Data Analytics Project

Yufei Wang

Introduction

The All-NBA Team is the annual National Basketball Association honor on the best players in the league for every NBA season. In general, the voting panel for All-NBA players consists of different media members and members from each of the league's 30 teams. Player statistics, such as points, assists, and rebounds, do matter for the selection. The five players with the highest point totals make the first team, and the next five makes the second and so forth. Therefore, fifteen players are selected to the three All NBA teams in each season.

In this project, I wanted to create a generalized model to predict the number of All-NBA selections for any given player at any point in their NBA career. It is a classification model because players can be classified as a 1 (All-NBA) or 0 (not All-NBA). Several key questions about this project are following: 1) where can I find reliable data sources of player statistics and All NBA selections? 2) which indicators (player statistics) are the most related to the All-NBA selection? 3) which classification algorithms work for the prediction? 4) How could I evaluate the classification model accuracy? 5) how could I use this statistical prediction to determine how a player might perform over the next few years, based on their trending performance? To answer these questions, let`s define the problem firstly.

Define the problem

Use the classification algorithm to classify All-NBA players and Non-All-NBA players based on reliable player statistics datasets. Evaluate the model accuracy and predict the number of All-NBA selections for the given player at defined NBA seasons.

Data preparation

In this project, I used two csv. files. One is the dataset of all NBA players and their statistics from 1950 – 2017, and another one is the data set of all NBA players selected to the All -NBA teams from 1988 – 2018 (sourced from basketball-reference.com). To create the nice tidying dataset, I reduced the number of years for the analysis to 1998-2007 because of the high data quality. Since the 'Year' variable in the NBA players data corresponds to the 'Season' variable in the All NBA data, I created an equivalent Year column in the All NBA dataset, and reduced the observations in both datasets to 1998-2017. Next, I checked the missing values of both datasets. There are 24225 missing data in the NBA players and 6 missing data in the All NBA dataset. I checked them by columns of NBA players dataset and listed the output below.

Figure 1

```
[1] 6
[1] 24225
     X   Year Player    Pos    Age     Tm      G     GS     MP    PER    TS.  X3PAr    FTr   ORB.
     0      0      0      0      0      0      0      0      0      5     53     57     57      5
  DRB.   TRB.   AST.   STL.   BLK.   TOV.   USG.    OWS    DWS     WS  WS.48   OBPM   DBPM    BPM
     5      5      5      5      5     43      5      0      0      0      5      0      0      0
  VORP     FG    FGA    FG.    X3P   X3PA   X3P.    X2P   X2PA   X2P.   eFG.     FT    FTA    FT.
     0      0      0     57      0      0   1876      0      0     80     57      0      0    478
   ORB    DRB    TRB    AST    STL    BLK    TOV     PF    PTS
     0      0      0      0      0      0      0      0      0
```

The key metrics I cared here are PER and USG. Player efficiency rating(PER) sums up all a player's positive accomplishments, subtracts the negative accomplishments, and returns a per-minute rating of a player's performance. The usage metric is an indicator of how involved a player is in his team's plays. As you can see, there are five NA values in both columns. To check the consistency of same five NA in both columns, I used the identical function in R, and got the

YES output. I removed the five players who did not play enough time (the PER is NA) and checked columns again. Another issue is the Double-counting from all players statistics. The row containing the total represents total seasons in which a player was trader, which is not relevant to this project, therefore, I removed the columns by doing the subset function in R.

Features and feature selection

Data cleaning is tedious, but it is important. The pregame statistics from All-NBA dataset make more sense than the original player statistics. Therefore, I created a new data frame that is based on per-game statistics. The new dataset contained 22 columns, rather than 51columns. Detailed descriptions of these columns are following:

1. Name
2. Position: (PG/SG/SF/PF/C)
3. Age
4. Year
5. Team
6. Games
7. Starts: How many of the games the player played in did they start?
8. Minutes (note that from minutes through to FGs, all values are converted to per-game)
9. Points
10. Rebounds
11. Assists
12. Steals
13. Blocks
14. Turnovers
15. Fouls
16. FTs: Number of made free-throws.
17. Threes: Number of made three-point shots.
18. FGs: Number of made field goals.
19. Usage: Statistic represented how involved the player was in his team's plays.
20. PER: Advanced statistic developed by Hollinger to calculate a player's overall efficiency/ output
21. Box Plus/Minus: statistical metric of overall player performance.
22. Shooting Percentage: Average shot percentage, weighted based on different values of different shots .

The references for these selected features came from All-NBA dataset and another analysis reports. To transform the data to per-game stats, I used the mutate function to divide each player's season totals by games played for the columns I mentioned above. I used the sapply function to loop apply that function to the NBA players dataset, and applied str function and summary function to the new dataset. The outputs from R are shown here:

Figure 2

```
'data.frame':   8355 obs. of  22 variables:
 $ PF              : int  121 137 77 117 73 78 72 89 98 75 ...
 $ PTS             : int  454 1152 261 856 409 412 701 427 134 319 ...
 $ Name            : chr  "Tariq Abdul-Wahad" "Shareef Abdur-Rahim" "Cory Alexander" "Ray Allen" ...
 $ Position        : chr  "SG" "SF" "PG" "SG" ...
 $ age             : int  24 22 25 23 24 28 31 25 24 31 ...
 $ year            : int  1999 1999 1999 1999 1999 1999 1999 1999 1999 1999 ...
 $ Team            : chr  "SAC" "VAN" "DEN" "MIL" ...
 $ Games           : num  49 50 36 50 38 34 47 50 41 50 ...
 $ Starts          : num  49 50 4 50 13 33 39 2 4 0 ...
 $ Minutes         : num  24.6 40.4 21.6 34.4 25.7 ...
 $ Points          : num  9.27 23.04 7.25 17.12 10.76 ...
 $ Rebounds        : num  3.8 7.48 2.06 4.24 2.87 3.03 5.89 2.64 2.37 1.26 ...
 $ Assists         : num  1.02 3.44 3.31 3.56 3.82 5.68 1.94 1.12 0.66 2 ...
 $ Steals          : num  1.02 1.38 0.97 1.06 1.26 0.97 1.36 0.78 0.44 1.32 ...
 $ Blocks          : num  0.33 1.1 0.14 0.14 0.11 0.06 0.32 0.2 0.32 0.06 ...
 $ Turnovers       : num  1.43 3.72 1.92 2.44 2.16 2.09 1.77 1.32 0.63 1.1 ...
 $ Fouls           : num  2.47 2.74 2.14 2.34 1.92 2.29 1.53 1.78 2.39 1.5 ...
 $ FTs             : num  1.92 7.38 1.03 3.52 3.63 2.47 2.11 1.78 0.83 1.24 ...
 $ Threes          : num  0.122 0.22 0.833 1.48 0.553 ...
 $ FGs             : num  3.61 7.72 2.69 6.06 3.29 ...
 $ Usage           : num  19 28.9 20.3 24.5 23.4 20.9 22.8 21.6 18.2 19.6 ...
 $ EfficiencyRating: num  11.8 20.7 11 18.9 16.5 16.7 15.3 12.8 9.2 16.6 ...
```

Figure 3

```
     Name             Position               age             year           Team               Games      
 Length:9722        Length:9722        Min.   :18.00   Min.   :1999   Length:9722        Min.   : 1.00  
 Class :character   Class :character   1st Qu.:23.00   1st Qu.:2004   Class :character   1st Qu.:24.00  
 Mode  :character   Mode  :character   Median :26.00   Median :2008   Mode  :character   Median :50.00  
                                       Mean   :26.87   Mean   :2008                      Mean   :47.28  
                                       3rd Qu.:30.00   3rd Qu.:2013                      3rd Qu.:73.00  
                                       Max.   :44.00   Max.   :2017                      Max.   :82.00  

     Starts         Minutes          Points          Rebounds         Assists           Steals      
 Min.   : 0.00   Min.   : 0.67   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   :0.0000  
 1st Qu.: 0.00   1st Qu.:11.50   1st Qu.: 3.330   1st Qu.: 1.690   1st Qu.: 0.510   1st Qu.:0.3000  
 Median : 8.00   Median :19.34   Median : 6.380   Median : 2.880   Median : 1.150   Median :0.5500  
 Mean   :23.06   Mean   :20.04   Mean   : 7.855   Mean   : 3.473   Mean   : 1.749   Mean   :0.6286  
 3rd Qu.:43.00   3rd Qu.:28.42   3rd Qu.:11.160   3rd Qu.: 4.620   3rd Qu.: 2.360   3rd Qu.:0.8800  
 Max.   :82.00   Max.   :43.70   Max.   :35.400   Max.   :18.000   Max.   :12.750   Max.   :3.0000  

     Blocks         Turnovers         Fouls            FTs             Threes            FGs        
 Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   : 0.000   Min.   :0.0000   Min.   : 0.000  
 1st Qu.:0.1000   1st Qu.:0.580   1st Qu.:1.230   1st Qu.: 0.500   1st Qu.:0.0000   1st Qu.: 1.260  
 Median :0.2400   Median :1.000   Median :1.850   Median : 1.020   Median :0.2300   Median : 2.430  
 Mean   :0.4006   Mean   :1.165   Mean   :1.832   Mean   : 1.459   Mean   :0.5109   Mean   : 2.942  
 3rd Qu.:0.5075   3rd Qu.:1.600   3rd Qu.:2.430   3rd Qu.: 1.970   3rd Qu.:0.8800   3rd Qu.: 4.170  
 Max.   :6.0000   Max.   :5.730   Max.   :6.000   Max.   :10.270   Max.   :5.0900   Max.   :12.220  

     Usage       EfficiencyRating  BoxPlusMinus      ShootingPercentage
 Min.   : 0.0   Min.   :-90.60   Min.   :-86.700   Min.   :0.0000    
 1st Qu.:15.1   1st Qu.:  9.60   1st Qu.: -4.100   1st Qu.:0.4330    
 Median :18.4   Median : 12.70   Median : -1.700   Median :0.4770    
 Mean   :18.7   Mean   : 12.45   Mean   : -2.247   Mean   :0.4658    
 3rd Qu.:22.0   3rd Qu.: 15.80   3rd Qu.:  0.400   3rd Qu.:0.5140    
 Max.   :88.3   Max.   :129.10   Max.   : 26.600   Max.   :1.5000    
                                                   NA's   :52        
```

As you can see, the summary of this new dataset introduced the extreme variability into some metrics, for example, the PER ranges from -90 to 129 for all players, which indicates that there are outliers in the new dataset. Those outliers are not helpful in predicting the All-NBA selection, and reduce the predictive power like PER. Therefore, I have to set up the minimum threhold for game and minute to minimize the outliers effect. I used the filter function by setting that the smallest game played is 10 and the mininum playing time is 5 minutes for inclusion in our analysis dataset.

After transforming the players data to a set of per-game statistics, I had to identify players who were selected for All-NBA teams in per-game statistics. Since the lengths of two tables are different and each players had multiple entries across seasons, the simple join of two tables did not work, and the name is not an unique identifier for each player. I had to create the unique udentifier for both data sets as the linkage key. I chose the combination of player name, year, team, and age because they are the common columns in both data sets. I used the substr function in R to combine these strings, and the unique identifier is the combination of the first three letters of player`s name, the player`s age, the first three letters of player`s team, the end two numbers of the season year. Once the unique identifier is created, I added a new column for the All-NBA indicator, and use an ifelse function to find if they are the same combinations of identifer in both datasets, and recorded 1 in this new column. I used the sum function to check the All-NBA indicator, and got 285 output which is the same result as the dim function on the All-NBA dataset. Therefore, I knew that all All-NBA players in the per-game statistics were indicated by 1 and the rest were indicated by 0.

Exploratory analysis and data visualization

All-NBA selection is the statistics driven game, which means all players statistics can be relevant to the final selection. In this exploratory analysis, I ignored the defensive stats here (steals and blocks). But I will include them in multivariate logistic regression model later.

Firstly, I used ggplot to plot the density graphs of player outputs. In these graphs, we can see that overall output (points, rebounds, assists and turnovers) is right skewed with positive skewness, and the overall output(minutes, threes, usage and shooting percentage) is not well skewed. The vertical dash line represents the mean of each metric, and I can assume that the best players are influential observations in our dataset from the overall outputs. Next, I was interested in the output of player`s age and efficiency rating. I used the groupby function and summarize function to get the age group with mean efficiency rating. These yellow dots represent the size of each age group. As you can see, this inverted u-shaped curve indicated that basketball players enter the league until they reach their prime at 27-30, and then steady decline afterward.

Figure 4

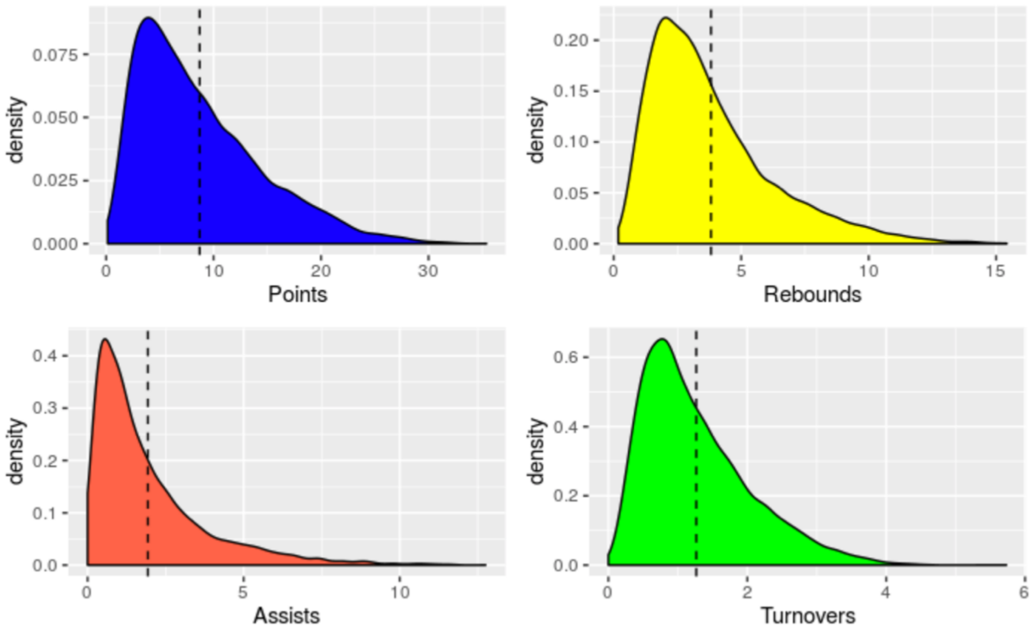


Figure 5

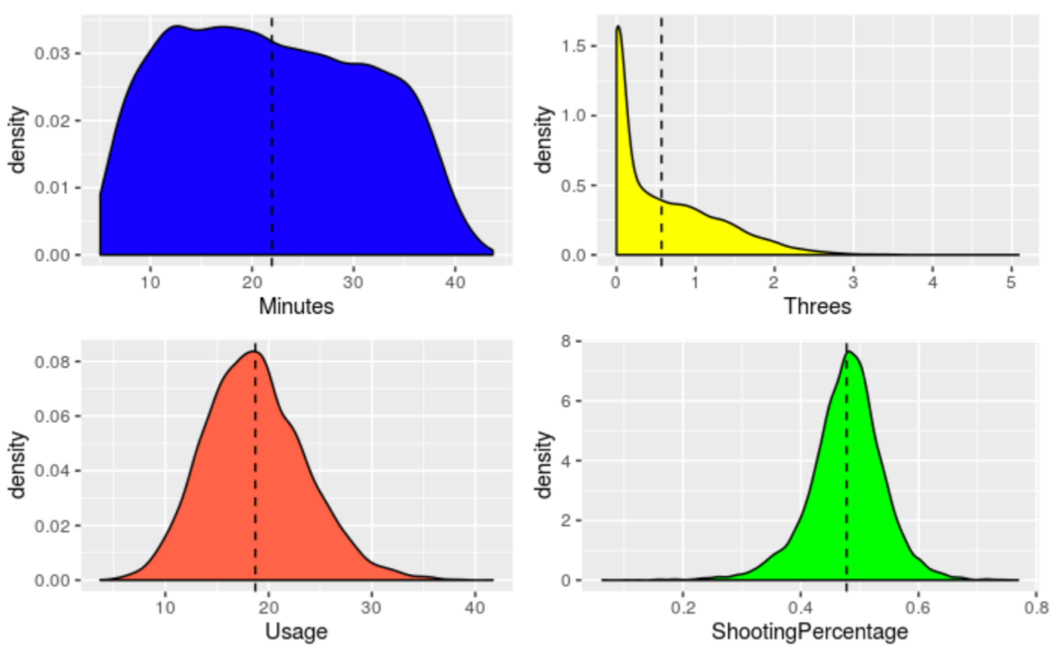
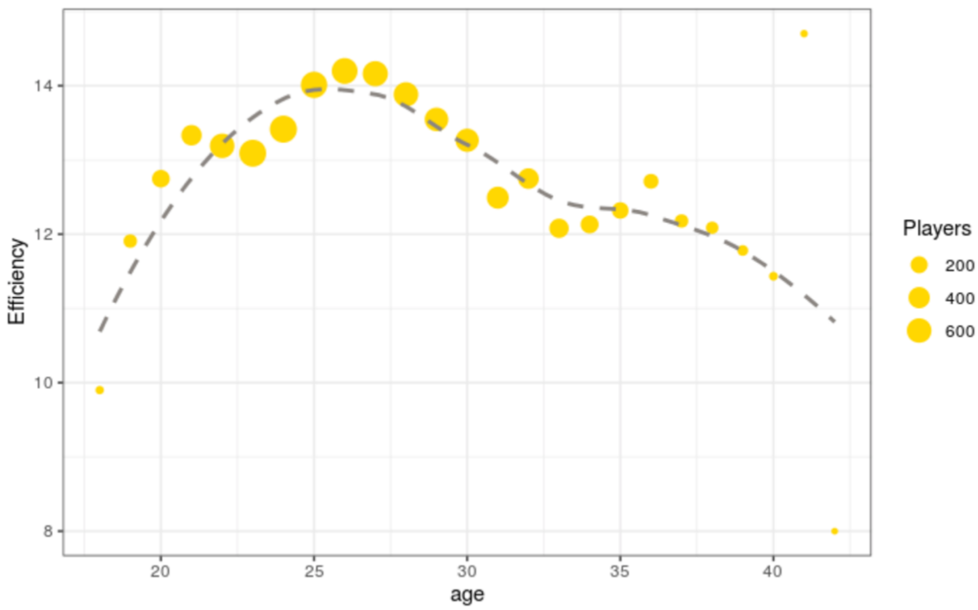


Figure 6



As I mentioned before, the efficiency rating (PER) is the most advanced metric to evaluate player`s performance. Therefore, it is important to find out how other variables are related to PER. I assumed that the correlations of these variables with PER will be high, and used the linear regression model (build in ggplot "lm" option in geom_smooth method) to show the straight-line fittings. As you can these plots below, all eight variables are positively related to PER, and these relationships seem linear. However, linear relationship between shooting percentage and PER seems out of the range when the percentage is around 50%. The cluster points of threes in 0 to 1 mostly, indicating that the linear model is not the best approximation of this relationship.

Figure 7

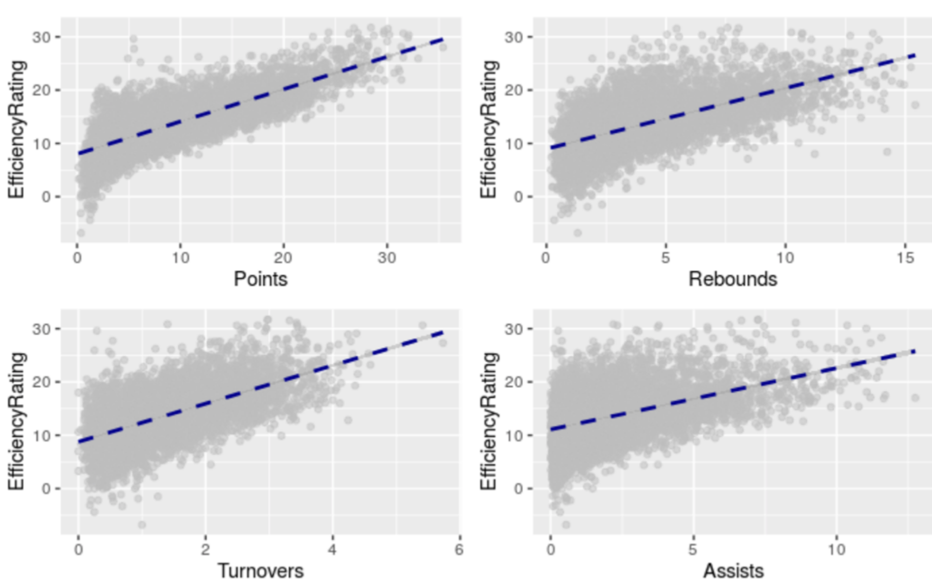
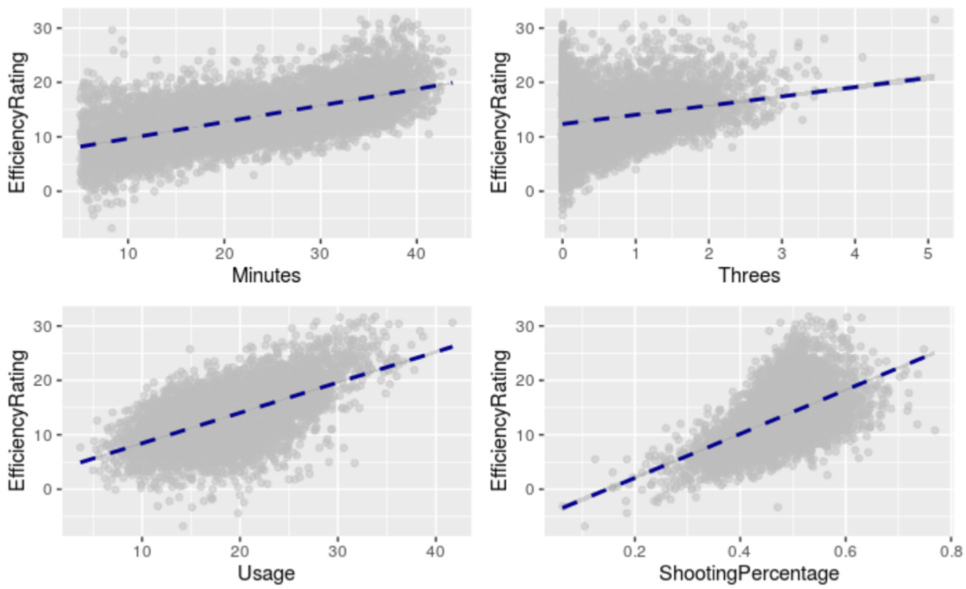
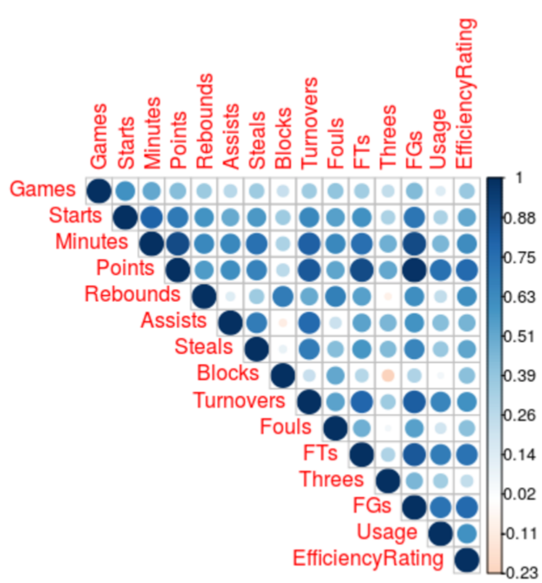


Figure 8



Finally, I plotted the correlation plot showing correlations between variables through these correlations vary in strength. There are expecting correlations across the board, for example, players who play more minutes have more chances to get high points and better statistics in the rest of variables. The negative correlations are expected to be the relationship between offensive and defensive stats, and I did not analyze them with efforts. I used the package corrplot in R to visualize the correlations with the directions and strengths from 0.23 to 1 by altering circles and colors.

Figure 9



Model selections

- **Logistic regression model**

Logistic regression model is selected to estimate the factors that best predict the All NBA selection. I used glm() command in R to fit the logistic model with binomial errors to investigate the relationships between variables (points, rebounds, assists, turnovers, minutes, threes, usage, and PER) and the All-NBA selection probability. I generated the basis of data frame to predict probability at each 0.1 point interval, and ran the prediction based on the glm model output. I graphed these predictions using a ggplot line graph method, and I added vertical lines to these graphs representing the mean and the 99th percentile for each predictor value. From **Figure 10** and **Figure 11 ,** at the mean level of output, players have almost zero chance of being selected for the All-NBA team. The vertical lines of 99th percentile of points plot and PER plot indicated that players have over 75% chance of the selection, but it is not the case for rebounds, assists, minutes, and threes. I saw that for usage and turnovers, output at this 99th percentile level gives players a 50% chance of selection. Therefore, from the univariable logistic regression analysis, point and PER are the most valuable metrices with high chance for All-NBA selection.

Figure 10

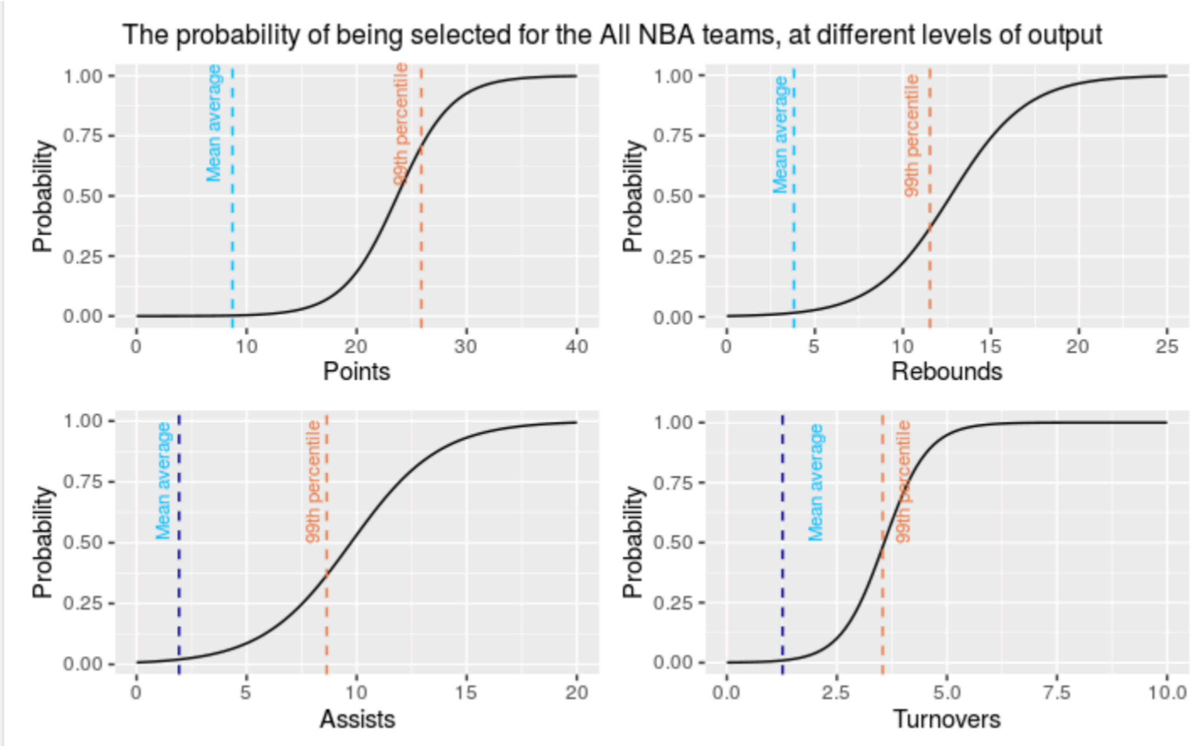
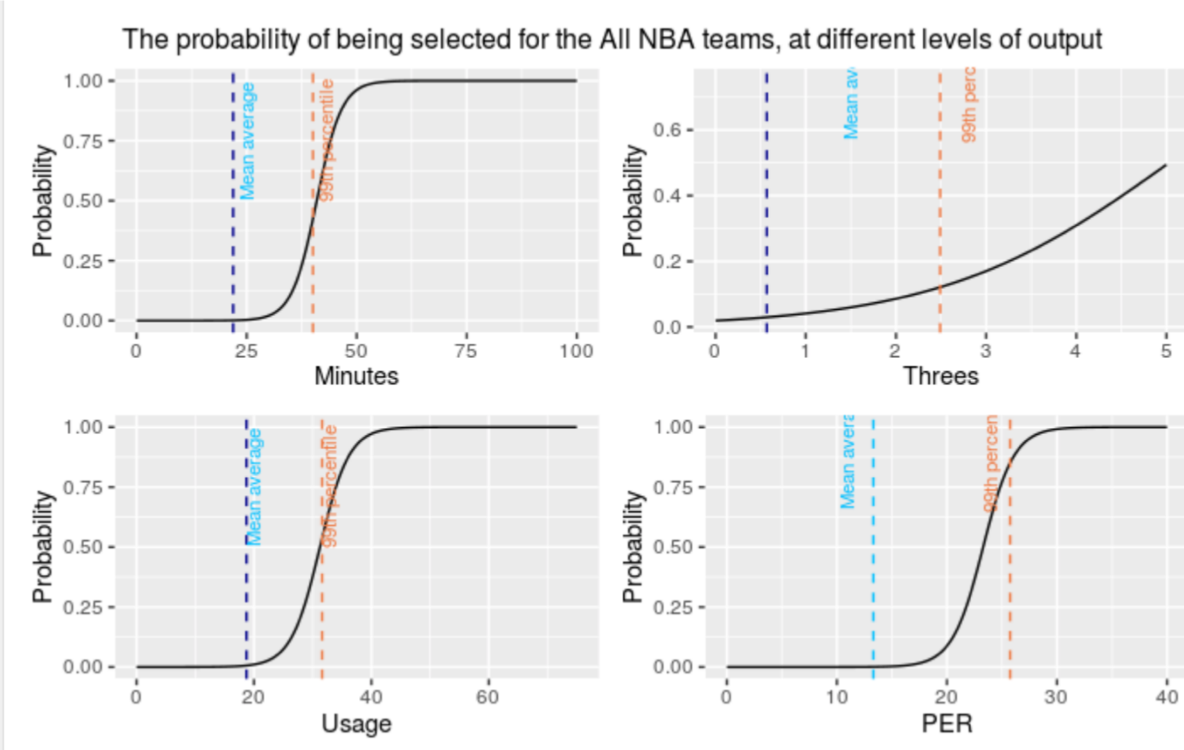


Figure 11



- Multivariate logistic regression

Next, I conducted the multivariate logistic regression model on all variables. The outputs are shown below:

Figure 12

Waiting for profiling to be done...

	Odds_Ratio	2.5 %	97.5 %
(Intercept)	0	0	0
Points	5.427e+07	0	9.495e+18
Rebounds	1.397	1.231	1.589
Assists	1.804	1.474	2.222
Usage	1.028	0.894	1.183
Threes	0	0	3183
FGs	0	0	1.294e+07
ShootingPercentage	1	1	1
EfficiencyRating	1.212	1.021	1.436
Steals	1.903	1.154	3.14
Blocks	2.833	1.883	4.305
Turnovers	0.56	0.293	1.058
Fouls	0.522	0.346	0.782
FTs	0	0	4469

```
Call:
glm(formula = All.NBA ~ Points + Rebounds + Assists + Usage +
    Threes + FGs + ShootingPercentage + EfficiencyRating + Steals +
    Blocks + Turnovers + Fouls + FTs, family = binomial, data = nba.pergame)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5637  -0.0423  -0.0115  -0.0038   4.5021

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.163e+01  2.440e+00  -8.864  < 2e-16 ***
Points             1.781e+01  1.316e+01   1.354  0.17588
Rebounds           3.343e-01  6.512e-02   5.133 2.85e-07 ***
Assists            5.899e-01  1.045e-01   5.643 1.67e-08 ***
Usage              2.809e-02  7.137e-02   0.394  0.69391
Threes            -1.769e+01  1.316e+01  -1.344  0.17887
FGs               -3.511e+01  2.631e+01  -1.335  0.18200
ShootingPercentage  1.252e-07  4.315e-08   2.902  0.00371 **
EfficiencyRating   1.923e-01  8.682e-02   2.215  0.02675 *
Steals             6.433e-01  2.551e-01   2.522  0.01166 *
Blocks             1.041e+00  2.107e-01   4.944 7.65e-07 ***
Turnovers         -5.802e-01  3.267e-01  -1.776  0.07573 .
Fouls             -6.495e-01  2.075e-01  -3.130  0.00175 **
FTs               -1.735e+01  1.316e+01  -1.318  0.18737
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2492.94  on 8362  degrees of freedom
Residual deviance:  764.44  on 8349  degrees of freedom
AIC: 792.44

Number of Fisher Scoring iterations: 10
```

The odds ratio is a statistic that quantifies the strength of the association between two events. Odds ratios center should be around 1. Values greater than 1 indicate a positive relationship, and values lower than 1 indicate a negative relationship. All the odd-ratios are positive indicating that it is reasonable to build accurate predictive model with these variables.

- **Random forest**

As I mentioned in define the problem section, this is the classification problem, which means that, the models classify players as a 1(All-NBA) or 0 (not ALL-NBA). This model also should return a probability for the remaining career path of a certain player. We can interpret of these as a certainty of sorts that a player with a 1.0 probability (100%) is a lock to make an All-NBA team. I chose the random forest algorithm to predict this selection. This algorithm ensures that the model is made of hundreds or thousands of decision trees using bootstrapping, random subsets of features, and averages results to make predictions. This cross-validation method ensures the less biased model. Moreover, I used an entire forest of trees, training each one on a random subsample of the training data. The final model then takes an average of all the individual decision trees to arrive at a classification.

I wanted this model to predict the future years of players of All-NBA selections, therefore, I separated the dataset based on year instead of the random samplings. The training dataset ranged from 1999 to 2011, which represents the past stats for each player`s career path. The testing dataset ranged from 2012 to 2017 ( six years) , which represents the remaining stats of each player`s career path. I used the dim function to get that 5584 rows and 24 columns are in the training dataset and 2779 rows and 24 columns are in the testing dataset.

Training and Validation

I used the RandomForest package to train the model using the training dataset. I set the seed of 100 to ensure the replicability, plotted the model errors and variables relative importance, and generated the summary report of this model. The outputs are shown in **Figure 13**.

From the generated summary report, there are total 500 trees in this model, and the OOB error at different number of decision trees including the random forest is 1.68%, meaning that an accuracy rate of 98.32%. Also, the confusion matrix is following:

	0	1	Class error
0	5358	31	0.005752
1	63	132	0.3230

These indicators showed optimism on the training dataset, but, most importantly, the real indicator will be how the model performs on the test dataset, and how many of the players can be selected it correctly predicts. The RF error plot matched the confusion matrix that the black line represents the overall OOB error rate, the red line represents errors for class 0 (not selected to All NBA; false positives) and the green line class 1 (selected to All NBA; i.e. false negatives). **Figure 15** confirmed that PER is a valuable metric for All-NBA selection.

Figure 13

```
                Length Class  Mode
call                3  -none- call
type                1  -none- character
predicted        5584  factor numeric
err.rate         1500  -none- numeric
confusion           6  -none- numeric
votes           11168  matrix numeric
oob.times        5584  -none- numeric
classes             2  -none- character
importance         18  -none- numeric
importanceSD        0  -none- NULL
localImportance     0  -none- NULL
proximity           0  -none- NULL
ntree               1  -none- numeric
mtry                1  -none- numeric
forest             14  -none- list
y                5584  factor numeric
test                0  -none- NULL
inbag               0  -none- NULL
terms               3  terms  call

Call:
 randomForest(formula = All.NBA ~ Points + Assists + Rebounds +      age + Games + Starts + Minutes + Steals + Blocks + Turnovers +      Fouls + FTs + Threes + FGs + Usage + EfficiencyRating + BoxPlusMinus +      ShootingPercentage, data = nba.train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 1.68%
Confusion matrix:
     0   1 class.error
0 5358  31 0.005752459
1   63 132 0.323076923
```

Figure 14

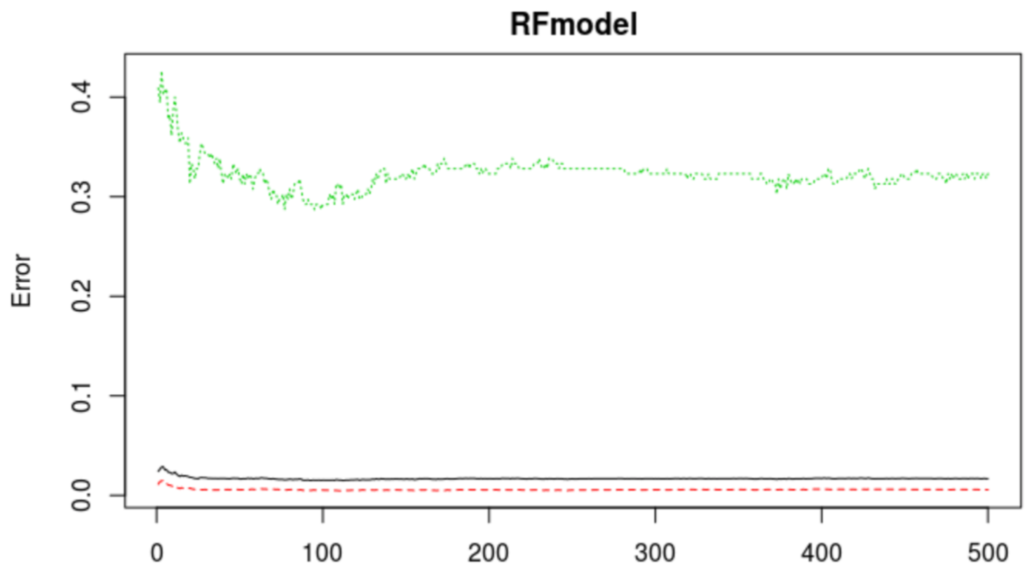
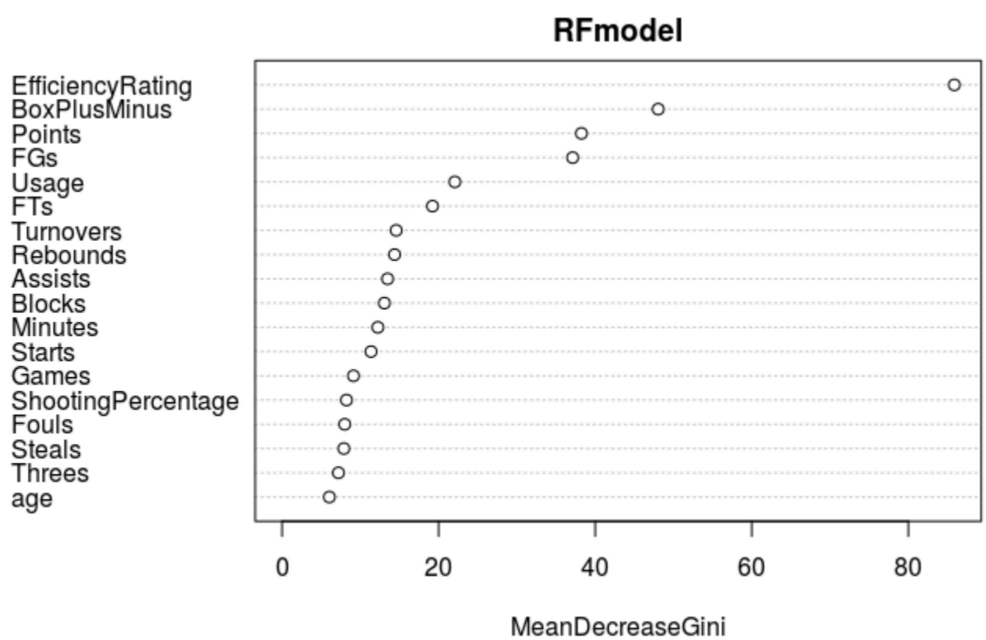


Figure 15



Evaluation of classification model accuracy

For validation, I used the trained algorithm on the trained per-game to make predictions on the test dataset, which is about the forecasting of which players were selected for All-NBA in the remaining career path. Next, I matched these predictions against the observed values to evaluate the accuracy of the predicted model. The true positives indicate that the algorithm correctly predicts selection, the true negatives indicate that the algorithm correctly predicts non-selection, the false positives indicate that the algorithm predicts selection, but the player was not selected, and the false negatives indicate the algorithm predicts non-selected, but the player was selected. These columns can be simply created with ifelse arguments in R. Finally, I created columns in the dataset for those parameters, and showed the result in **Figure 16**:

Figure 16

Type	Count
True Positive	53
True Negative	2675
False Positive	14
False Negative	37

Also, I interpreted the precision, sensitivity, true negative rate, and false negative rate. The algorithm correctly predicted 2675 of a possible 2712 non-selections, resulting in the high true negative rate (98.64%). But the algorithm only correctly identify 53 of the 90 All-NBA selections, resulting in moderate true positive rate(58.9%).

Figure 17

```
##precision; sensitivity (recall); specificity
p1 <- 53/(53+14)
p1

#true positive rate
TPR <- 53/(53+37)
TPR

TNR <- 2675/(2675+37)
TNR

FNR <- 1-TNR
FNR

table1 <- c(p1,TPR,TNR,FNR)
formattable(table1)
```

```
[1] 0.7910448
[1] 0.5888889
[1] 0.9863569
[1] 0.01364307
[1] 0.791   0.5889  0.9864  0.01364
```

Because of the moderate true positive rate, I have to do some adjestments on the current algorithm. One challenge for the algorithm is to build the season-sensitive option and interpret the probability of each player of selected All-NBA. Instead of type= binary response in the predict function, I put probability here and added the new column in dataset. Since the total numbers of selected All-NBA are fifteen, I used "top_n" function to create a new data frame of the 15 players in each season with highest probability values. Then, I calculted the average accuracy, resulting in 76.7%. It is a big improvement on previous true positive rate(58.9%). Moreover, I wanted to know how this algorithm matched up with advanced metric (e.g PER) in prediction. The PER metric related accuracy of top fifteen players is 58.7%, which showed the optimism on predicting All-NBA selections.

Figure 18

```
dim(nba.test.prob)
dim(nba.top15)

##season-specific model accuracy
which(nba.test.prob$All.NBA == 0 & nba.test.prob$Probability > 0.75)
length(which(nba.test.prob$All.NBA == 0 & nba.test.prob$Probability > 0.75))
percentage1 <- (4/2779)
percentage1

which(nba.test.prob$All.NBA == 1 & nba.test.prob$Probability < 0.5)
length(which(nba.test.prob$All.NBA == 1 & nba.test.prob$Probability < 0.5))
percentage2 <- (35/2779)
percentage2

accuracy <- 1-(percentage1+percentage2)
accuracy
|
```

```
[1] 76.6667
[1] 58.6957
[1] 2779   26
[1] 90 26
[1] 1007 2021 2405 2728
[1] 4
[1] 0.001439367
 [1]   14   78  166  310  321  346  593  599  613  644  700  911 1036 1069 1118 1169 1227 1239 1377 1502 1529 1553 1592 1615
[25] 1804 1861 1965 1997 2009 2086 2112 2268 2471 2480 2547
[1] 35
[1] 0.01259446
[1] 0.9859662
```

Figure 18 showed result of this season-specific algorithm accuracy. Based on the business rule, I found the numbers of selections where the algorithm is confident, but it is wrong:
1) The probability is larger than 0.75 , but the player is not selected.
2) The probability is less than 0.5, but the player is selected.
I used the length function to get the numbers, and calculated the accuracy rate, resulting in 98.5%.

Discussion

- Use the model to estimate for the players listed below

(1) the total number of All-NBA selections remaining in each of these player's careers.
(2) the likelihood of each player having the greatest number of All-NBA selections remaining among this group.
• Luka Doncic
• Karl Anthony-Towns
• Kyrie Irving
• Stephen Curry

I used the filter function on the prediction model, and got the results in **Figure 20:**
The length of the testing dataset contains 6 years, therefore, I can only use this model to predict the number of All-NBA selections for each player in the next 6 years.
If the probability is larger than 0.5, the player will be selected for All-NBA.

• Luka Doncic
He entered the game at 2018, but the testing dataset here is from 2012 to 2017. Therefore, his stats are blank here.

• Karl Anthony-Towns
1.The total number is 1 that he will be selected in All-NBA team in the next 6 years
2. The highest likelihood is 0.9 that he will be selected in All-NBA team in the next 6 years

• Kyrie Irving
1.The total number is 1 in the next six years that he will be selected in All-NBA team
2. The highest likelihood is 0.556 that he will be selected in All-NBA team in the next 6 years

• Stephen Curry
1.The total number is 5 in the next six years that he will be selected in All-NBA team
2. The highest likelihood is 0.898 that he will be selected in All-NBA team in the next 6 years

Figure 20

~~~~
```{r}
## predict the total numbers of All-NBA selections remaining in player`s career path
## predict the highest likelihood of each player for the selection
nba.top15$Name
filter(nba.test.prob,Name=='Luka Doncic')
filter(nba.test.prob,Name=='Karl-Anthony Towns')
filter(nba.test.prob,Name=='Kyrie Irving')
filter(nba.test.prob,Name=='Stephen Curry')

```
~~~~

R Console | data.frame 0 x 26 | data.frame 2 x 26 | data.frame 6 x 26 | data.frame 6 x 26

0 rows | 1-10 of 26 columns


~~~~
## predict the total numbers of All-NBA selections remaining in player`s career path
## predict the highest likelihood of each player for the selection
nba.top15$Name
filter(nba.test.prob,Name=='Luka Doncic')
filter(nba.test.prob,Name=='Karl-Anthony Towns')
filter(nba.test.prob,Name=='Kyrie Irving')
filter(nba.test.prob,Name=='Stephen Curry')

```
~~~~


R Console | data.frame 0 x 26 | data.frame 2 x 26 | data.frame 6 x 26 | data.frame 6 x 26

Threes	FGs	Usage	EfficiencyRating	BoxPlusMinus	ShootingPercentage	ID	All.NBA	0	Probability
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<fctr>	<dbl>	<dbl>
0.37	7.62	24.9	22.5	2.8	0.555	Kar20MIN16	0	0.518	0.482
1.23	9.78	27.5	26.0	5.0	0.576	Kar21MIN17	0	0.100	0.900

2 rows | 17-26 of 26 columns

````r
## predict the total numbers of All-NBA selections remaining in player`s career path
## predict the highest likelihood of each player for the selection
nba.top15$Name
filter(nba.test.prob,Name=='Luka Doncic')
filter(nba.test.prob,Name=='Karl-Anthony Towns')
filter(nba.test.prob,Name=='Kyrie Irving')
filter(nba.test.prob,Name=='Stephen Curry')

```
````

R Console | data.frame 0 x 26 | data.frame 2 x 26 | data.frame 6 x 26 | data.frame 6 x 26

Threes <dbl>	FGs <dbl>	Usage <dbl>	EfficiencyRating <dbl>	BoxPlusMinus <dbl>	ShootingPercentage <dbl>	ID <chr>	All.NBA <fctr>	0 <dbl>	Probability <dbl>
1.43	6.86	28.7	21.4	3.3	0.517	Kyr19CLE12	0	0.852	0.148
1.85	8.20	30.2	21.4	3.3	0.503	Kyr20CLE13	0	0.756	0.244
1.73	7.49	28.2	20.1	3.2	0.480	Kyr21CLE14	0	0.952	0.048
2.09	7.71	26.2	21.5	3.3	0.532	Kyr22CLE15	1	0.792	0.208
1.58	7.43	29.5	19.9	1.6	0.496	Kyr23CLE16	0	0.984	0.016
2.46	9.32	30.8	23.0	2.5	0.535	Kyr24CLE17	0	0.444	0.556


````r
## predict the total numbers of All-NBA selections remaining in player`s career path
## predict the highest likelihood of each player for the selection
nba.top15$Name
filter(nba.test.prob,Name=='Luka Doncic')
filter(nba.test.prob,Name=='Karl-Anthony Towns')
filter(nba.test.prob,Name=='Kyrie Irving')
filter(nba.test.prob,Name=='Stephen Curry')

```
````

R Console | data.frame 0 x 26 | data.frame 2 x 26 | data.frame 6 x 26 | data.frame 6 x 26

Threes <dbl>	FGs <dbl>	Usage <dbl>	EfficiencyRating <dbl>	BoxPlusMinus <dbl>	ShootingPercentage <dbl>	ID <chr>	All.NBA <fctr>	0 <dbl>	Probability <dbl>
2.12	5.58	24.0	21.2	3.4	0.583	Ste23GSW12	0	0.972	0.028
3.49	8.03	26.4	21.3	5.4	0.549	Ste24GSW13	0	0.488	0.512
3.35	8.36	28.3	24.1	7.4	0.566	Ste25GSW14	1	0.184	0.816
3.58	8.16	28.9	28.0	9.9	0.594	Ste26GSW15	1	0.162	0.838
5.09	10.19	32.6	31.5	12.5	0.630	Ste27GSW16	1	0.102	0.898
4.10	8.54	30.1	24.6	7.3	0.580	Ste28GSW17	1	0.178	0.822

6 rows | 17-26 of 26 columns

- **Limitations**

1. Each NBA team has 5 players: 2 guards, 2 forwards, and 1 center. But this model cannot predict 5 players for each team and to have the predictions follow the positional restrictions. Because prediction probabilities here give more insight than the classes themselves.
2. Since I can only use this model to predict the number of All-NBA selections for each player in the next 6 years, I need to find other reasonable datasets with larger year ranges than the original datasets. In that way, the model can be improved to predict All-NBA selections for each player in longer remaining time.