Machine Learning Report

Joshua Palamuttam, Michael Domke, Albert Udom

Week 1 – Introduction to Machine Learning:

The NBA All star game has been run from 1950 to present day. Currently there are 300 or so active players in the NBA and only 24 NBA all star roster spots. What is interesting about this is that the starting 5 from each team is voted in by NBA fans and the other 7 per team are voted in by coaches around the league, meaning that the all star play is not necessarily directly related to statistics in the NBA, but it would be interesting to note if there is a significant correlation between college play statistics and NBA all star caliber players.

This project could be used for NBA team scouts and general managers before the NBA draft to find out which players are worth drafting to the team because they will become an all star.

It would be interesting to note if a player would become an all star within a relatively short time after their draft date. There are only a handful who were voted into the all star game from their rookie NBA year.

# Week 2 – Learning to Manage Data:

1. We can load the data we found to R as a data frame.

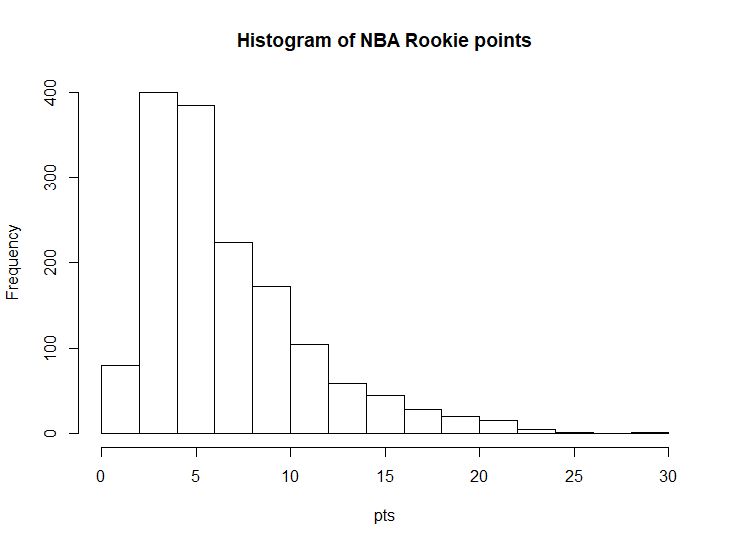
2. We found our data on kaggle for NBA all all stars, and data world for NBA rookie year statistics

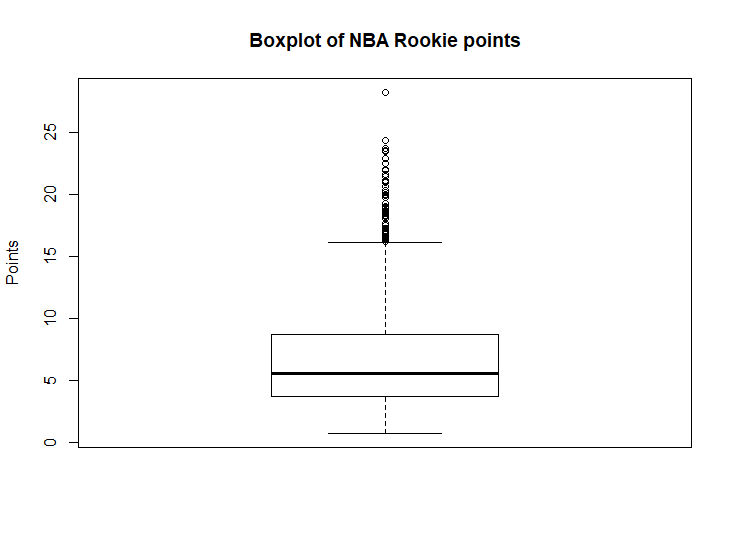
3.

summary(NBA\_Rookies$min)  
Min. 1st Qu. Median Mean 3rd Qu. Max.   
3.10 11.10 16.10 17.55 22.80 40.90   
> summary(NBA\_Rookies$pts)  
Min. 1st Qu. Median Mean 3rd Qu. Max.   
0.700 3.700 5.500 6.708 8.675 28.200

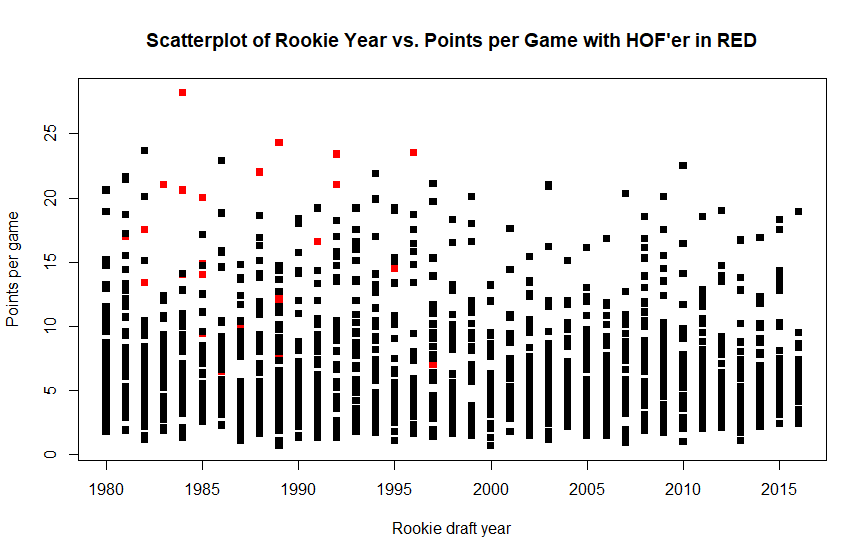
Here are some examples of what we tried to see correlations between amount of time played and points made during thier rookie year.

4. The Histogram and BoxPlot diagram attached to this response show more of a story for rookies. Obviously rookies did not make many points, but there are quite a few outliers shown in the boxplot diagram.

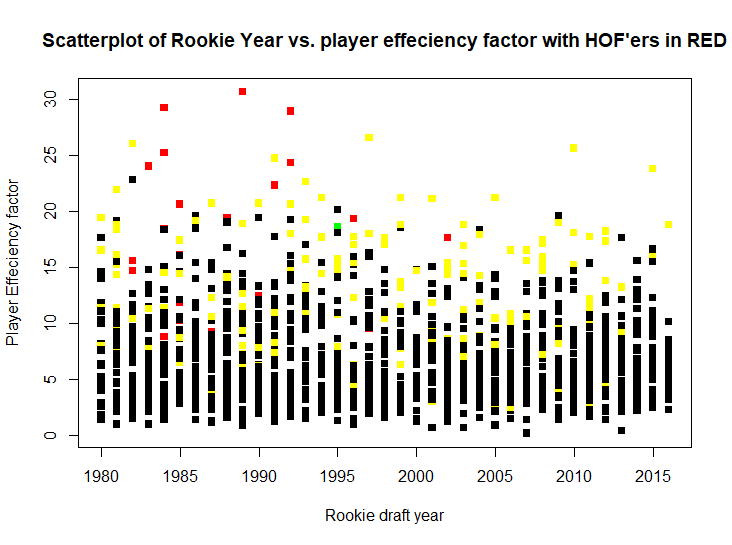
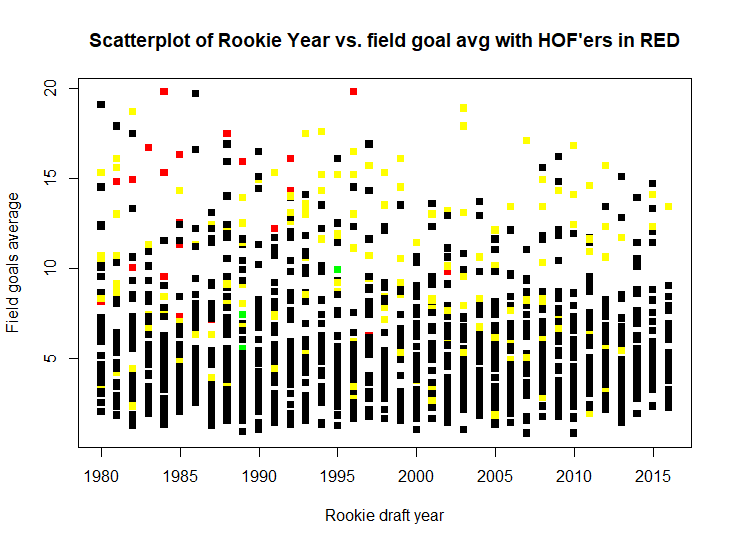


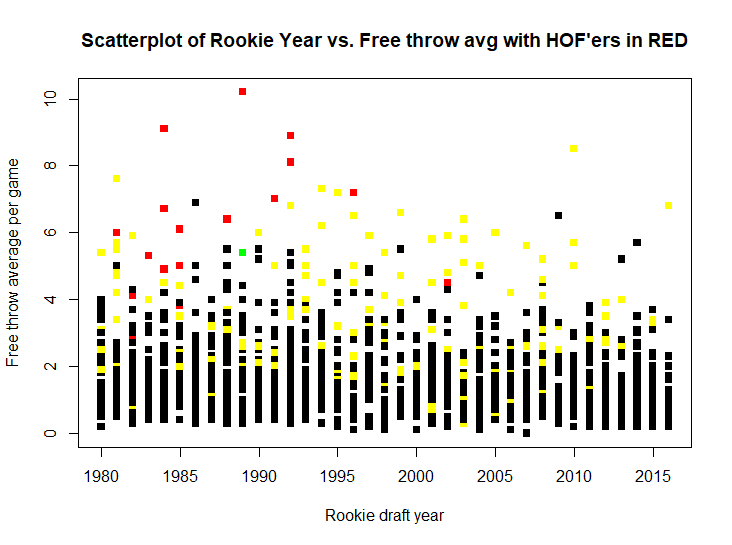
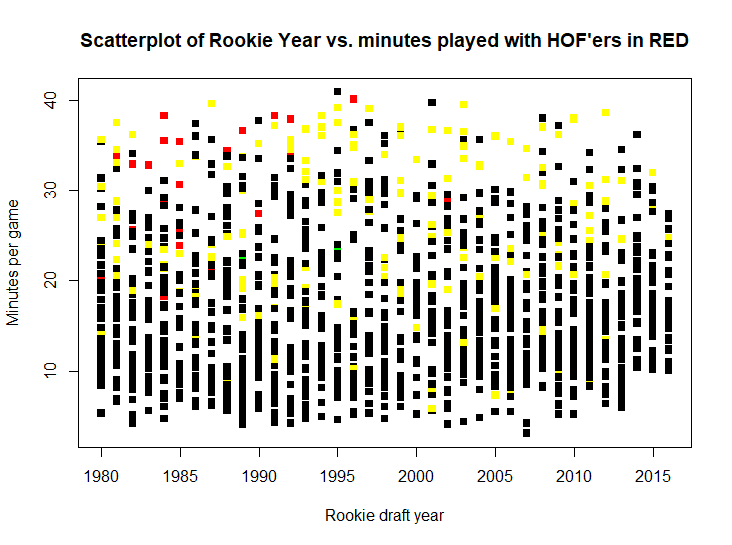
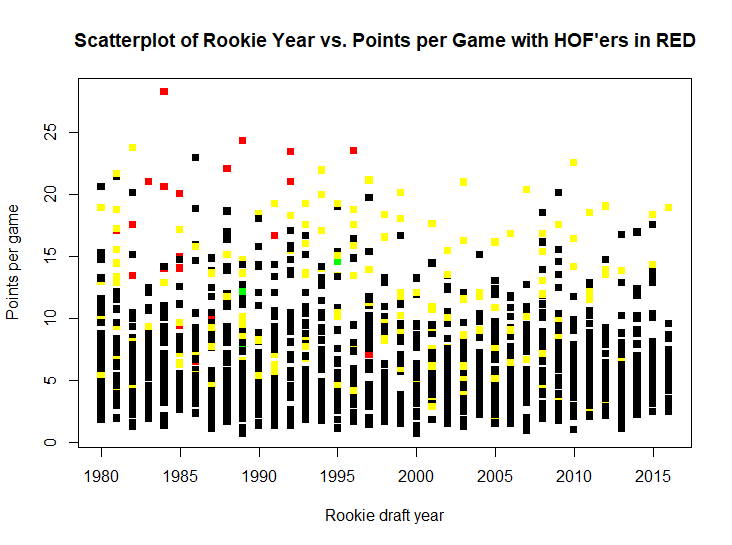
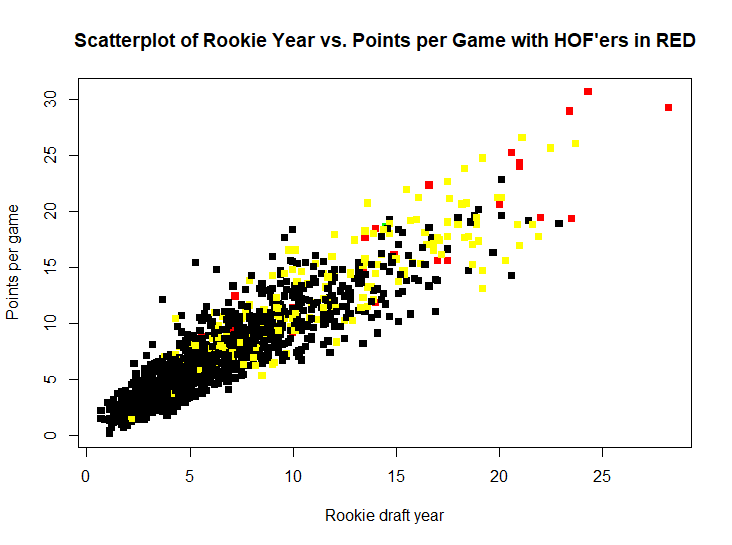
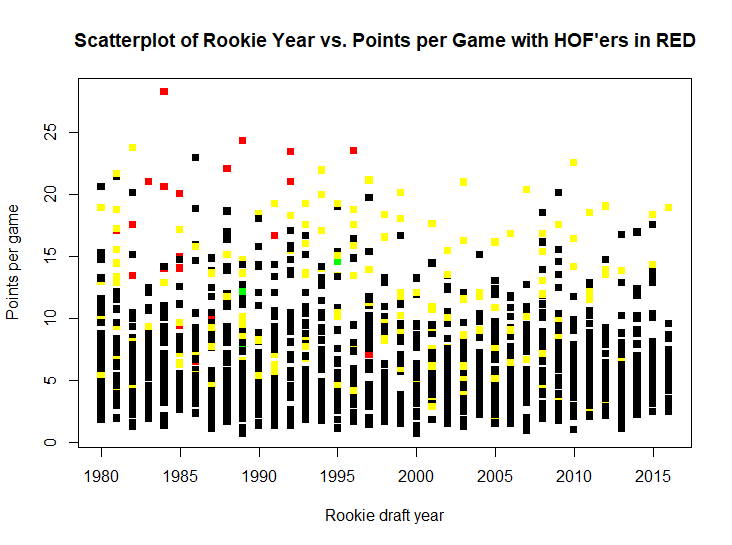
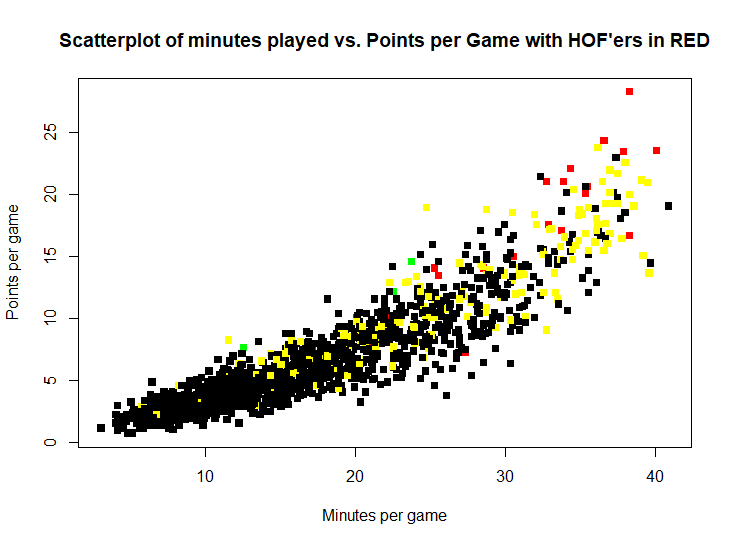


# Week 3 – Lazy Learning:

 The scatter plot above highlights HOF inductees when rookie year and points per game are compared. There is a cutoff of HOFers before 2000 because the rookies from this point on are still playing.

here is some more data that we have added since:

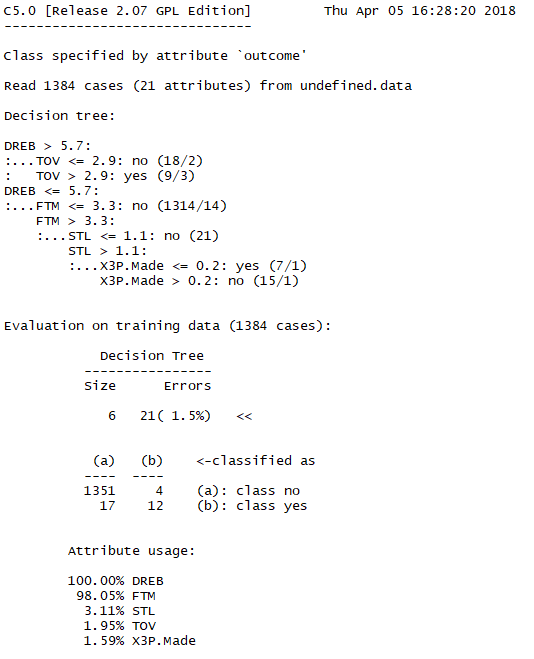
     

# Week 4 – Naïve Bayes:

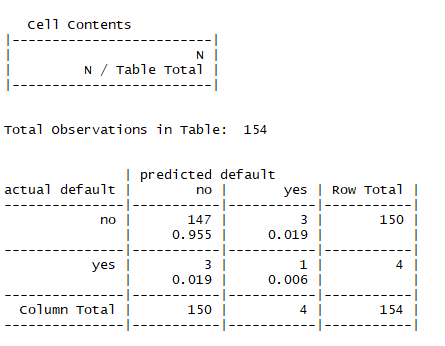
Because our data is completely numeric because we are dealing with statistics in the NBA, we could not implement Naive Bayes Algorithm as it calls for some semblance of categorical data, which we do not have.

# Week 5 – Decision Trees and Rules:

This algorithm may have been the best one so far for our prediction cases. Something interesting we noticed was that when training with 90% of our data, we got values saying that the model predicted Hall of Famers. But with 85% and 95% of our data in the training set, we did not get the model to predict any Hall of Famers.



The photo above is a screenshot of our decision tree with 90% of the data in the training class. Below is the test output.



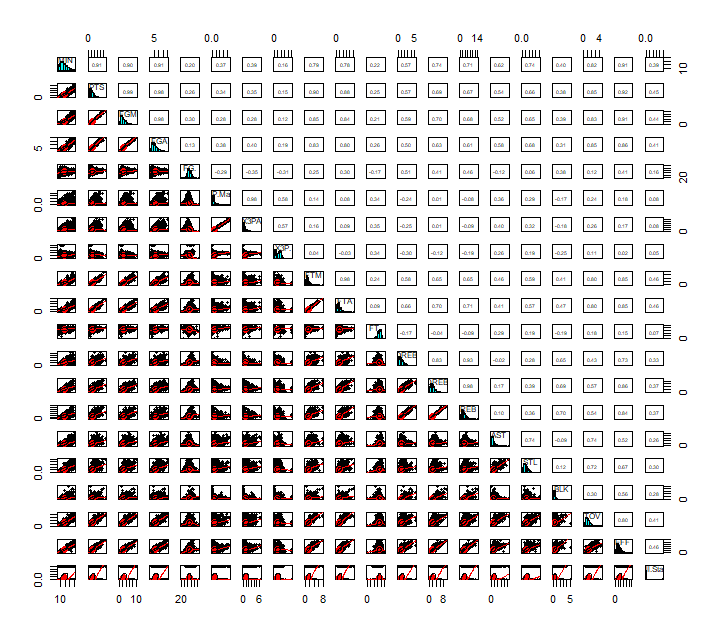
It can be seen here that our model successfully predicted 147 non-hall of famers and 1 hall of famer. Interestingly, it noted 3 false positives and 3 false negatives. Maybe when changing the seed or changing the amount of test data, we cold find better results. We will probably try boosting as well to see what changes we get from that.

# Week 6 - Regression:

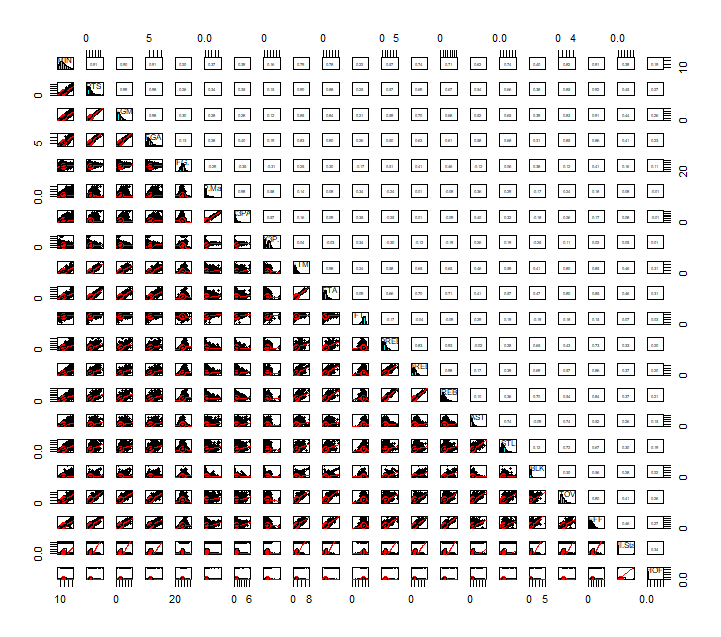
This week was a very successful week for us, Linear regression yielded some good results, but we think that the regression trees yielded the most beneficial results in our goal which is to predict with just stats who will be in the Hall of fame or who will be an All-star in the NBA.

First we used Linear Regression to come up with a plot for

All Stars:

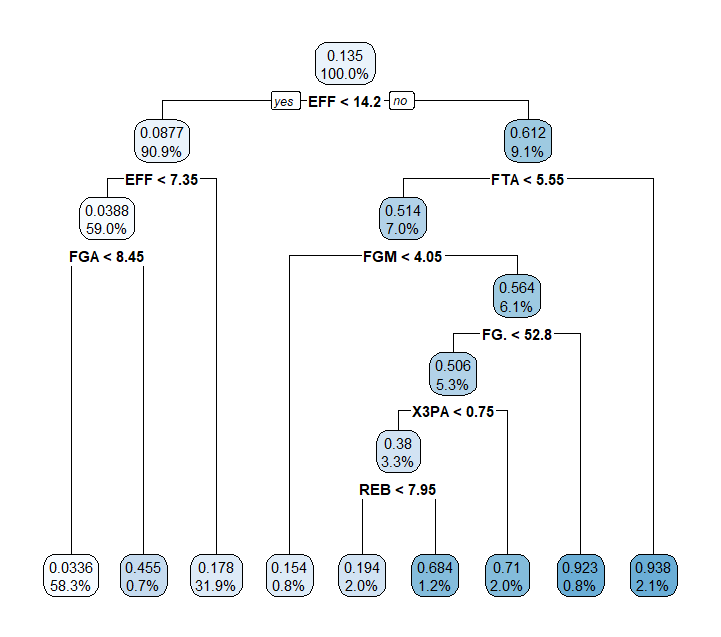


HOF:

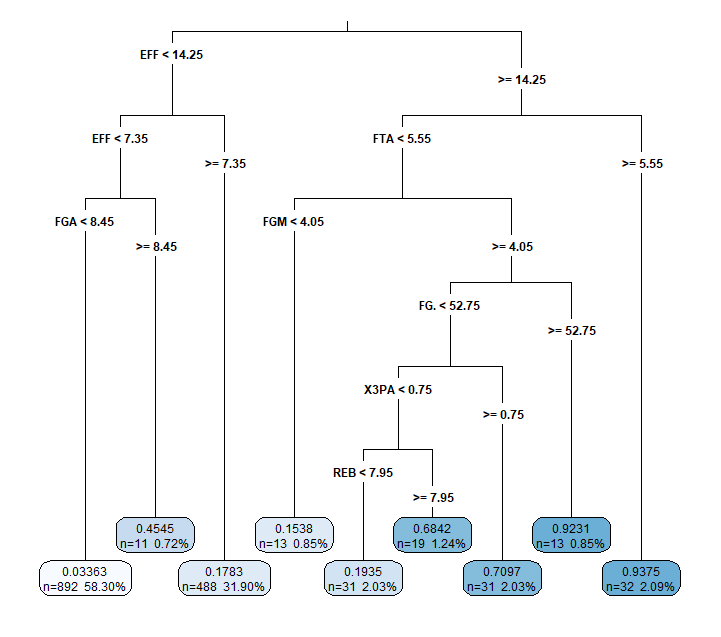


We then started to make a Regression Tree for each:

Our initial All-star regression tree looks like this:

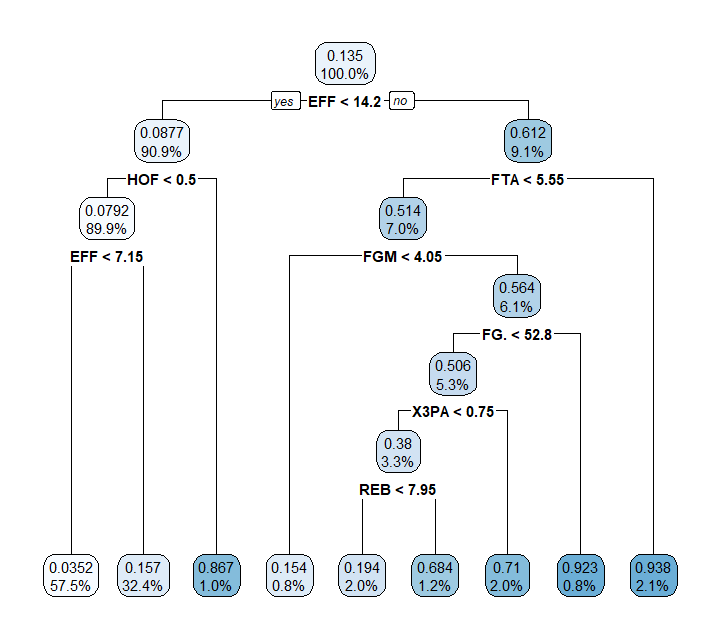


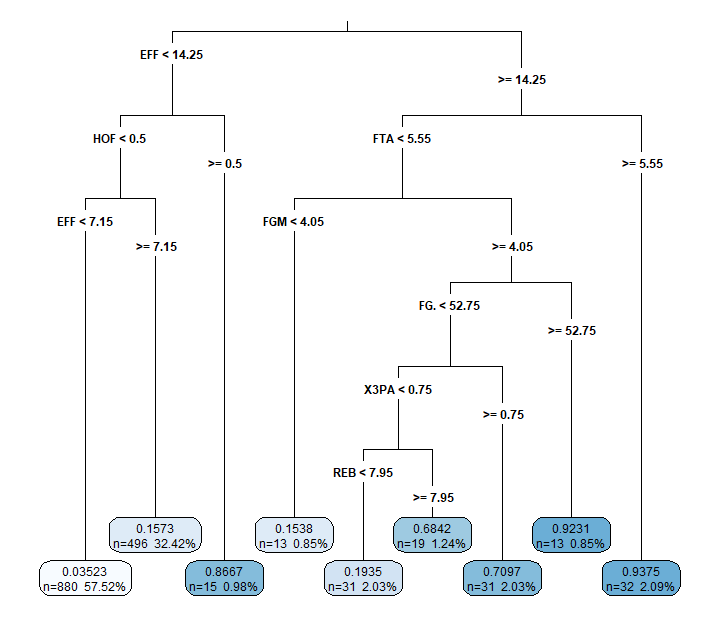
The final Regression tree looks like this:



We did the same for Hall of famers

initial:



Final: 

This type of machine learning algorithm fits our data well because our data is entirely numeric. As you can see the regression tree for hall of famer, there is a slight error where if you have a less than 14.25 effectiveness rating but are in the hall of fame, then you are in the hall of fame. We changed that, and in our new tree it values effectiveness, on the left side of that tree. What we thought was kind of interesting was that free throws were valued highly in this regression tree, we think it might be due to the fact that the player is playing more, and very active on the court. Other than that, the tree pretty much explains itself.

# Week 7 – Neural Nets and Support Vector Machines: C:\Users\Administrator\Downloads\NN 5 hidden plot.png

The other neural nets that we created are in the folder dataWithPictures.

# Week 8 – Market Basket Analysis:

Due to the numerical nature of our dataset, the Market Basket Analysis didn't work. Instead we cleaned up R script file we've used for our project throughout the quarter.

# Week 9 – Clustering with K-means:

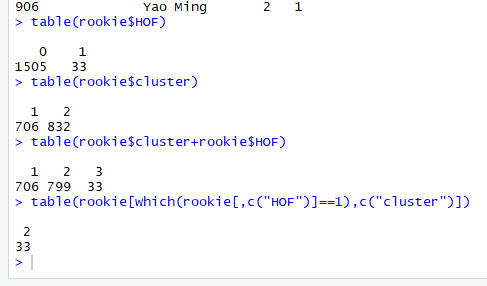
We started out using the clusted k-means algorithm and got these results for 2 clusters.

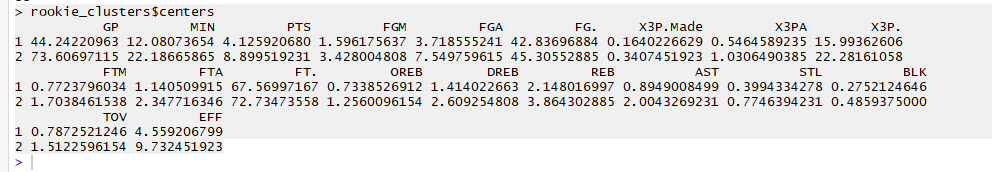
rookie\_clusters<-kmeans(stats,2)  
> rookie\_clusters$centers  
GP MIN PTS FGM FGA FG. X3P.Made X3PA X3P. FTM FTA FT. OREB DREB REB AST  
1 44.24221 12.08074 4.125921 1.596176 3.718555 42.83697 0.1640227 0.5464589 15.99363 0.7723796 1.140510 67.56997 0.7338527 1.414023 2.148017 0.8949008  
2 73.60697 22.18666 8.899519 3.428005 7.549760 45.30553 0.3407452 1.0306490 22.28161 1.7038462 2.347716 72.73474 1.2560096 2.609255 3.864303 2.0043269  
STL BLK TOV EFF  
1 0.3994334 0.2752125 0.7872521 4.559207  
2 0.7746394 0.4859375 1.5122596 9.732452

> rookie\_clusters$size  
[1] 706 832

> rookie$cluster<-rookie\_clusters$cluster  
> rookie[,c("cluster","Name", "PTS","All.Star","HOF")]  
cluster Name PTS All.Star HOF  
1 2 Kevin McHale 10.0 1 1  
2 2 James Donaldson 5.3 1 0  
3 2 Bill Laimbeer 9.8 1 0  
4 1 Kiki Vandeweghe 11.5 1 0  
5 2 Andrew Toney 12.9 1 0  
6 2 Joe Barry Carroll 18.9 1 0  
7 1 Jeff Wilkins 4.7 0 0  
8 2 Johnny Moore 7.4 0 0  
9 2 Larry Smith 9.6 0 0  
10 1 Reggie Carter 2.8 0 0  
11 1 Mike Harper 2.7 0 0  
12 2 Don Collins 11.5 0 0  
13 1 Earl Cureton 4.2 0 0  
14 2 Calvin Garrett 6.1 0 0  
15 2 Darwin Cook 11.2 0 0  
16 1 Lorenzo Romar 4.1 0 0  
17 1 James Wilkes 4.1 0 0  
18 1 Charles Whitney 7.4 0 0  
19 2 Larry Drew 6.6 0 0  
20 2 Kelvin Ransey 15.2 0 0  
21 1 Billy Reid 3.2 0 0  
22 2 Bill Hanzlik 5.4 0 0  
23 2 Mike Gminski 13.2 0 0  
24 2 Wayne Robinson 7.9 0 0  
25 2 Mike Woodson 4.7 0 0  
26 2 Darrell Griffith 20.6 0 0  
27 1 Rick Mahorn 4.8 0 0  
28 1 Dewayne Scales 4.9 0 0  
29 2 Sam Worthen 3.7 0 0  
30 1 Rickey Brown 4.0 0 0  
31 2 Reggie Johnson 10.2 0 0  
32 1 Mike Niles 2.6 0 0  
33 2 Lowes Moore 7.0 0 0  
34 1 John Duren 1.8 0 0  
35 2 Cedrick Hordges 8.4 0 0  
36 2 Michael Brooks 14.7 0 0  
37 1 Jerry Sichting 2.0 0 0  
38 2 Kyle Macy 8.1 0 0  
39 1 Edgar Jones 8.7 0 0  
40 1 Butch Carter 5.6 0 0  
41 2 Mike O'Koren 11.0 0 0  
42 2 Louis Orr 10.5 0 0  
43 1 Craig Shelton 4.3 0 0  
44 2 Isiah Thomas 17.0 1 1  
45 2 Steve Johnson 12.8 1 0  
46 2 Tom Chambers 17.2 1 0  
47 2 Mark Aguirre 18.7 1 0  
48 2 Kelly Tripucka 21.6 1 0  
49 2 Larry Nance 6.6 1 0  
50 2 Eddie Johnson 9.3 1 0  
51 2 Buck Williams 15.5 1 0  
52 2 Jeff Ruland 14.4 1 0  
53 2 Rolando Blackman 13.3 1 0  
54 1 Danny Ainge 4.1 1 0  
55 1 Charles Bradley 3.0 0 0  
56 1 Kurt Nimphius 5.3 0 0  
57 2 Darnell Valentine 6.4 0 0  
58 2 Jim Smith 2.9 0 0  
59 1 Franklin Edwards 3.6 0 0  
60 2 Gene Banks 9.6 0 0  
61 1 Kurt Rambis 4.6 0 0  
62 2 Herb Williams 11.5 0 0  
63 1 Jeff Lamp 4.6 0 0  
64 2 Rudy Macklin 7.0 0 0  
65 2 Jay Vincent 21.4 0 0  
66 1 Kevin McKenna 1.9 0 0  
67 1 Mark Radford 3.4 0 0  
68 1 Charlie Davis 3.8 0 0  
69 2 Elston Turner 8.3 0 0  
70 2 Danny Vranes 4.9 0 0  
71 1 Garry Witts 2.9 0 0  
72 1 Alex Bradley 3.5 0 0  
73 2 Orlando Woolridge 7.3 0 0  
74 1 Samuel Williams 6.1 0 0  
75 2 John Douglas 7.0 0 0  
76 1 Howard Wood 3.4 0 0  
77 2 Jim Brogan 6.3 0 0  
78 2 Peter Verhoeven 4.9 0 0  
79 2 Alton Lister 4.5 0 0  
80 2 Petur Gudmundsson 3.2 0 0  
81 2 Albert King 12.1 0 0  
82 1 Ray Blume 4.6 0 0  
83 1 Bobby Cattage 3.1 0 0  
84 1 Mike McGee 4.9 0 0  
85 2 Frank Johnson 10.7 0 0  
86 2 Kevin Loder 6.9 0 0  
87 2 Dan Schayes 7.9 0 0  
88 1 Ed Rains 3.9 0 0  
89 2 Dominique Wilkins 17.5 1 1  
90 2 James Worthy 13.4 1 1  
91 2 Sleepy Floyd 8.1 1 0  
92 2 Mark Eaton 4.3 1 0  
93 1 Ricky Pierce 2.2 1 0  
94 2 Terry Cummings 23.7 1 0  
95 1 Chris Engler 1.5 0 0  
96 1 Mark McNamara 2.2 0 0  
97 1 Cliff Levingston 5.6 0 0  
98 1 Bruce Flowers 4.9 0 0  
99 2 Walker Russell 2.7 0 0  
100 2 Lafayette Lever 7.8 0 0  
101 1 Darren Tillis 3.2 0 0  
102 2 Craig Hodges 9.9 0 0  
103 2 Terry Teagle 10.4 0 0  
104 2 Richard Anderson 5.2 0 0  
105 1 David Thirdkill 4.0 0 0  
106 2 Rob Williams 7.0 0 0  
107 2 LaSalle Thompson 5.4 0 0  
108 2 Lester Conner 4.9 0 0  
109 1 Bryan Warrick 4.0 0 0  
110 2 Quintin Dailey 15.1 0 0  
111 2 Rod Higgins 10.3 0 0  
112 2 Trent Tucker 8.4 0 0  
113 2 Paul Pressey 6.7 0 0  
114 1 Rory White 5.0 0 0  
115 1 Eddie Phillips 3.2 0 0  
116 2 Clark Kellogg 20.1 0 0  
117 1 Rickey Williams 3.3 0 0  
118 2 Jerry Eaves 9.3 0 0  
119 2 Bill Garnett 6.3 0 0  
120 1 Jose Slaughter 3.6 0 0  
121 2 Russ Schoene 6.2 0 0  
122 1 Oliver Robinson 2.9 0 0  
123 2 John Bagley 5.7 0 0  
124 1 Jeff Taylor 3.6 0 0  
125 1 Brook Steppe 4.0 0 0  
126 2 Mitchell Anderson 7.4 0 0  
127 1 Scott Hastings 1.2 0 0  
128 1 Dave Batton 3.3 0 0  
129 2 Marc Iavaroni 5.1 0 0  
130 1 Corny Thompson 2.8 0 0  
131 1 Linton Townes 4.5 0 0  
132 2 Ed Nealy 4.4 0 0  
133 2 Edmund Sherod 6.2 0 0  
134 1 Jim Johnstone 1.3 0 0  
135 2 Clyde Drexler 7.7 1 1  
136 2 Ralph Sampson 21.0 1 1  
137 2 Jeff Malone 12.1 1 0  
138 2 Doc Rivers 9.3 1 0  
139 2 Dale Ellis 8.2 1 0  
140 2 Mitchell Wiggins 12.4 0 0  
141 1 Russell Cross 3.7 0 0  
142 1 Pace Mannion 2.1 0 0  
143 2 Derek Harper 5.7 0 0  
144 1 Sidney Green 5.2 0 0  
145 2 Rodney McCray 10.8 0 0  
146 1 Larry Micheaux 3.1 0 0  
147 1 Randy Breuer 2.9 0 0  
148 1 Dane Suttle 6.5 0 0  
149 2 Jim Thomas 6.3 0 0  
150 2 Darren Daye 6.1 0 0  
151 1 Bob Hansen 2.7 0 0  
152 2 Ennis Whatley 8.4 0 0  
153 1 Bruce Kuczenski 1.9 0 0  
154 2 Byron Scott 10.6 0 0  
155 2 Thurl Bailey 8.5 0 0  
156 2 Fred Roberts 7.3 0 0  
157 2 Steve Stipanovich 12.0 0 0  
158 1 Leroy Combs 4.5 0 0  
159 1 Greg Kite 1.9 0 0  
160 2 Roy Hinson 5.5 0 0  
161 2 Jon Sundvold 6.9 0 0  
162 1 Sedale Threatt 3.3 0 0  
163 1 John Paxson 2.9 0 0  
164 1 Stewart Granger 4.5 0 0  
165 2 Granville Waiters 3.6 0 0  
166 2 Rod Foster 8.3 0 0  
167 2 Paul Thompson 9.0 0 0  
168 2 Sidney Lowe 4.2 0 0  
169 2 Darrell Walker 7.9 0 0  
170 1 Howard Carter 6.2 0 0  
171 1 Scooter McCray 2.7 0 0  
172 2 Randy Wittman 4.5 0 0  
173 2 Michael Jordan 28.2 1 1  
174 2 Hakeem Olajuwon 20.6 1 1  
175 2 Charles Barkley 14.0 1 1  
176 2 John Stockton 5.6 1 1  
177 2 Kevin Willis 9.3 1 0  
178 2 Alvin Robertson 9.2 1 0  
179 2 Otis Thorpe 12.8 1 0  
180 2 Sam Perkins 11.0 0 0  
181 2 Jeff Turner 5.8 0 0  
182 1 Cory Blackwell 3.4 0 0  
183 1 Rick Carlisle 1.8 0 0  
184 2 Othell Wilson 4.4 0 0  
185 2 Ken Bannister 6.8 0 0  
186 1 Stuart Gray 2.0 0 0  
187 2 Michael Holton 8.4 0 0  
188 2 Mel Turpin 10.6 0 0  
189 1 Jim Petersen 3.2 0 0  
190 1 Leon Wood 3.2 0 0  
191 2 Terence Stansbury 7.1 0 0  
192 2 Chuck Aleksinas 5.1 0 0  
193 2 Tim McCormick 9.3 0 0  
194 2 Jerome Kersey 6.1 0 0  
195 1 Ozell Jones 3.7 0 0  
196 1 Charlie Sitton 2.1 0 0  
197 2 Vern Fleming 14.1 0 0  
198 1 Eddielee Wilkins 5.5 0 0  
199 2 Antoine Carr 8.0 0 0  
200 2 Charles Jones 8.4 0 0  
[ reached getOption("max.print") -- omitted 1338 rows ]

Using k-means and 2 clusters, we were able to sort all of the HOFers into a single cluster:



 We then tried doubling the clusters to 4: 