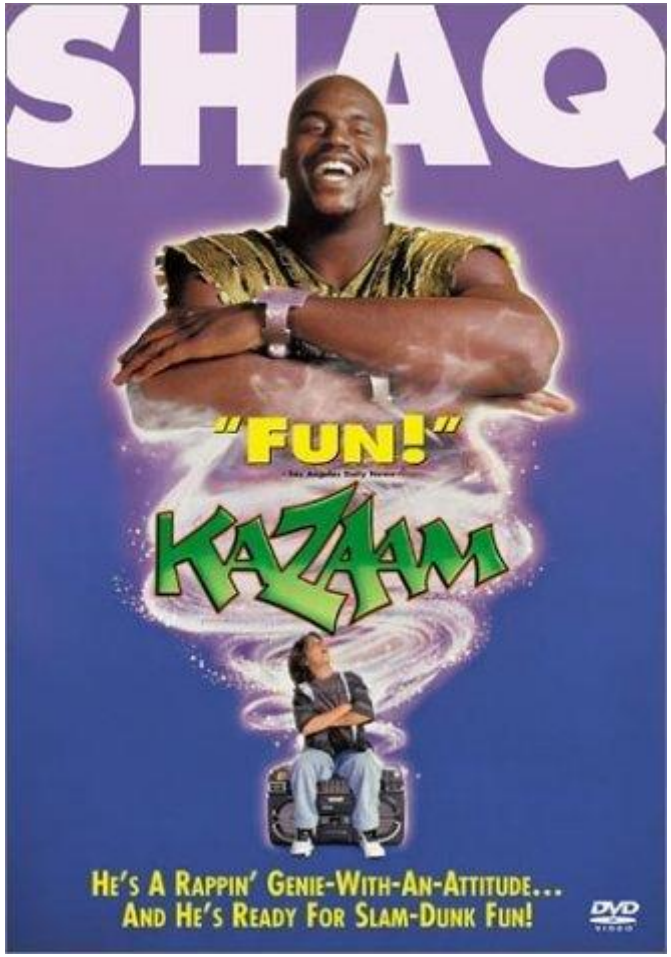
Forecasting the 2015 All NBA Teams

Background


LAKERS


MIAMI
6


SHAQ
"FUN!"
KAZAAM
He's A Rappin' Genie-With-An-Attitude...
And He's Ready For Slam-Dunk Fun!
DVD

Motivation

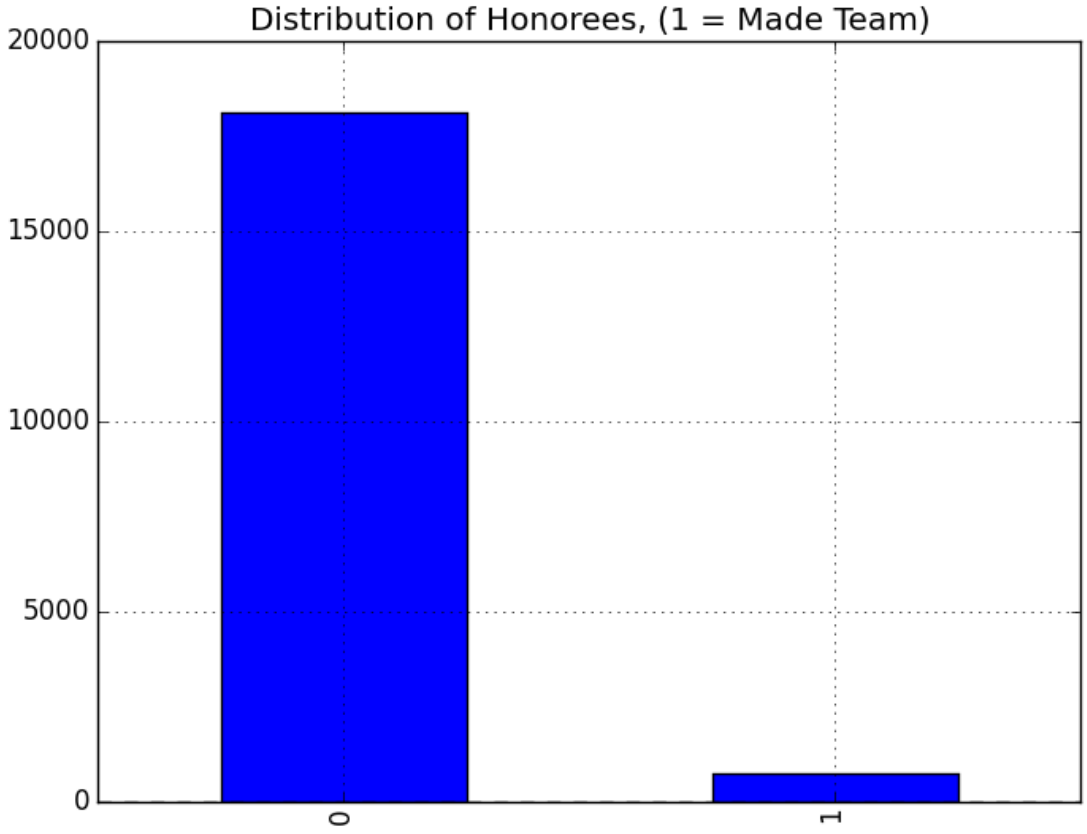
- These honors are supposed to reflect the performance of the best players in their respective positions but is this always true?
  - Do the best players(statistically) have a higher probability of making the team?
- Can I use novel data mining techniques to predict which players will be voted onto the 2015 NBA All-NBA Teams?

The Data

- Every Single NBA player from 1952 – Present
  - 18904 rows(each a player in a given a season)
  - 52 variables
  - Basic Statistics:
    - 25 Continuous Variables
      - I.E.: Field goals, Steals, Blocks, Assists, Turnovers
  - Advance Statistics:
    - 18 Continuous Variables
      - I.E: Player Efficiency Rating, Win Shares , Box Plus/Minus
  - All-NBA Teams(1st, 2nd, and 3rd)

Pre-Processing

1) Collect data
   - Web Scrape via Beautiful Soup
2) Join Data
   - Joined three datasets together by player name and season
     - Basic, Advanced, Historic Team data
3) Cleaned data
   - Filled missing values in with 0 or median value
   - Re-mapped positions
     - PG,SG : G
     - SF, PF : F


Distribution of Honorees, (1 = Made Team)
20000
15000
10000
5000
0
0
1

Exploration/Feature Selection

g	gs	mp	fg	fga	fg_	x3p	x3pa	x3p_	x2p	x2pa	x2p_	ft	fta	ft_	orb	drb	trb	ast	stl
0.183	-0.272	0.0627	0.129	0.148	0.112	-0.421	-0.445	-0.561	0.226	0.284	0.0412	0.169	0.184	0.147	-0.163	-0.282	0.151	0.0885	-0.2
0.0359	0.0739	0.106	0.0344	0.0341	0.0183	0.0892	0.0834	0.0322	0.0159	0.0117	0.0231	0.0216	0.00116	0.0755	0.00703	0.0894	0.0404	0.0887	0.03
1	0.443	0.63	0.543	0.526	0.453	0.133	0.121	0.0611	0.531	0.517	0.438	0.445	0.442	0.458	0.301	0.336	0.446	0.392	0.35
0.443	1	0.611	0.521	0.49	0.261	0.366	0.376	0.197	0.457	0.405	0.278	0.423	0.415	0.204	0.479	0.598	0.363	0.414	0.56
0.63	0.611	1	0.891	0.895	0.27	0.322	0.327	0.0941	0.848	0.846	0.265	0.784	0.781	0.321	0.416	0.552	0.687	0.671	0.56
0.543	0.521	0.891	1	0.982	0.302	0.246	0.249	0.0317	0.977	0.959	0.287	0.856	0.848	0.305	0.376	0.483	0.631	0.594	0.49
0.526	0.49	0.895	0.982	1	0.21	0.296	0.306	0.0559	0.948	0.962	0.202	0.853	0.839	0.317	0.296	0.417	0.599	0.619	0.48
0.453	0.261	0.27	0.302	0.21	1	-0.0866	-0.111	0.00115	0.33	0.253	0.961	0.225	0.239	0.437	0.322	0.278	0.275	0.11	0.18
0.133	0.366	0.322	0.246	0.296	-0.0866	1	0.985	0.524	0.0344	0.027	-0.00707	0.148	0.097	0.152	-0.0929	0.146	-0.0757	0.313	0.37
0.121	0.376	0.327	0.249	0.306	-0.111	0.985	1	0.52	0.0409	0.0335	-0.0102	0.156	0.105	0.149	-0.0877	0.154	-0.0821	0.333	0.40
0.0611	0.197	0.0941	0.0317	0.0559	0.00115	0.524	0.52	1	-0.0825	-0.0913	0.0515	-0.0141	-0.0451	0.159	-0.0813	0.099	-0.0947	0.0838	0.15
0.531	0.457	0.848	0.977	0.948	0.33	0.0344	0.0409	-0.0825	1	0.983	0.297	0.85	0.853	0.281	0.408	0.466	0.667	0.544	0.43
0.517	0.405	0.846	0.959	0.962	0.253	0.027	0.0335	-0.0913	0.983	1	0.215	0.851	0.85	0.29	0.336	0.394	0.652	0.554	0.38
0.438	0.278	0.265	0.287	0.202	0.961	-0.00707	-0.0102	0.0515	0.297	0.215	1	0.205	0.213	0.45	0.29	0.271	0.232	0.128	0.21
0.445	0.423	0.784	0.856	0.853	0.225	0.148	0.156	-0.0141	0.85	0.851	0.205	1	0.982	0.321	0.31	0.39	0.597	0.543	0.37
0.442	0.415	0.781	0.848	0.839	0.239	0.097	0.105	-0.0451	0.853	0.85	0.213	0.982	1	0.251	0.353	0.413	0.659	0.508	0.35
0.458	0.204	0.321	0.305	0.317	0.437	0.152	0.149	0.159	0.281	0.29	0.45	0.321	0.251	1	0.0307	0.0897	0.0944	0.28	0.21
0.301	0.479	0.416	0.376	0.296	0.322	-0.0929	-0.0877	-0.0813	0.408	0.336	0.29	0.31	0.353	0.0307	1	0.859	0.609	0.00555	0.41
0.336	0.598	0.552	0.483	0.417	0.278	0.146	0.154	0.099	0.466	0.394	0.271	0.39	0.413	0.0897	0.859	1	0.636	0.179	0.53
0.446	0.363	0.687	0.631	0.599	0.275	-0.0757	-0.0821	-0.0947	0.667	0.652	0.232	0.597	0.659	0.0944	0.609	0.636	1	0.189	0.18
0.392	0.414	0.671	0.594	0.619	0.11	0.313	0.333	0.0838	0.544	0.554	0.128	0.543	0.508	0.28	0.00555	0.179	0.189	1	0.58

With large set of variables:

Scatterplot Matrix ≠ easy to interpret

Correlation Matrix = easy to interpret

Didn't use features that were inherently correlated like:

FGs Made, Missed and Percentage

Offensive and Defensive Rebounds

Wins Shared(WS), DWS, OWS

Analysis Plan

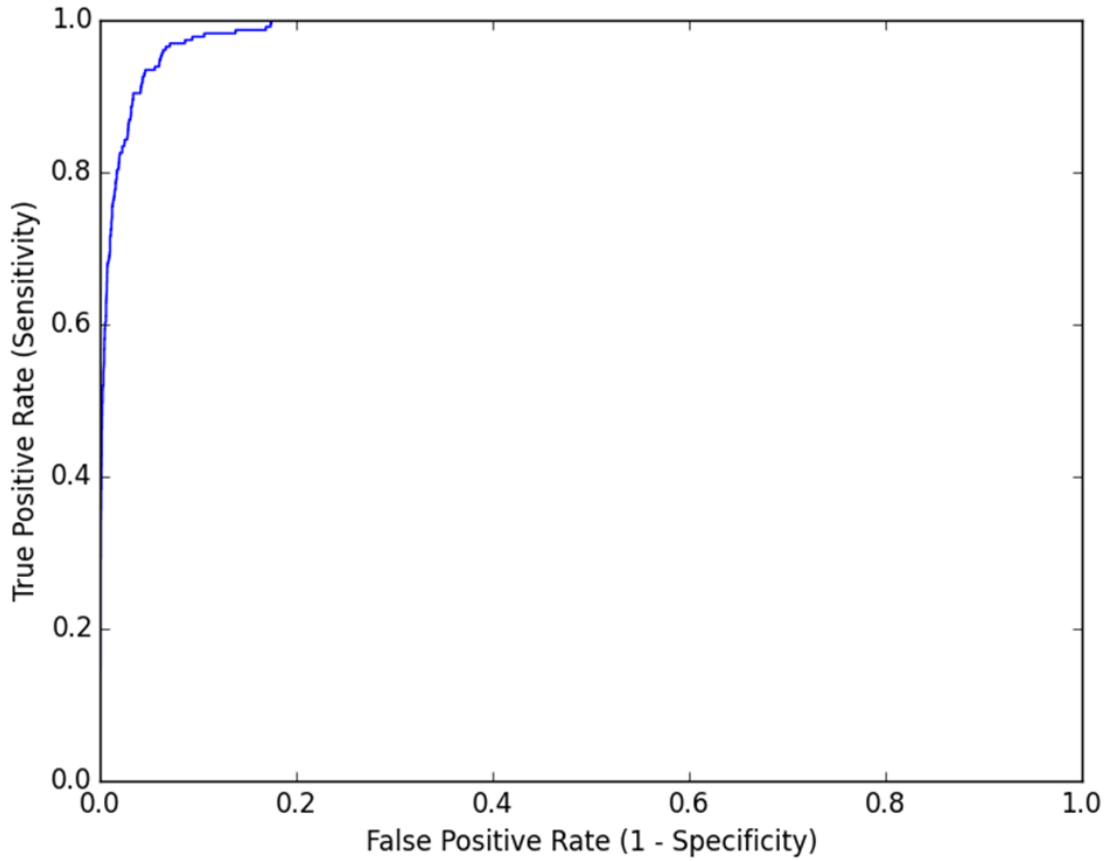
- Use logistic Regression to predict the probability that a player makes a team
- Input Variables:
  - First Model: Basic Stats
  - Second Model: Advanced stats
  - **Third Model: Combination**
- Y variables:
  - Converted Historic Team Variable to binary variable
    - 1 = made a team
- Using the best variables from logistic regression in:
  - Decision Trees and Random Forests

Results

- **Third Model:** Combination of Basic and Advanced Stats
  - **Specificity: .992**
  - **Sensitivity: .655**
  - **Overall Accuracy: .978**

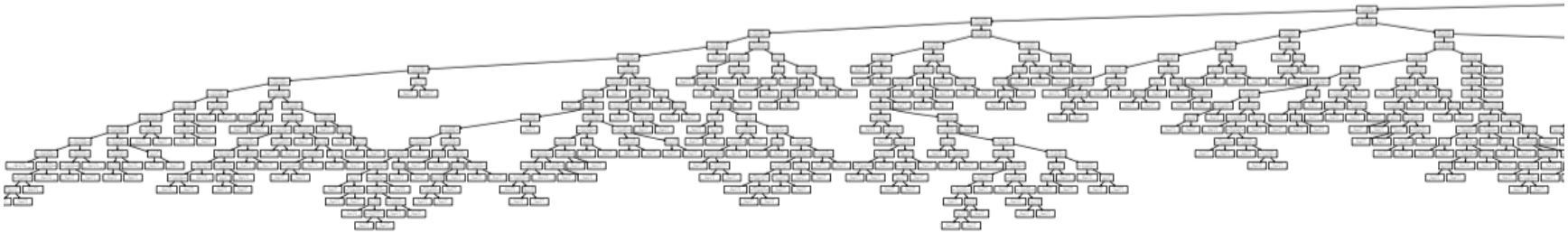
	Predicted: Did not make team	Predicted: Did make team
Actual: Did not make the team	TN =5135	FP = 37
Actual: Did make the team	FN = 79	TP = 150

ROC Curve

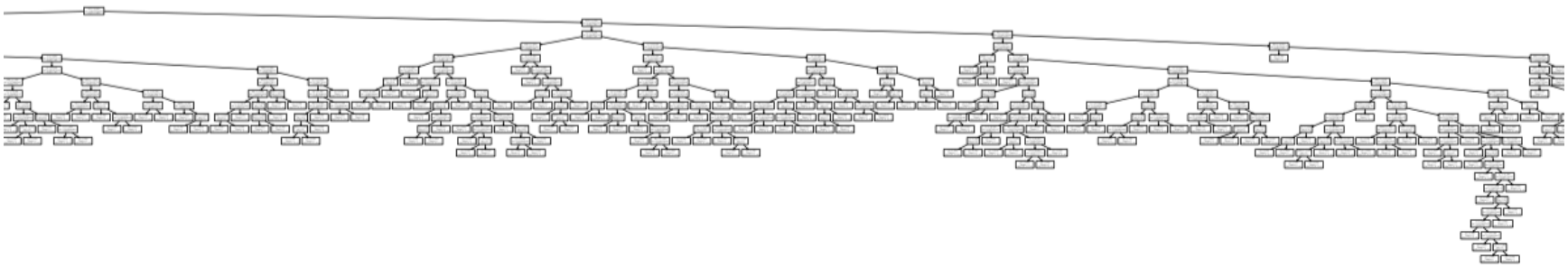

1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate (Sensitivity)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate (1 - Specificity)

Can I use decision trees or random forests to predict what team they make?

First Part……



Second Part……



- Win shares
- Player Efficiency Rating
- Minutes Per game

Important Features

Decision Tree	
feature	importance
g	0.0143158
mp	0.03826871
fg_	0.0499174
ft_	0.0246495
orb	0.00723026
drb	0.00749169
ast	0.02866476
stl	0.00892662
blk	0.01452553
tov	0.02591866
pts	0.12042197
DBPM	0.01333131
OBPM	0.01576599
PER	0.07895327
DWS	0.04513165
WS	0.45274044
WS/48	0.03045382
AST%	0.02329263

Random Forest	
feature	importance
g	0.02971754
mp	0.06046784
fg_	0.05413964
ft_	0.04405769
orb	0.02204337
drb	0.03102316
ast	0.0562888
stl	0.0243098
blk	0.01953337
tov	0.03252539
pts	0.10623309
DBPM	0.02709483
OBPM	0.0330902
PER	0.10744835
DWS	0.08226314
WS	0.16525255
WS/48	0.0678062
AST%	0.03670504

2014 Predictions

Player	Actual	Tree Prediction	Random Forest	Logistic	Probability
Kevin Durant	1st Team	1st Team	1st Team	1	0.999692898
LeBron James	1st Team	1st Team	1st Team	1	0.985469944
James Harden	1st Team	1st Team	1st Team	1	0.947061389
Chris Paul	1st Team	1st Team	1st Team	1	0.880137885
Joakim Noah	1st Team	None	None	0	0.466037131
Kevin Love	2nd Team	1st Team	1st Team	1	0.986704305
Stephen Curry	2nd Team	1st Team	1st Team	1	0.964781235
Blake Griffin	2nd Team	1st Team	1st Team	1	0.805085652
Dwight Howard	2nd Team	None	None	0	0.082847791
Tony Parker	2nd Team	None	None	0	0.010452498
Paul George	3rd Team	2nd Team	2nd Team	1	0.673016981
Damian Lillard	3rd Team	None	None	0	0.293724519
Goran Dragic	3rd Team	None	None	0	0.261742236
LaMarcus Aldridge	3rd Team	None	None	0	0.255747141
Al Jefferson	3rd Team	None	None	0	0.171375588
Noteworthy:					
Carmelo Anthony	None	3rd Team	1st Team	1	0.772540908
DeMarcus Cousins	None	None	2nd Team	0	0.407137468
John Wall	None	2nd Team	None	0	0.299123899
Dirk Nowitzki	None	2nd Team	None	0	0.259351986
Russell Westbrook	None	3rd Team	None	0	0.223400131
Kyle Lowry	None	None	None	1	0.609623766
Anthony Davis	None	None	None	1	0.543741623

Green = All three models correct predicted if player made team and what team they made

Yellow = Predicted who would make it team but didn't not correctly predict what team

Red = Made team but did not correctly predict

2015 predictions!!!

Who is this guy????

Index	player	pred2015	probs2015
187	James Harden	1	1
104	Stephen Curry	1	1
109	Anthony Davis	1	1
455	Russell Westbrook	1	1
224	LeBron James	1	1
347	Chris Paul	1	1
131	Kevin Durant	1	0.999
271	Damian Lillard	1	0.999
94	Jack Cooley	1	0.998
157	Marc Gasol	1	0.997
180	Blake Griffin	1	0.997
158	Pau Gasol	1	0.997
8	LaMarcus Aldridge	1	0.997

Who is Jack Cooley?

Probability of making an All NBA Team is: .998

2015 stats:

- 2 minutes played(total)
- Only has 4 pts
- 100% FGp
- PER = 81.1(the highest in the league right now)

Next Steps

- Investigate problems with predictions
  - Lead to extremely high probabilities of making the team.
  - Might be due to using partial season data
  - Need to predict using less number of variables
    - Having so many variables may be impacting the most important variables in the model distorting the results.
- Would like to see if making the team in the past increases the likelihood of making the team? Is it a popularity contest?

Questions or better yet…Suggestions?