Finding Clarity in the Madness

An Introductory Look at Sports Analytics

Jonathan Ashbrock

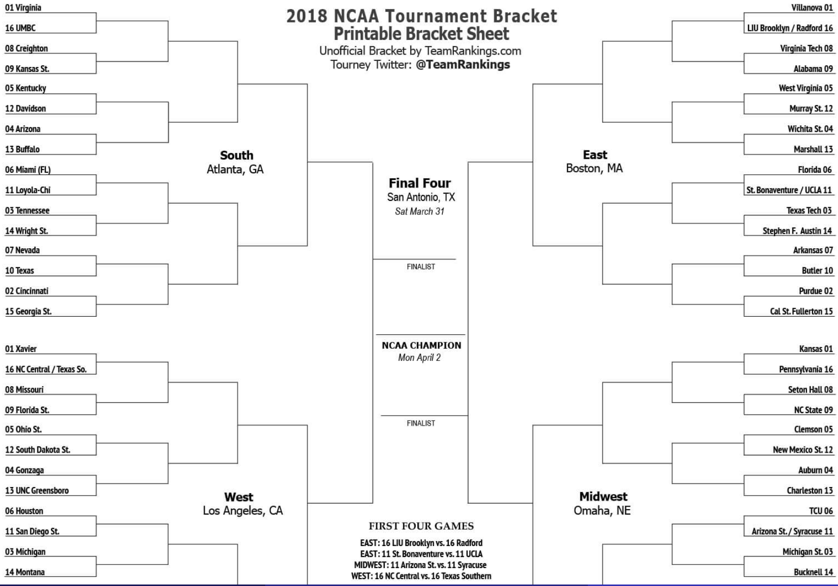
Vanderbilt University

April 3, 2018

Outline

1. **Problem Statement**
2. Initial Data and Creating Meaningful Data
3. The Models We Will Use
4. Analysis and Model Performance

The Setting


2018 NCAA Tournament Bracket
Printable Bracket Sheet
Unofficial Bracket by TeamRankings.com
Tourney Twitter: @TeamRankings
01 Virginia
16 UMBC
08 Creighton
09 Kansas St.
05 Kentucky
12 Davidson
04 Arizona
13 Buffalo
06 Miami (FL)
11 Loyola-Chi
03 Tennessee
14 Wright St.
07 Nevada
10 Texas
02 Cincinnati
15 Georgia St.
South
Atlanta, GA
01 Xavier
16 NC Central / Texas So.
08 Missouri
09 Florida St.
05 Ohio St.
12 South Dakota St.
04 Gonzaga
13 UNC Greensboro
06 Houston
11 San Diego St.
03 Michigan
14 Montana
West
Los Angeles, CA
Villanova 01
LIU Brooklyn / Radford 16
Virginia Tech 08
Alabama 09
West Virginia 05
Murray St. 12
Wichita St. 04
Marshall 13
Florida 06
St. Bonaventure / UCLA 11
Texas Tech 03
Stephen F. Austin 14
Arkansas 07
Butler 10
Purdue 02
Cal St. Fullerton 15
East
Boston, MA
Kansas 01
Pennsylvania 16
Seton Hall 08
NC State 09
Clemson 05
New Mexico St. 12
Auburn 04
Charleston 13
TCU 06
Arizona St. / Syracuse 11
Michigan St. 03
Bucknell 14
Midwest
Omaha, NE
Final Four
San Antonio, TX
Sat March 31
FINALIST
NCAA CHAMPION
Mon April 2
FINALIST
FIRST FOUR GAMES
EAST: 16 LIU Brooklyn vs. 16 Radford
EAST: 11 St. Bonaventure vs. 11 UCLA
MIDWEST: 11 Arizona St. vs. 11 Syracuse
WEST: 16 NC Central vs. 16 Texas Southern

Problem Statement

We want to answer the following questions:

1. How well does a team's seed predict their performance?
2. Using 'more advanced' data, can we make a better predictor than a team's seed?
3. Is the tournament any more unpredictable than the regular season?
4. Can statistics help you win your bracket pool?

Outline

1. Problem Statement
2. **Initial Data and Creating Meaningful Data**
3. The Models We Will Use
4. Analysis and Model Performance

The Given Data

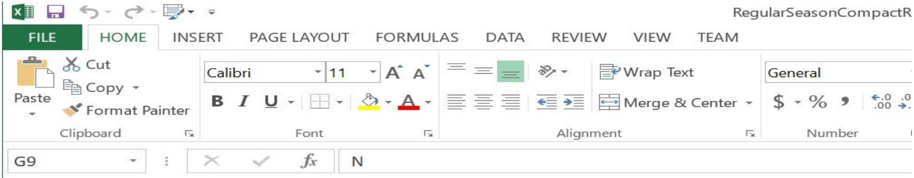
The data is from Kaggle: https://www.kaggle.com/c/mens-machine-learning-competition-2018/data

We are provided:

1. Score, location, teams, and date for every game since 1984
2. Results from tournament games including seeds, scores, and locations for every game since 1984

What the file Looks Like

150,685 rows which look like this:



	A	B	C	D	E	F	G	H	I	J
1	Season	Daynum	Wteam	Wscore	Lteam	Lscore	Wloc	Numot		
2	1985	20	1228	81	1328	64	N	0		
3	1985	25	1106	77	1354	70	H	0		
4	1985	25	1112	63	1223	56	H	0		
5	1985	25	1165	70	1432	54	H	0		
6	1985	25	1192	86	1447	74	H	0		
7	1985	25	1218	79	1337	78	H	0		
8	1985	25	1228	64	1226	44	N	0		
9	1985	25	1242	58	1268	56	N	0		
10	1985	25	1260	98	1133	80	H	0		
11	1985	25	1305	97	1424	89	H	0		
12	1985	25	1307	103	1288	71	H	0		
13	1985	25	1344	75	1438	71	N	0		
14	1985	25	1374	91	1411	72	H	0		
15	1985	25	1412	70	1397	65	N	0		
16	1985	25	1417	87	1225	58	H	0		
17	1985	26	1116	65	1368	62	H	0		
18	1985	26	1120	92	1391	50	H	0		

Computing Meaningful Data

We start with this:

G9 | N

	A	B	C	D	E	F	G	H
1	Season	Daynum	Wteam	Wscore	Lteam	Lscore	Wloc	Numot
2	1985	20	1228	81	1328	64	N	0
3	1985	25	1106	77	1354	70	H	0
4	1985	25	1112	63	1223	56	H	0
5	1985	25	1165	70	1432	54	H	0
6	1985	25	1192	86	1447	74	H	0

And get to this:

O10

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Wscore	Lscore	Winning_team	Team1	Team2	Game_num1	Game_num2	Avg_score1	Avg_score2	Avg_allow1	Avg_allow2	dif	Team1_wins	Team2_wins
2	70	53	2	1157	1104	1	1	0	0	0	0	-17	0	0
3	87	81	2	1336	1107	1	1	0	0	0	0	-6	0	0
4	65	63	1	1112	1277	1	1	0	0	0	0	2	0	0
5	88	70	2	1340	1113	1	1	0	0	0	0	-18	0	0
6	92	83	2	1236	1116	1	1	0	0	0	0	-9	0	0

Computing Meaningful Data

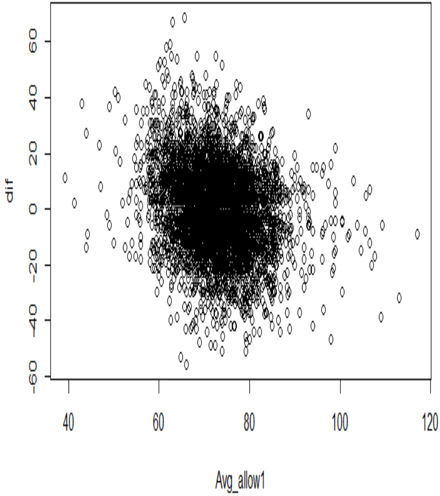
The full data set:

O10

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Wscore	Lscore	Winning_team	Team1	Team2	Game_num1	Game_num2	Avg_score1	Avg_score2	Avg_allow1	Avg_allow2	dif	Team1_wins	Team2_wins
2	70	53	2	1157	1104	1	1	0	0	0	0	-17	0	0
3	87	81	2	1336	1107	1	1	0	0	0	0	-6	0	0
4	65	63	1	1112	1277	1	1	0	0	0	0	2	0	0
5	88	70	2	1340	1113	1	1	0	0	0	0	-18	0	0
6	92	83	2	1236	1116	1	1	0	0	0	0	-9	0	0
⋮														
5387	56	53	2	1107	1436	33	33	71.59375	73.9375	66.5625	63.15625	-3	20	27
5388	74	60	2	1166	1437	33	34	82.21875	77.7878788	72.5625	62.87878788	-14	24	30
5389	76	48	1	1458	1321	33	33	72.3125	72	61.53125	64.59375	28	24	0
5390	73	71	1	1463	1217	27	26	73.7307692	71.72	70.7692308	66.92	2	16	0
5391	82	65	2	1116	1246	34	34	80.2424242	86.0606061	73.7575758	71.78787879	-17	25	28
5392	71	56	2	1458	1276	34	35	72.4242424	74.9117647	61.1212121	65.82352941	-15	25	23
5393	71	59	2	1463	1343	28	28	73.7037037	70.8148148	70.7777778	62.14814815	-12	17	21
5394	70	63	2	1433	1348	34	33	75.030303	73.46875	66.3030303	64.9375	-7	26	23
5395	71	56	2	1153	1374	34	34	75.030303	74.6363636	60.5151515	59.93939394	-15	29	29
5396	59	53	2	1402	1407	31	34	67.1333333	76.2424242	65.0666667	72.6969697	-6	18	19

Why were these changes made?

Visualizing the Noise



Outline

1. Problem Statement
2. Initial Data and Creating Meaningful Data
3. **The Models We Will Use**
4. Analysis and Model Performance

The Dependent and Independent Variables

Recall: Want to predict game winners. There are two approaches

1. Predict game winner (Categorical or Discrete, "Team1" or "Team2")

The Dependent and Independent Variables

Recall: Want to predict game winners. There are two approaches

1. Predict game winner (Categorical or Discrete, "Team1" or "Team2")
2. Predict difference (Continuous, Team1 Score-Team2 Score=dif $\in \mathbb{R}$)

The Dependent and Independent Variables

Recall: Want to predict game winners. There are two approaches

1. Predict game winner (Categorical or Discrete, "Team1" or "Team2")
2. Predict difference (Continuous, Team1 Score-Team2 Score=dif $\in \mathbb{R}$)

Using:

1. Team1 average scored/allowed (Continuous)
2. Team2 average scored/allowed (Continuous)

The Dependent and Independent Variables

Recall: Want to predict game winners. There are two approaches

1. Predict game winner (Categorical or Discrete, "Team1" or "Team2")
2. Predict difference (Continuous, Team1 Score-Team2 Score=dif $\in \mathbb{R}$)

Using:

1. Team1 average scored/allowed (Continuous)
2. Team2 average scored/allowed (Continuous)
3. Home team (Discrete)
4. Seeding of Team (Discrete)

The Theory of Regression

- Used for continuous variables predicting continuous variables
- The so-called "line of best fit"

The Theory of Regression

- Used for continuous variables predicting continuous variables
- The so-called "line of best fit"
- The line minimizes the vertical distance between points and the line

The Theory of Regression

- Used for continuous variables predicting continuous variables
- The so-called "line of best fit"
- The line minimizes the vertical distance between points and the line
- If we have points $\{(x_i, y_i)\}_{i=1}^{n}$ and a line $y = mx + b$, the regression line solves the minimization problem:

The Theory of Regression

- Used for continuous variables predicting continuous variables
- The so-called "line of best fit"
- The line minimizes the vertical distance between points and the line
- If we have points $\{(x_i, y_i)\}_{i=1}^{n}$ and a line $y = mx + b$, the regression line solves the minimization problem:

$$\min_{m,b} \sum_{i=1}^{n} (y_i - (m * x_i + b))^2$$

The Theory of Regression

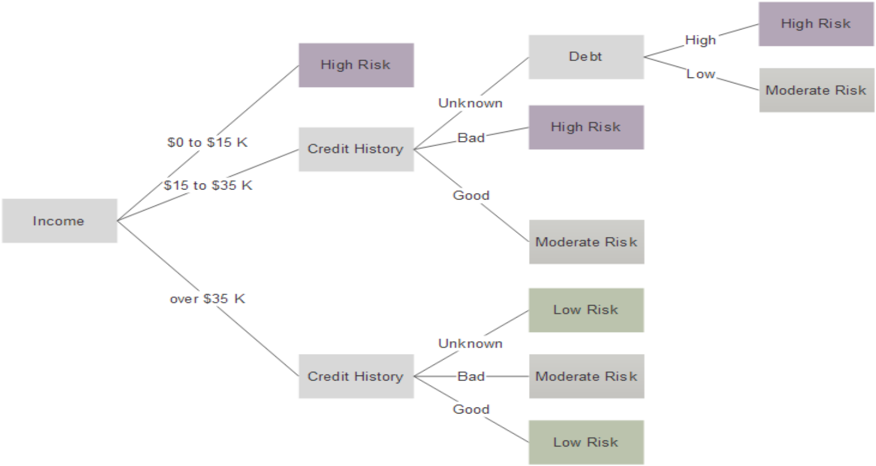
- Used for continuous variables predicting continuous variables
- The so-called "line of best fit"
- The line minimizes the vertical distance between points and the line
- If we have points $\{(x_i, y_i)\}_{i=1}^{n}$ and a line $y = mx + b$, the regression line solves the minimization problem:

$$\min_{m,b} \sum_{i=1}^{n} (y_i - (m * x_i + b))^2$$

- Can also do multivariable regression when x_i contains more than 1 variable. The "line" then is: $y = m_1 x_1 + m_2 x2 + \cdots + m_k x_k + b$.

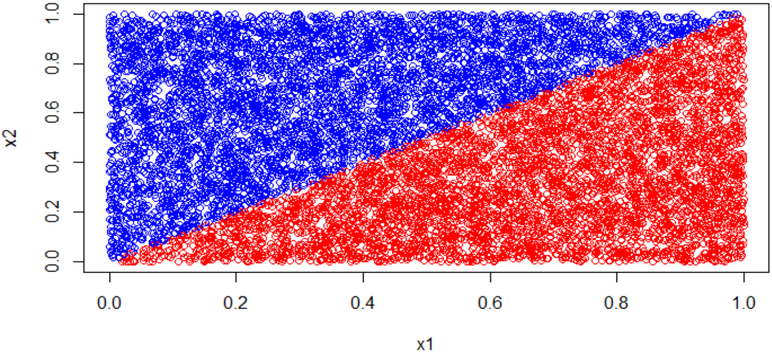
Decision Trees

- Used for continuous and/or discrete variables predicting discrete variables



Decision Trees Have a Problem

- If we notice, decision trees consider one variable at a time
- Consider the following classification example: Points in the plane with $y > x$ are called "blue" and those with $x \leq y$ are called "red":



Decision Trees Have a Problem

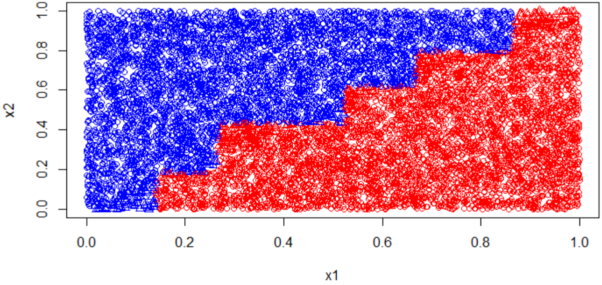
- The following is the classification we get when classifying points into red/blue using a decision tree:

Decision Trees Have a Problem

- The following is the classification we get when classifying points into red/blue using a decision tree:
- Since Decision trees only consider one variable at a time, we only "split" the data parallel to one of the two axes

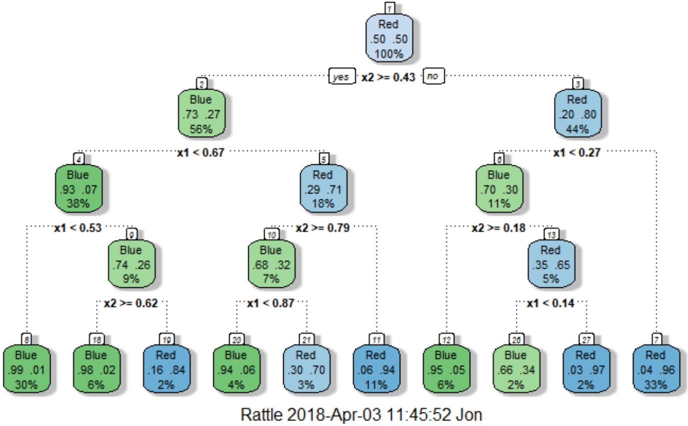
Decision Trees Have a Problem

- The following is the classification we get when classifying points into red/blue using a decision tree:
- Since Decision trees only consider one variable at a time, we only "split" the data parallel to one of the two axes



Decision Trees Have a Problem

- And here is the decision tree:

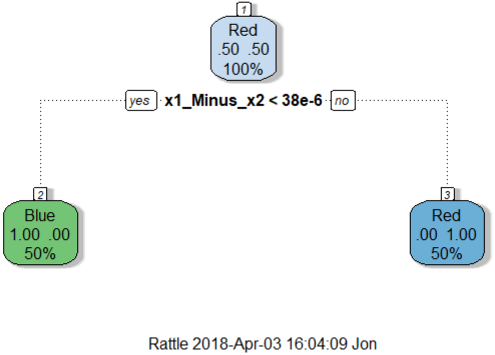


Pre-Processing Makes Decision Trees Better

- Doing one pre-processing step makes our tree simpler and 100% accurate.

Pre-Processing Makes Decision Trees Better

- Doing one pre-processing step makes our tree simpler and 100% accurate.



Plan for Our Predictor

Taking into account the limitations of decision trees, our Predictor will do 2 things:

Plan for Our Predictor

Taking into account the limitations of decision trees, our Predictor will do 2 things:

1. Use a regression model on the continuous variables to predict the point differential (Team1-Team2=dif)

Plan for Our Predictor

Taking into account the limitations of decision trees, our Predictor will do 2 things:

1. Use a regression model on the continuous variables to predict the point differential (Team1-Team2=dif)
   - This takes care of the problem discussed in decision trees. This "combines" the continuous variables into 1 value.

Plan for Our Predictor

Taking into account the limitations of decision trees, our Predictor will do 2 things:

1. Use a regression model on the continuous variables to predict the point differential (Team1-Team2=dif)
   - This takes care of the problem discussed in decision trees. This "combines" the continuous variables into 1 value.
2. Use a decision tree to predict the winner using:

Plan for Our Predictor

Taking into account the limitations of decision trees, our Predictor will do 2 things:

1. Use a regression model on the continuous variables to predict the point differential (Team1-Team2=dif)
    - This takes care of the problem discussed in decision trees. This "combines" the continuous variables into 1 value.
2. Use a decision tree to predict the winner using:
    1. The output of the regression

Plan for Our Predictor

Taking into account the limitations of decision trees, our Predictor will do 2 things:

1. Use a regression model on the continuous variables to predict the point differential (Team1-Team2=dif)
   - This takes care of the problem discussed in decision trees. This "combines" the continuous variables into 1 value.
2. Use a decision tree to predict the winner using:
   1. The output of the regression
   2. Location of game

Regression Model Summary

The computed regression model to predict the point differential is:

Regression Model Summary

The computed regression model to predict the point differential is:

$$dif = 0.725(x_1) - 0.693(x_2) - 0.627(x_3) + 0.710(x_4) - 8.332$$

Where,

Regression Model Summary

The computed regression model to predict the point differential is:

$$dif = 0.725(x_1) - 0.693(x_2) - 0.627(x_3) + 0.710(x_4) - 8.332$$

Where,

- dif = Team1 Score - Team2 Score
- x_1 is Team1 average points scored
- x_2 is Team1 average points allowed

Regression Model Summary

The computed regression model to predict the point differential is:

$$dif = 0.725(x_1) - 0.693(x_2) - 0.627(x_3) + 0.710(x_4) - 8.332$$

Where,

- dif = Team1 Score - Team2 Score
- x_1 is Team1 average points scored
- x_2 is Team1 average points allowed
- x_3 is Team2 average points scored
- x_4 is Team2 average points allowed

Regression Model Summary

The computed regression model to predict the point differential is:

$$dif = 0.725(x_1) - 0.693(x_2) - 0.627(x_3) + 0.710(x_4) - 8.332$$

Where,

- dif = Team1 Score - Team2 Score
- x_1 is Team1 average points scored
- x_2 is Team1 average points allowed
- x_3 is Team2 average points scored
- x_4 is Team2 average points allowed

Whenever you get a model, stop and think:
Do we think this model makes sense?

Regression Model Summary

The computed regression model to predict the point differential is:

$$dif = 0.725(x_1) - 0.693(x_2) - 0.627(x_3) + 0.710(x_4) - 8.332$$

Where,

- dif = Team1 Score - Team2 Score
- x_1 is Team1 average points scored
- x_2 is Team1 average points allowed
- x_3 is Team2 average points scored
- x_4 is Team2 average points allowed

Whenever you get a model, stop and think:
Do we think this model makes sense?
Are there any problems here?

Regression Model Summary

The computed regression model to predict the point differential is:

$$dif = 0.725(x_1) - 0.693(x_2) - 0.627(x_3) + 0.710(x_4) - 8.332$$

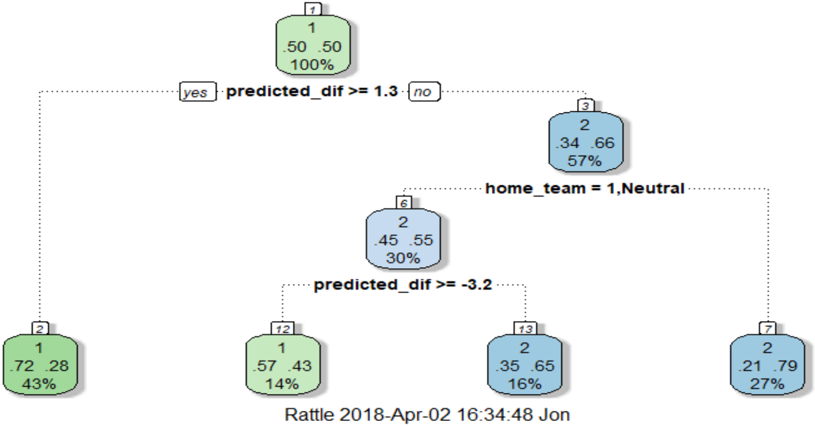
Where,

- dif = Team1 Score - Team2 Score
- x_1 is Team1 average points scored
- x_2 is Team1 average points allowed
- x_3 is Team2 average points scored
- x_4 is Team2 average points allowed

Whenever you get a model, stop and think:
Do we think this model makes sense?
Are there any problems here? Symmetry?

Computed Decision Tree

Decision tree is computed allowing the tree to use: Home Team, Predicted Difference (output of regression), both team's winning percentages, and both team's scoring statistics



Why didn't the tree use the other variables?

Outline

1. Problem Statement
2. Initial Data and Creating Meaningful Data
3. The Models We Will Use
4. **Analysis and Model Performance**

Model Accuracy

- We want to see not only how accurate our model is, but whether or not the "added complexity" of using both regression and a decision tree made for a better model.

Model Accuracy

- We want to see not only how accurate our model is, but whether or not the "added complexity" of using both regression and a decision tree made for a better model.
- Therefore we test the accuracy of "predicted difference" as a model

Model Accuracy

- We want to see not only how accurate our model is, but whether or not the "added complexity" of using both regression and a decision tree made for a better model.
- Therefore we test the accuracy of "predicted difference" as a model

Model: If "predicted difference" is positive, predict team 1 to win.
Otherwise, "predicted difference" is negative, predict team 2 to win

Model Accuracy

- We want to see not only how accurate our model is, but whether or not the "added complexity" of using both regression and a decision tree made for a better model.
- Therefore we test the accuracy of "predicted difference" as a model

Model: If "predicted difference" is positive, predict team 1 to win. Otherwise, "predicted difference" is negative, predict team 2 to win

- Predicted Difference Accuracy: 68%

Decision Tree Accuracy

- Now we test the "full model" accuracy

Decision Tree Accuracy

- Now we test the "full model" accuracy
- That is, use the set of rules defined by the decision tree to predict which team will win the game.

Decision Tree Accuracy

- Now we test the "full model" accuracy
- That is, use the set of rules defined by the decision tree to predict which team will win the game.

- Decision Tree Accuracy: 71%

Decision Tree Accuracy

- Now we test the "full model" accuracy
- That is, use the set of rules defined by the decision tree to predict which team will win the game.

- Decision Tree Accuracy: 71%
- Is a 3% improvement worth it?

Problem Statement

Recall our questions:

1. **How well does a team's seed predict their performance?**
2. **Using "more advanced" data, can we make a better predictor than a team's seed?**
3. Is the tournament any more unpredictable than the regular season?
4. Can statistics help you win your bracket pool?

Seed as Predictor

- We want to compare our "prediction accuracy" to the accuracy generated by picking the higher seed to win

Seed as Predictor

- We want to compare our "prediction accuracy" to the accuracy generated by picking the higher seed to win
- Going back to 1984, how often does the higher seed win?
- When two teams have the same seed we choose randomly between the two

Seed as Predictor

- We want to compare our "prediction accuracy" to the accuracy generated by picking the higher seed to win
- Going back to 1984, how often does the higher seed win?
- When two teams have the same seed we choose randomly between the two
- Guesses?

Seed as Predictor

- We want to compare our "prediction accuracy" to the accuracy generated by picking the higher seed to win
- Going back to 1984, how often does the higher seed win?
- When two teams have the same seed we choose randomly between the two
- Guesses?

- 66% of games were won by the higher seed

Seed as Predictor

- We want to compare our "prediction accuracy" to the accuracy generated by picking the higher seed to win
- Going back to 1984, how often does the higher seed win?
- When two teams have the same seed we choose randomly between the two
- Guesses?

- 66% of games were won by the higher seed
- Our model performs "5%" better...

Seed as Predictor

- We want to compare our "prediction accuracy" to the accuracy generated by picking the higher seed to win
- Going back to 1984, how often does the higher seed win?
- When two teams have the same seed we choose randomly between the two
- Guesses?

- 66% of games were won by the higher seed
- Our model performs "5%" better...
- Our model performs **ONLY** "5%" better???

Summary of Model Accuracies

- Predicted difference accuracy: 68%
- Decision tree accuracy: 71%
- Higher seed accuracy: 66%

Summary of Model Accuracies

- Predicted difference accuracy: 68%
- Decision tree accuracy: 71%
- Higher seed accuracy: 66%
- Other methods I tried which I didn't present:
  - Support Vector Machine Accuracy: 69%

Summary of Model Accuracies

- Predicted difference accuracy: 68%
- Decision tree accuracy: 71%
- Higher seed accuracy: 66%
- Other methods I tried which I didn't present:
  - Support Vector Machine Accuracy: 69%
  - Random forest accuracy: 69% accuracy

Problem Statement

Recall our questions:

1. How well does a team's seed predict their performance?
2. Using "more advanced" data, can we make a better predictor than a team's seed?
3. **Is the tournament any more unpredictable than the regular season?**
4. Can statistics help you win your bracket pool?

The Madness of the Tournament

- I pose the following method to determine whether or not the tournament is any more unpredictable than the regular season

  Method: Compare the percentage accuracy of picking the "higher seed" method to the accuracy of picking the team who has the better record in the regular season

- Recall, higher seed accuracy: 66%

The Madness of the Tournament

- I pose the following method to determine whether or not the tournament is any more unpredictable than the regular season

  Method: Compare the percentage accuracy of picking the "higher seed" method to the accuracy of picking the team who has the better record in the regular season

- Recall, higher seed accuracy: 66%
- Guess how often the better record wins?

The Madness of the Tournament

- I pose the following method to determine whether or not the tournament is any more unpredictable than the regular season

Method: Compare the percentage accuracy of picking the "higher seed" method to the accuracy of picking the team who has the better record in the regular season

- Recall, higher seed accuracy: 66%
- Guess how often the better record wins?
- Better record accuracy: 67%

The Madness of the Tournament

- I pose the following method to determine whether or not the tournament is any more unpredictable than the regular season

Method: Compare the percentage accuracy of picking the "higher seed" method to the accuracy of picking the team who has the better record in the regular season

- Recall, higher seed accuracy: 66%
- Guess how often the better record wins?
- Better record accuracy: 67%
- Conclusion: March Madness is not that mad

Final Conclusions

- It is likely that our 5% improvement is not a result of overfitting:
  - The model was trained on regular season data and tested on tournament data.

Final Conclusions

- It is likely that our 5% improvement is not a result of overfitting:
  - The model was trained on regular season data and tested on tournament data.
- We could have used more advanced data (shooting percentages, rebounding statistics, etc.) to perhaps come up with a better predictor

Final Conclusions

- It is likely that our 5% improvement is not a result of overfitting:
  - The model was trained on regular season data and tested on tournament data.
- We could have used more advanced data (shooting percentages, rebounding statistics, etc.) to perhaps come up with a better predictor
  - It could be interesting to use "expert rankings" as a predictor.

Final Conclusions

- It is likely that our 5% improvement is not a result of overfitting:
  - The model was trained on regular season data and tested on tournament data.
- We could have used more advanced data (shooting percentages, rebounding statistics, etc.) to perhaps come up with a better predictor
  - It could be interesting to use "expert rankings" as a predictor.
- We have one unanswered question still:

Problem Statement

Recall our questions:

1. How well does a team's seed predict their performance?
2. Using "more advanced" data, can we make a better predictor than a team's seed?
3. Is the tournament any more unpredictable than the regular season?
4. **Can statistics help you win your bracket pool?**

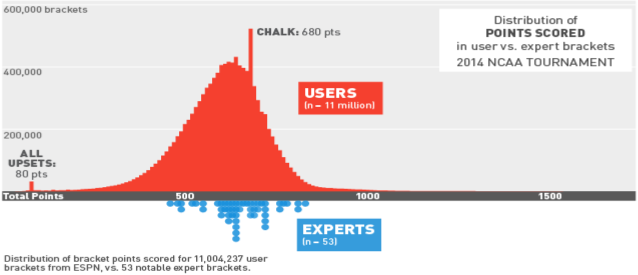


Figure 1: Taken from Stephen Pettigrew's Blog

- The red line labeled "Chalk" is our "higher seed" bracket.

Can Statistics Help you Win your Bracket Pool?

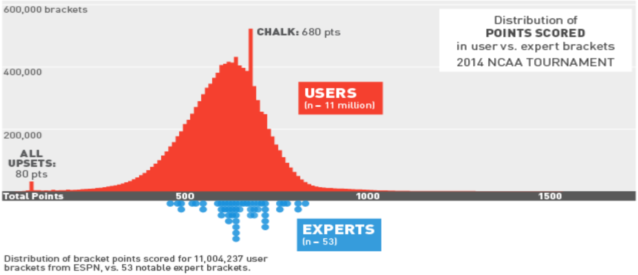


Figure 1: Taken from Stephen Pettigrew's Blog

- The red line labeled "Chalk" is our "higher seed" bracket.
- What do we think? If we add an additional "5%" accuracy to our bracket we will be even further up the curve
- Can Statistics help you win your bracket pool?

Can Statistics Help you Win your Bracket Pool?

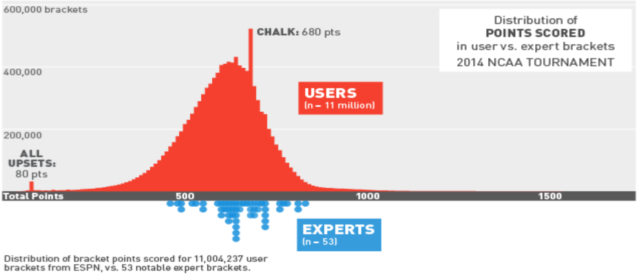


Figure 1: Taken from Stephen Pettigrew's Blog

- The red line labeled "Chalk" is our "higher seed" bracket.
- What do we think? If we add an additional "5%" accuracy to our bracket we will be even further up the curve
- Can Statistics help you win your bracket pool? Maybe