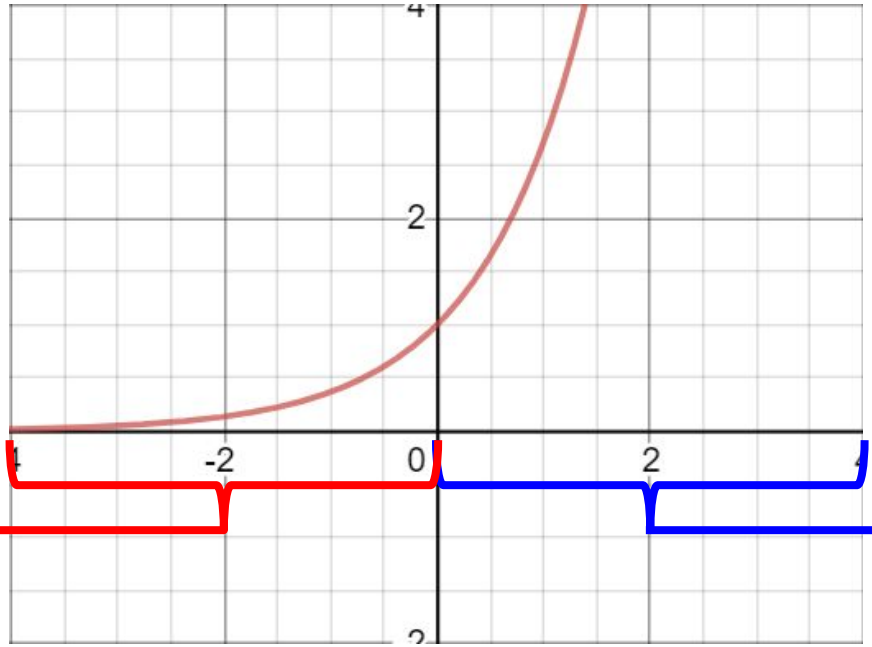
Strength of Schedule Difference =

(Strength of schedule on Selection Sunday - Strength of schedule the day before conference tournament)

`ct_perform =`

$$e^{\text{Strength of Schedule Difference}} * \#\text{ of games played in C.T.}$$

Exponential Function



Decrease in SOS: fractional multiplier

Increase in SOS: multiplier > 1

$e^{\text{Strength of Schedule Difference}} * \#\text{ of games played in C.T.}$

Example: 2019 ACC Conference Tournament

#10 Georgia Tech

- Played one game against **#15 Notre Dame**, and lost
- `ct_perform` = 0.904

#8 NC State

- Played two games against **#9 Clemson**, and **#1 Virginia**
- `ct_perform` = 3.297

Conference Tournament Performance Analysis

- `ct_perform` was found **not** to be significant or important *as a main effect* in any of our models for **field prediction**

- It *was* found to significantly affect probability of an at-large bid when interacted with:
  - `conferenceACC`
  - `conferenceBIGEAST`

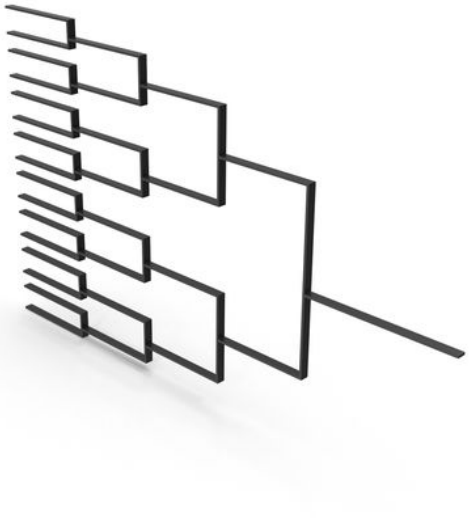
- **Interpretation**: conference tournament performances were found significant **only** in the ACC and Big East conferences for at-large bids.

Conference Tournament Performance Analysis

- `ct_perform` was found **not** to be significant *as a main effect* in our model for seeding.

- It was found to significantly affect probability of an at-large bid when interacted with:
  - `midmajor` (indicator variable)
    - Sign of coefficient: **positive**

- **Interpretation**: conference tournament performance significantly increases the likelihood of a higher seed, but only in mid-major conferences.



Conclusion

Key Takeaways

- Given the fact that our classification rate is similar to the experts, we imagine that the key factors included in our model adequately capture the committee's guidelines
- Given the expertise of the bracketologists and a lack of other contextual information, we can presume a model won't end up doing better than the bracketologist, though it can get pretty close

Limitations

- There are numerous minute details the committee takes into account in seeding that our model did not. These include:
  - Teams cannot play in a region they host
  - The committee tends to avoid rematches from regular season meetings
  - Opening weekend matchups between teams in the same conference are avoided
- The committee's guiding metric recently changed, from RPI to the NET. Those emphasize slightly different components of teams and could cause predictive error.

Q&A