Forecasting the All-NBA Teams
By Kerry Jones
General Assembly DC –DAT4

1.0 Introduction

At the end of every NBA season teams, players and coaches are recognized for their performance through the NBA season. Outside of the 12 annual awards given at the end of year, players' are anointed to All-Pro teams (All NBA, All Rookie and All Defensive Teams) on the basis of their individual performance. It's an honor that is bestowed to the best players in the league at their respective positions: guard, forwards and centers.

Each year three teams composed of two guards, two forwards, and a center voted on by sportswriters and broadcasters across North America. Players receive five points for the first team, three points for the second team, and one point for the third. The five players with the highest total scores make the first team, then next five make the second team, and so forth.

Below is a list of players who made the 2014 teams:

> **First Team:** Lebron James, Kevin Durant, James Harden, Joakim Noah, and Chris Paul
> **Second team:** Blake Griffin, Kevin Love, Dwight Howard, Steph Curry and Tony Parker
> **Third team:** Paul George, LaMarcus Aldridge, Al Jefferson, Goran Dragic, Damian Lillard

To my knowledge, I don't believe there is some criteria or method to the way voters vote for players to receive honors. The award is supposed to reflect the best players in their respective positions but is this always true? Given the availability of historical basketball data, can we predict which players will be voted into each team by the end of the 2014-15 Season? Do the best players' (statistically) have a higher probability of making the All-NBA Teams? I will try to answer these questions using predictive and classification modeling first on the 2014 data to evaluate the models then take the best model to predict the 2015 teams.

2.0 Data Mining

Data mining is the process of automatically discovering information from large data. In general, data mining techniques predict the bad, identify the good, automate an existing process or identify patterns in data. These techniques are useful finding non-trivial patterns that might have otherwise remained unknown.

2.1 Data Mining Tasks

Data mining tasks can be divided into two major categories: Predictive and Descriptive

Predictive Tasks: the goal of this is to predict the value of a particular attribute (dependent variable) based on the values of other attributes (independent variables). This type of

modeling is also known as supervised learning. If the dependent variable being assessed is continuous, the predictive modeling technique is regression. For example, using a set a variables to predict price range or blood pressure. Values under this modeling technique are infinite. On the hand, if the response variable is categorical, classification techniques are used to predict the function of the explanatory variables. For example, predicting a team will win or lose, or determining cancer class of tissue sample.

Descriptive Tasks: The goal under these tasks is to understand patterns that summarize the relationships in the data. There are no dependent variables, just a vector of predictors. The objective is to find groups features that have similar characteristics or find a combination of features that explain the variation in the data. These tasks are useful explanatory tasks that can be used as a preprocessing step for supervised learning procedures. For example: Correlations, trends, clusters, and anomalies.

2.2 Data Mining meets basketball

Defining the problem

Using data mining techniques, can we predict what players will make the 2015 All-NBA team? This question can be assessed using classification techniques. The input variables into these modeling techniques are the basketball statistical attributes. Using basketball attributes as input variables can we predict what individual players make the All-NBA teams, the categorical dependent variable.

3. 0 Data and Methods

3.1 *The dependent variable: team*

Team is a categorical variable that has four values: 1st team, 2nd team, 3rd team or none. I am interested in assessing two things. My first goal is to use the independent variables to predict the probability that the best players in the NBA in a given season make the All- NBA team. I re-code the variable into a binary variable; 1 meaning they made the team, 0 meaning they did not. Second, I would like to use the team variable as a categorical variable to predict what specific teams a player may make based on their season performance.

3.2 *The Independent Variables: Basic Stats and Advance Stats.*

These variables characterize the performance of individual players.

Traditional (Basic) Stats evaluate a player or team's performance

- G, GS: games played, games started
- Pts: Points scores
- Fg, fga. Fg. Field goals made, attempted and percentage.
- FT, FTA, FT.: free throws made, attempted and percentage
- x3p, x3pa, x3p.: three-point field goals made, attempted and percentage
- TRB, ORB, DRB: Total rebounds, Offensive rebounds, Defensive rebounds
- AST: Assists
- STL: Steals

- BLK: Blocks
- TO: Turnovers
- PF: Personal Fouls
- MIN: Minutes

Advance Statistics evaluate a player or teams performance in relation to other players or teams. It's a way to evaluate a players' performance per possession instead of per game like traditional statistics.

- PER: Player Efficiency Rating
- TS: True Shooting percentage(a measure of shooting efficiency)
- 3PAr: Percentage of FGs shot from 3 pt range
- Ftr : Number of Free throws attempts per FG attempt
- ORB%: An estimate of the percentage of available offensive rebounds a player grabbed while he was on the floor
- DRB%: An estimate of the percentage of available defensive rebounds a player grabbed while he was on the floor
- TRB% : An estimate of the percentage of available total rebounds a player grabbed while he was on the floor
- AST% : An estimate of the percentage of teammate field goals a player assisted while he was on the floor
- STL%: An estimate of the percentage of opponent possessions that end with a steal by the player when he is on the floor
- Tov%: An estimated number of turnovers committed per 100 plays
- Blk%: An estimate of the percentage of opponent two point field goals attempts blocked by player when he was on the floor.
- WS: An estimated number of wins a player contributed to.
- OWS: An estimated number of wins a player contributed to due his offense.
- DWS: An estimated number of wins a player contributed to due his defense.
- Ws_per_48 : An estimated number of wins a player contributed to per 48 mins.
- OBPM: Offensive Box Plus /Minus
- DBPM: Defensive Box Plus /Minus
- BPM: Box Plus/ Minus

All data from this model was collected using a scrape function from Basketball Reference.com, a repository for present and historical basketball data. The temporal coverage of the dataset ranges from 1952 – 2014.

4.0 Preprocessing

The original data I collected was from github. It only contained the traditional NBA statistics variables mentioned above. I developed two functions in python using the beautiful soup library to scrape more data from Basketball-Reference.com. The first function was developed to gather the advance metric data and the other pulled together the names of every player that made an All-NBA team. After I collected the data, both datasets were put into data frames and joined together in pandas with the traditional basketball data.

Next, I assessed the data frame for missing values. All the shooting percentage variables were missing. This occurred when players in a given season actually didn't take a single shot attempt (free throw, two pointer, or a three pointer). These values were converted to 0. Certain attributes such as three point field goals, steals, turnovers, offensive and defensive rebounds and blocks were not recorded throughout the course of NBA history. Although, I considered taking the average of the variables by position, I decided that for the initial model I would also fill the missing variables with 0. The same is true for there corresponding advance metric data (i.e. ORB%, STL**%).** For the other variables with missing values, such as Win Shares and BPM, I took the median of this as well because they are more or less assessments of average players contribution towards offensive, defense, and total.

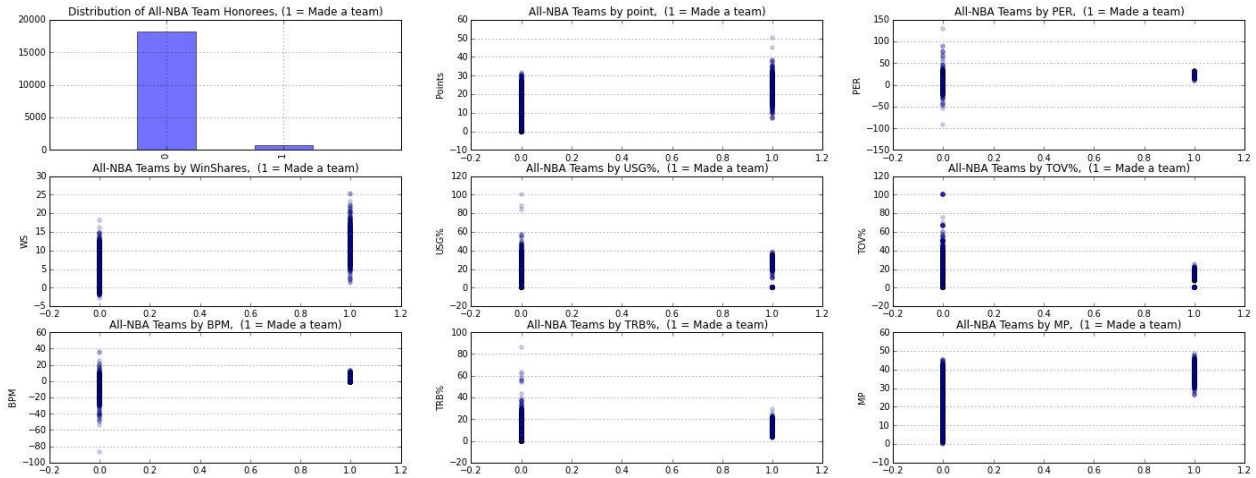
Teams in the NBA are typically broken up into five positions; Point Guards, Shooting Guards, Small Forwards, Power Forwards, and Centers. These honorary teams are classified under more generic positions (Guards, Forwards, and Centers). So for example, two point guards can make the first team or two shooting guards. In the data, the subtypes of forwards and guards were mapped to their generic classes (Small Forward: Forward).
** I did not consider using the position variable in the model but will in later steps*

Lastly, for players that didn't make the team, I converted their missing values in the team variable to "None".

5.0 Descriptive and Exploratory Analysis

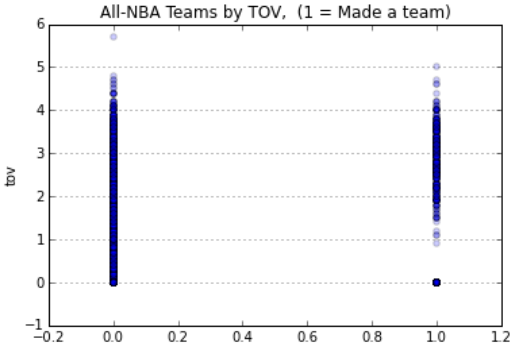
Initially, I thought it was important to get an idea of what was in the data; for instance the average points a player scores per game by position. My initial findings weren't very surprising. On average, each player by position scores 8 – 9 points per game, has three to one assist, three to 5 five rebounds, and with an average PER of about 12.9. As you would assume, the distribution of most attributes is skewed to left; typically below the average. This makes since because most player will not average over 20 points a game or over 12 rebounds. This is most likely because they are not getting a lot of playing time and if they are, they probably are not the best players on the court.

Since I didn't get anything particularly interesting from this initial analysis, I decided it might good to look at a few variables I thought could distinguish players who didn't make the All-NBA team historically to players who have. Below is a bar plot alongside several scatterplot comparing these two groups of players.

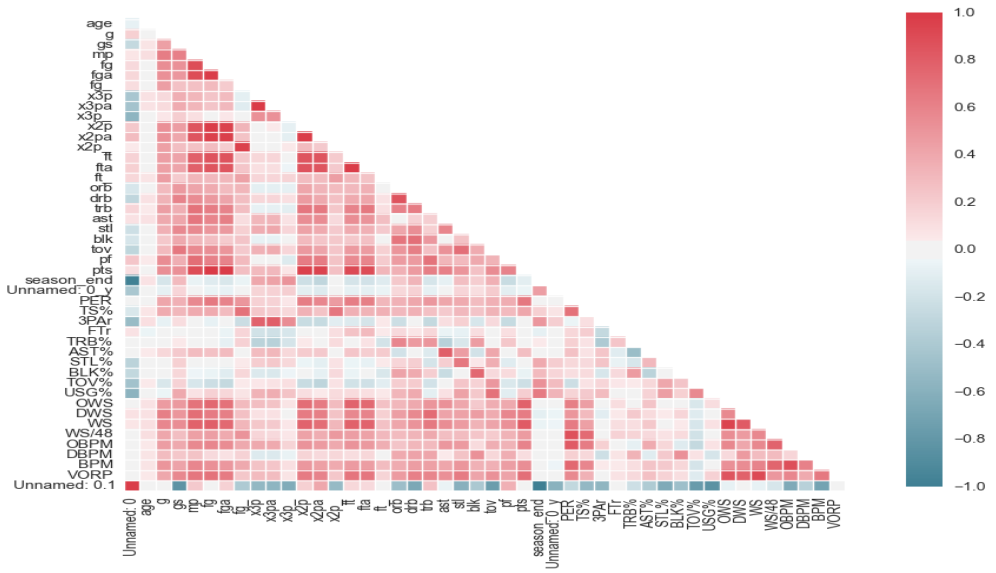


Only 750 players in the history of the award have received this award. Nearly half of the players who make the team average over 20 points a game and rarely is their a case of someone who averaged under 10 points a game and still made the team. The variance of PER for players who made the team is extremely small in relation to players that have not. Extremely productive players in the NBA won't have an abnormally high PER like 129 but will have a rating that is positive and above average.

As expected Win Shares and minutes played were higher for players who made the team. However, BPM, USG% and TOV% were lower than I anticipated. I assumed USG% would be higher for players that made the team because I suspected that more players would be ran for the best players on the floor. TOV% is lower than I expected, initially I thought the estimated number of turnovers committed per 100 plays for players who made the team would have showed slightly more variance because even the best guards turn the ball over at an above average rate (its assumed because they are the main ball handler) Steph average 3.5 turnovers last year; even Allen Iverson and Magic averaged 4.6 in 2004 -2005 1983-1984 respectively. That being said average turnovers per game is as I assumed



Below is a heat map that visualizes the relationship between the different variables in the dataset.



Given that there over 45 variables in this data set, determining the relationship between the variables can be incredibly challenging. The heat map helps me visually assess the correlated relationship between the variables. The darker values (red or blue), the higher the correlation amongst the two or more variables. So inherently variables such as Field Goals Made and Field Goals Attempted, Offensive and Defensive rebounds, and Offensive and Defensive Win Shares. I avoided using these highly correlated variables in he models to avoid multicolinarity.

6.0 Prediction and machine learning

The goal of prediction modeling is driven by the data. The algorithms in this discipline use feature data to make the best guess or estimate about the value of the outcome variables. The machine essentially "learns the data" and then makes a guess about possible outcomes when given new data or out of sample data. In order to make a model that generalizes well it should follow the following steps:

1) **Split Dataset**
2) **Train model**
3) **Test Model**
4) **Parameter Tuning**
5) **Choose best model**
6) **Train on all data**
7) **Make predications on new data**

A typical data split is 70% training and 30% in validation. The validation is set aside until the very end until all the parts of the model are finalized. Once that it done, the validation set is used to estimate the error rate of the prediction. Training error occurs when the model under or overfits the data. In order to evaluate the model, specifically in binary responses, we assess the accuracy, sensitivity, and specificity. How accurate does model predict outcomes, sensitivity, and specificity. Different splits give us different test errors so in order to fully assess the model we use cross- validation; which is essentially splitting,

training, and validating you data multiple times and then takes the average test error rate. Averaging multiple accuracies can reduce variability without giving up bias. Below I will walk you through this framework I used to predict the probability a given NBA player makes the All-NBA team.

6.1 Classification models: *Logistic Regression, Decision Trees and Random Forests*

6.1.1 Logistic Regression

Logistic Regression is a supervised classification model typically used for categorical data. It's essentially a generalization of the linear regression model for classification problems. In order to assess my question, I use logistic regression to predict the probability of a player making the All-NBA team. Using this method, I will try three different approaches. In the first model, I will use just the basic statistics to predict dependent variable. I will then only the advance variables in the second, and a combination of the two in the third.

Results

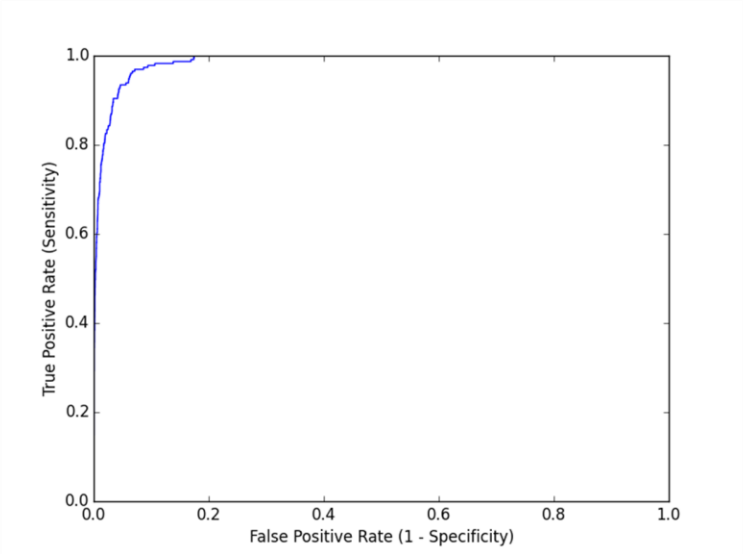
Model	Variables	Specificity	Sensitivity	Overall Accuracy
First Model	Basic	.991	.572	.973
Second Model	Advanced	.994	.54	.975
Third Model	Combination	.992	.655	.978

6.1.2 Model Evaluation

The combination of the basic and advanced statistics resulted in the best model. This model did really well at predicting who didn't make the team at a specificity level of .992. Surprisingly, the model is significantly better at predicting who made the team in comparison to the first two models at a sensitivity level of .655. Given my question, the true positive rate (how often the model correctly predicted who made the team) is most important metric in my analysis. Using the cross-validation technique, my average AUC is .986. A near perfect model is one, so the closer my score is to one the better the model is. Below is the Confusion table and ROC Curve:

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

the team		



6.1.3 Decision Tree

Decision Trees are tree-based method for both regression and classification. It is a tree-like graph with a model of decisions and possible outcomes. The predictors are segmented into regions based on some split value. Predictions are made by using the mean of the training observation it belongs to. Essentially, it is like asking the model a question, then each time it receives an answer, you ask it a follow up question until no further information can be extracted and a conclusion can be made about the class label. So in terms of this project, I used decision trees to predict what All-NBA teams (labels) players' make.

Using this model, I found the most important features are: Win Shares, PER, Minutes, Points per game, Assist per game, and Field Goal and Free throw percentage. In other words, these are the features the model use to make a series of splitting decisions to categorize the players. For instance, the first split, win shares. So players with a Win Share value greater or equal to 10.5, were assigned to right side of branch, while the other players were assigned to the left. This segmentation continues to occur until no more information can be extracted from the given variables.

Given that Win Shares is the estimated number of wins an individual player contributed. The performance of a player is typically going to effect how many games the team wins. The best player on their team is going to be doing all the tangible things (rebounds, steals, scoring) necessary to help their team win and doing them efficiently. So yes, scoring a lot of points is good but if a player shoots 8 for 30, that's not effective. Those missed shots can potentially lead to lost opportunities that would have evidently helped his team main or get

lead. Furthermore, the best players in the league are typically going to get most minutes in a game. The more minutes the player is left on the court to produce through the game, creates more efficient production via scoring and/or assists that and could potentially increase number of wins.

6.1.4 Random Forests

Random Forests is a type of ensemble method use to improve classification accuracy. It is specifically designed for tree classifiers. It takes the combination of multiple decision trees that were independently generated based on an independent set of random vectors. Each time a split is considered during the building tree process, a random sample of predictors is chosen as split candidates (split is only allowed to use one predictor). Random forest reduces variance in a mode because at each split a random predictor is left out, decorrelating the team.

The results from this model did not improve my model by all but did the results from my decision tree. I actually found all the same important features from my decision tree model was also found in random forests too, except put less emphasis on Win Shares.

Random Forest		Decision Tree	
feature	importance	feature	importance
g	0.02971754	g	0.0143158
mp	0.06046784	mp	0.03826871
fg_	0.05413964	fg_	0.0499174
ft_	0.04405769	ft_	0.0246495
orb	0.02204337	orb	0.00723026
drb	0.03102316	drb	0.00749169
ast	0.0562888	ast	0.02866476
stl	0.0243098	stl	0.00892662
blk	0.01953337	blk	0.01452553
tov	0.03252539	tov	0.02591866
pts	0.10623309	pts	0.12042197
DBPM	0.02709483	DBPM	0.01333131
OBPM	0.0330902	OBPM	0.01576599
PER	0.10744835	PER	0.07895327
DWS	0.08226314	DWS	0.04513165
WS	0.16525255	WS	0.45274044
WS/48	0.0678062	WS/48	0.03045382
AST%	0.03670504	AST%	0.02329263

7.0 Results: *So did the best players in their respective positions get voted to the All-NBA Teams in 2014?*

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

The logistics regression accurately predicted 8/ 15 of players that made the team in 2014 at a probability threshold of .5. The Decision Trees and Random Forests gave basically the same results, so not sure if using the Random Forest model actually improves the decision tree predictions. All three models that the best player in their given positions do not always make the All-NBA teams. The three models did extremely well at predicting the first four players. I am not surprised that Joakim Noah came up short. He is not particular known for he's offensive but he does a lot of things on the defensive side of the floor that cannot be easily captured in the variables I chose to us.

The predicted results for Tony Parker and Dwight Howard are particular interesting because the models suggests that both of them had an extremely high unlikelihood of making any team yet they made they both made 2nd team. These players clearly were not the best players in their respective positions last year but is there another factor that should be considered such as the team record. Houston Rockets and San Antonio Spurs were had over 50 wins last season and were ranked fourth and first respectively amongst teams in their conference.

Further evidence of this of the need to explore the importance of team records can be seen with a player like Carmelo Anthony. The model suggest that had a 77 percent chance of making a team and could have probably made any of the All-NBA teams given his stats last

year. The model actually suggests he should have made it over both Paul George and LaMarcus Aldridge. However, his team, the Knicks only won 37 out of 82 games last year and did not make the playoffs in a horrific Eastern Conference.

Several other players that did not make an All-NBA Team but higher probability of making the team over actual players selected: Anthony Davis, Demarcus Cousins, John Wall, and Kyle Lowry. Does this suggest that there is another factor I'm not considering? Is making the All-NBA Team a popularity contest? Does making a team in a previous effect whether you make it in a present year? Does the career length in the NBA also a determining factor? I will need to investigate these factors at a later date to see if they are determining factors of whether someone makes the All-NBA team.

7.1 2015 Predictions

I used the logistic model I used to predict the 2014 All-NBA team on the current 2015 data. The results were not as impressive as what I had received from the 2014 data. From the results below, its obvious that something is either wrong with my 2015 season data or the model. Yes, all of the players who I thought had a probability of making the team are at the top of the table but their probabilities are unreasonably high. Another error I found in my predictions was index 94, player name, Jack Cooley.

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

Jack Cooley has a .998 probability of making the All-NBA team. As exciting as the news is for Jack, he has only played two minutes this year, has a total of four points on 100% Field Goal percentage and the highest PER in the league at 81.1. Too much weight is being put on his PER. However, I did find a similar player in the 2014 data. DeAndre Liggins had highest PER in the league last year at 129.1, however his probability of making team was only .143. So something is up.

8.0 Conclusion/ Next Steps

My initial analysis is promising, but I will to lessen the number of predictor variables I am using to predict NBA players. I will continue to investigate the problems in my 2015 predictions. It may be due to data partiality or might also be due to the number of variables. Having so many variables may be impacting the most important variables in the model distorting the results. Next I want to address some of the factors I mentioned above: Is making the All-NBA Team a popularity contest? Does making a team in a previous effect whether you make it in a present year? Does the career length in the NBA also a determining factor?