Daniel Trevino

04/17/19

Predicting NBA Champions for the 2018-2019 Season

Purpose: In order to identify an NBA team winning a game, multiple decisional events must occur. This project is designed to develop an understanding which events and statistics per NBA Conference lead to winning a game and losing a game. Upon identification of the key attributes and decisional events, the aim is to predict the probability of an Eastern or Western Conference NBA team winning.

Problem: How to develop an accurate classification model that will allow for a prediction of winning greater than 75%?

Data Questions:

(a) What characteristics or events need to occur for an NBA team to win a game?
(b) What is the level of significance of each variable, related to winning basketball games?
(c) Are all attributes in the data required to predict the outcome of an NBA game?

What Data do we need:

(1) Basic Team Box scores
   (a) Points per Game, Shooting percentages, Assists, Steals, Blocks, Rebounds, Turnovers
(2) Structured data set to perform analysis
   (a) The data is structured in tables on the following website. A python script was written to pull the NBA season box scores for every NBA game in 2018-2019. Below is a sample of the data set after it was translated to the .csv.
   www.basketball-reference.com/leagues/NBA_2019.htm

Example Team Data Set:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	Rk	Team	G	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
2	1	New Orleans Pelicans	62	240	43.4	91	0.477	10	28.9	0.345	33.4	62.1	0.538	18.9	24.5	0.769	11	36	47.1	26.9	7.3	5.7	14.4	21.4	115.6
3	2	Golden State Warriors	60	241.7	44.2	90.1	0.49	13	33.9	0.384	31.2	56.2	0.555	17.5	21.6	0.811	10.1	36.4	46.5	29.6	7.5	6.6	14	21.6	118.8
4	3	Los Angeles Clippers	62	242	40.8	87	0.469	9.6	25.1	0.382	31.2	61.9	0.504	22.8	28.7	0.794	9.5	35.5	45	23.4	6.6	4.8	14.7	23.3	114
5	4	Philadelphia 76ers	61	242	41.4	87.4	0.474	11.1	31	0.36	30.3	56.5	0.537	21.6	28	0.773	10.5	36.5	47	27.3	7.6	5.5	15.5	21.7	115.7
6	5	Milwaukee Bucks	60	240.8	43.3	90.3	0.479	13.2	37.8	0.351	30	52.6	0.572	17.2	22.4	0.77	9	39.9	48.9	26.1	7.7	5.9	14.1	19.8	117
7	6	Toronto Raptors	61	242	42	89.8	0.468	11.6	33.2	0.349	30.5	56.6	0.538	18.5	23	0.805	10.2	34.9	45.2	24.7	8.4	5.2	13.8	21.3	114.1
8	7	San Antonio Spurs	62	241.6	42.1	88.4	0.476	10	25	0.401	32	63.3	0.506	17.9	21.6	0.826	9.4	34.9	44.3	24.5	6.1	4.5	12.4	18.5	112.1
9	8	Brooklyn Nets	62	244	40.2	89.3	0.45	12.6	35.4	0.357	27.6	53.9	0.511	19	25.2	0.752	11.1	35	46.1	24	6.5	4.2	15.2	21.7	112
10	9	Oklahoma City Thunder	59	242.1	43.3	93.9	0.461	11	31.3	0.352	32.2	62.5	0.515	18.4	25.6	0.719	12.2	35.8	48.1	23.4	10.2	5.3	14.1	22.5	116
11	10	Washington Wizards	60	243.3	42.1	89.7	0.47	11.6	33.5	0.346	30.5	56.2	0.543	18	23.6	0.762	9	31.8	40.8	26.6	8.7	4.8	14.2	21.3	113.8

Data Mining Tasks Overview:

The following tasks are used to assess the data and evaluate prediction results.

(1) Summary Statistics
   (a) The goal of summary statistics is to give an overview of what the data is trying to portray. It shows the Mean, Median, Min, Max, Totals and Distributions of the data.

Daniel Trevino

04/17/19

(2) Linear Regression
    (a) The goal for linear regression is to find the attributes that have have a high level of significance <0.05, in predicting wins for a team.
(3) Association Rules Mining
    (a) The goal for association rules mining is to find the attributes that create the best confidence and lift that result in team wins.
(4) Naive Bayes
    (a) The goal for naive bayes is to find the probability of specific conditions with a goal of direct correlation to wins.
(5) Support Vector Machines
    (a) The goal of support vector machines is to use the data to perform regression analysis, the historic data should allow us to be able to predict possible future results.

Evaluate the Data:

Master Structured Data

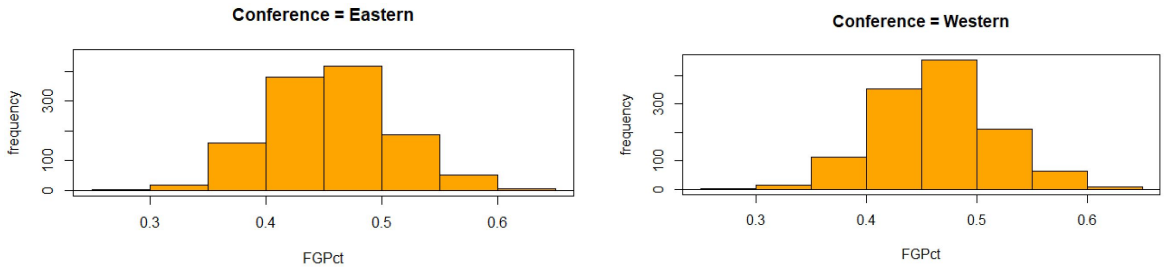
(1) Create a master data file, then seperate into the two conferences (Western / Eastern)
(2) Compare Conference teams with class attribute set to Win (1) or Loss (0)
(3) Create Win / Loss column in the data for 2019 season

Perform Summary Statistics

(1) The following output describes the descriptive statistics from the master NBA data file.

```
                Team         Conference        Division        MP              FG              FGA             FGPct            X3P             X3PA            X3PPct           FT
 Boston Celtics     :  82   Eastern:1219   Atlantic :407   Min.   :240.0   Min.   :25.00   Min.   : 64.00   Min.   :0.2780   Min.   : 2.00   Min.   :12.00   Min.   :0.115   Min.   : 2.0
 Cleveland Cavaliers:  82   Western:1219   Central  :406   1st Qu.:240.0   1st Qu.:38.00   1st Qu.: 85.00   1st Qu.:0.4260   1st Qu.: 9.00   1st Qu.:27.00   1st Qu.:0.296   1st Qu.:13.0
 Houston Rockets    :  82                  Northwest:405   Median :240.0   Median :41.00   Median : 89.00   Median :0.4600   Median :11.00   Median :32.00   Median :0.351   Median :17.0
 Los Angeles Lakers :  82                  Pacific  :407   Mean   :241.6   Mean   :41.05   Mean   : 89.16   Mean   :0.4614   Mean   :11.35   Mean   :31.98   Mean   :0.355   Mean   :17.7
 New Orleans Pelicans:  82                 Southeast:406   3rd Qu.:240.0   3rd Qu.:44.00   3rd Qu.: 94.00   3rd Qu.:0.4950   3rd Qu.:14.00   3rd Qu.:37.00   3rd Qu.:0.409   3rd Qu.:22.0
 Phoenix Suns       :  82                  Southwest:407   Max.   :340.0   Max.   :61.00   Max.   :123.00   Max.   :0.6490   Max.   :27.00   Max.   :70.00   Max.   :0.842   Max.   :44.0
 (Other)            :1946
      FTA             FTPct              ORB              DRB              TRB              AST              STL              BLK              TOV              PF              PTS             WinLoss
 Min.   : 4.0   Min.   :0.2630   Min.   : 1.00   Min.   :18.0   Min.   :22.00   Min.   :10.00   Min.   : 0.000   Min.   : 0.000   Min.   : 3.00   Min.   : 9.00   Min.   : 68.0   Min.   :0.0
 1st Qu.:18.0   1st Qu.:0.7000   1st Qu.: 8.00   1st Qu.:31.0   1st Qu.:41.00   1st Qu.:21.00   1st Qu.: 6.000   1st Qu.: 3.000   1st Qu.:11.00   1st Qu.:18.00   1st Qu.:103.0   1st Qu.:0.0
 Median :23.0   Median :0.7720   Median :10.00   Median :35.0   Median :45.00   Median :24.00   Median : 7.000   Median : 5.000   Median :13.00   Median :21.00   Median :111.0   Median :0.5
 Mean   :23.1   Mean   :0.7671   Mean   :10.34   Mean   :34.8   Mean   :45.15   Mean   :24.57   Mean   : 7.647   Mean   : 4.963   Mean   :13.56   Mean   :20.93   Mean   :111.2   Mean   :0.5
 3rd Qu.:28.0   3rd Qu.:0.8400   3rd Qu.:13.00   3rd Qu.:38.0   3rd Qu.:50.00   3rd Qu.:28.00   3rd Qu.:10.000   3rd Qu.: 6.000   3rd Qu.:16.00   3rd Qu.:24.00   3rd Qu.:119.0   3rd Qu.:1.0
 Max.   :54.0   Max.   :1.0000   Max.   :26.00   Max.   :55.0   Max.   :71.00   Max.   :42.00   Max.   :20.000   Max.   :19.000   Max.   :27.00   Max.   :38.00   Max.   :168.0   Max.   :1.0
```

Generating a few histograms of the data shows the data falls within normal standard distribution..



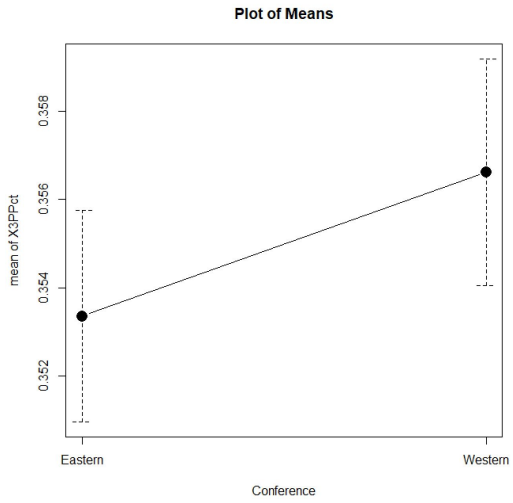
The following structured table shows the data collected and how it is organized. There are 2329 observations (rows) in the dataset “teamtotals”.

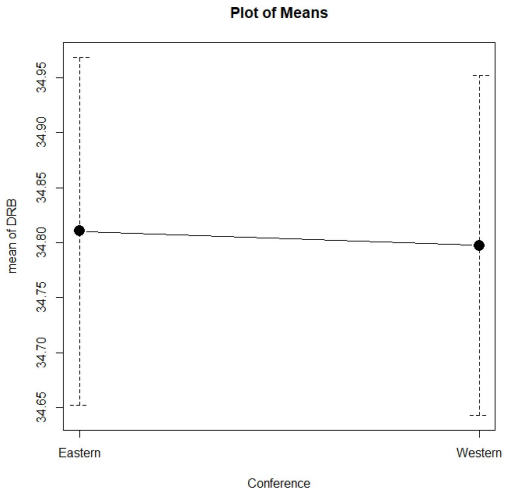
Team	Conference	Division	MP	FG	FGA	FGPct	3P	3PA	3PPct	FT	FTA	FTPct	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	WinLoss
Philadelphia 76ers	Eastern	Atlantic	240	34	87	0.391	5	26	0.192	14	23	0.609	6	41	47	18	8	5	16	20	87	0
Boston Celtics	Eastern	Atlantic	240	42	97	0.433	11	37	0.297	10	14	0.714	12	43	55	21	7	5	14	20	105	1
Oklahoma City Thunder	Western	Northwest	240	33	91	0.363	10	37	0.27	24	37	0.649	16	29	45	21	12	6	14	21	100	0
Golden State Warriors	Western	Pacific	240	42	95	0.442	7	26	0.269	17	18	0.944	17	41	58	28	7	7	21	29	108	1
Milwaukee Bucks	Eastern	Central	240	42	85	0.494	14	34	0.412	15	20	0.75	11	46	57	26	5	4	21	25	113	1

Daniel Trevino

04/17/19

This plots below show the statistical differences between the two major attributes of the conferences (3 Point Percentage and Defensive Rebounds).





Linear Regression

(1) To proceed with further analysis and model development, the data must be linear. The following RESET test was performed Is the NBA team total dataset linear. The Ramsey Regression Equation Error Test (RESET) as seen from the output below maintains a p-value greater than 0.05 and therefore the dataset is linear.

```
Rcmdr>  resettest(WinLoss ~ AST + BLK + DRB + FG + FGA + FGPct + FT + FTA + FTPct +
Rcmdr+    ORB + PF + PTS + STL + TOV + TRB + X3P + X3PA + X3PPct, power=2:3,
Rcmdr+    type="regressor", data=teamtotals)

        RESET test

data:  WinLoss ~ AST + BLK + DRB + FG + FGA + FGPct + FT + FTA + FTPct +     ORB + PF + PTS + STL + TOV + TRB + X3P + X3PA + X3PPct
RESET = 1.3209, df1 = 36, df2 = 2273, p-value = 0.09661
```

(2) The output to the right shows the linear model from the dataset. The linear regression highlights the most statistically significant attributes within the data are Defensive Rebounds, Blocks, Field Goal Attempts, Offensive Rebounds, Steals and Turnovers. The model does not fully incorporate the varying differences between the two conferences and therefore does not give an accurate depiction of the different playing styles in conferences.

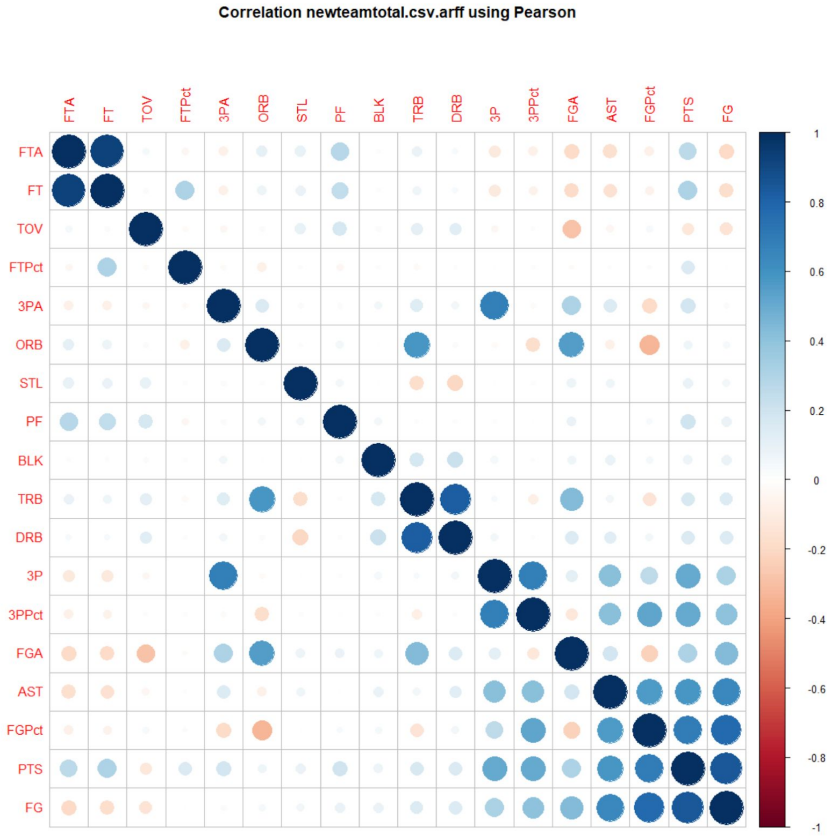
```
lm(formula = WinLoss ~ AST + BLK + DRB + FG + FGA + FGPct + FT +
    FTA + FTPct + MP + ORB + PF + PTS + STL + TOV + TRB + X3P +
    X3PA + X3PPct, data = newteamtotals)

Residuals:
     Min       1Q   Median       3Q      Max
-1.08596 -0.27578  0.00307  0.26753  1.05484

Coefficients: (2 not defined because of singularities)
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.069671   0.831842   0.084  0.933258
AST          0.002749   0.001952   1.408  0.159299
BLK          0.011258   0.002962   3.800  0.000148 ***
DRB          0.041503   0.001481  28.025   < 2e-16 ***
FG           0.017200   0.032358   0.532  0.595085
FGA         -0.035686   0.008987  -3.971 0.0000738 ***
FGPct        0.251769   1.691573   0.149  0.881695
FT           0.007563   0.015139   0.500  0.617415
FTA         -0.008978   0.007382  -1.216  0.224036
FTPct        0.141275   0.210365   0.672  0.501919
MP          -0.001342   0.001111  -1.208  0.227254
ORB          0.039937   0.002633  15.170   < 2e-16 ***
PF          -0.007743   0.001852  -4.181 0.0000301 ***
PTS          0.010309   0.011776   0.875  0.381408
STL          0.044232   0.002608  16.962   < 2e-16 ***
TOV         -0.036870   0.002255 -16.352   < 2e-16 ***
TRB               NA         NA      NA        NA
X3P               NA         NA      NA        NA
X3PA         0.001858   0.004390   0.423  0.672197
X3PPct       0.611176   0.356219   1.716  0.086339 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3523 on 2420 degrees of freedom
Multiple R-squared:  0.5073,    Adjusted R-squared:  0.5038
F-statistic: 146.6 on 17 and 2420 DF,  p-value: < 2.2e-16
```

Daniel Trevino

04/17/19

Correlation between Attributes

The correlation table below shows most of the data is centered around the bottom right of the table, with attributes centered around scoring points, a primary objective of basketball. Thus, we will use further classification algorithms to develop a further understanding of the data.



Classification Development

The following models are evaluated with the data, decision tree, SVM, random forest, and Naive Bayes, but the best performing model (Support Vector Machines) is used to evaluate the NBA conferences. Of the models listed, SVM returned the highest model prediction at 84% correct instance classification. See the appendix for the other model details.

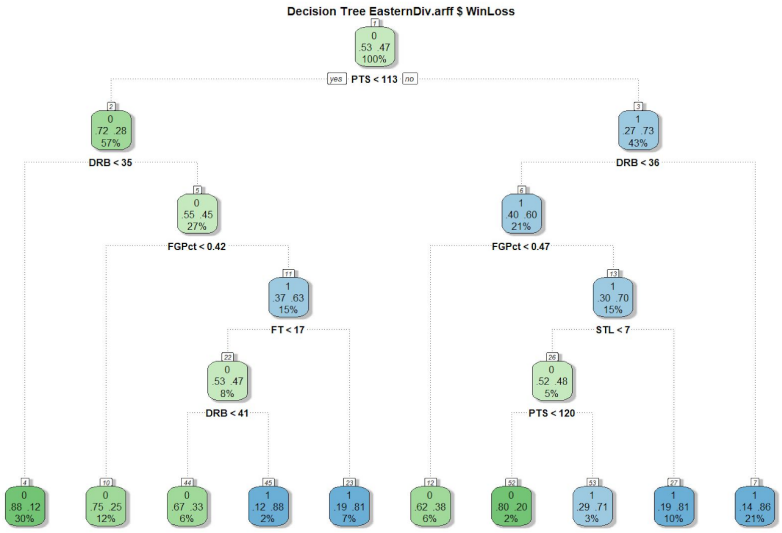
The excerpt on the right from Weka shows the weight each attribute. This equation, similar to linear regression, can be used to predict how the conference compete against each other. As seen, more field goal attempts leads to a negative weight.

```
SMO

Kernel used:
  Linear Kernel: K(x,y) = <x,y>

Classifier for classes: 0, 1

BinarySMO

Machine linear: showing attribute weights, not support vectors.

        0.7975 * (normalized) FG
 +     -5.27   * (normalized) FGA
 +      4.8171 * (normalized) FGPct
 +      1.7389 * (normalized) 3P
 +     -0.3745 * (normalized) 3PA
 +      1.8605 * (normalized) 3PPct
 +      0.8459 * (normalized) FT
 +      0.2315 * (normalized) FTA
 +      1.2149 * (normalized) FTPct
 +      2.0181 * (normalized) ORB
 +      4.2938 * (normalized) DRB
 +      4.2719 * (normalized) TRB
 +      0.5309 * (normalized) AST
 +      3.8668 * (normalized) STL
 +      0.7577 * (normalized) BLK
 +     -3.8252 * (normalized) TOV
 +     -1.5138 * (normalized) PF
 +      1.3468 * (normalized) PTS
 -      7.7833
```

Daniel Trevino

04/17/19

Conference Comparisons
Eastern Conference
Which stats are most valid for win in eastern conf. (arules) / decision tree



When trying to predict a win in the eastern conference, we can see through the decision tree that first, the ideal point range is greater than 113 points scored in the game. Next defense is most important, with defensive rebounds being a key attribute to winning in the East, ideally greater than 36 defensive rebounds. Finally steals and free throws made are the final deciding keys to winning in the East.

```
Apriori
=======

Minimum support: 0.25 (145 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Size of set of large itemsets L(2): 11

Best rules found:

 1. FGA='(89-94]' 189 ==> WinLoss=1 189    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 2. BLK='(4.8-6.4]' 170 ==> WinLoss=1 170    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 3. FTA='(20.4-25.2]' 167 ==> WinLoss=1 167    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 4. PTS='(113.4-121.2]' 164 ==> WinLoss=1 164    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 5. 3PA='(32.2-37]' 157 ==> WinLoss=1 157    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 6. FT='(16.7-20.6]' 157 ==> WinLoss=1 157    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 7. FGA='(84-89]' 154 ==> WinLoss=1 154    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 8. 3PPct='(0.3317-0.3796]' 152 ==> WinLoss=1 152    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 9. DRB='(34-37.5]' 151 ==> WinLoss=1 151    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
10. STL='(7.6-9.5]' 149 ==> WinLoss=1 149    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
```

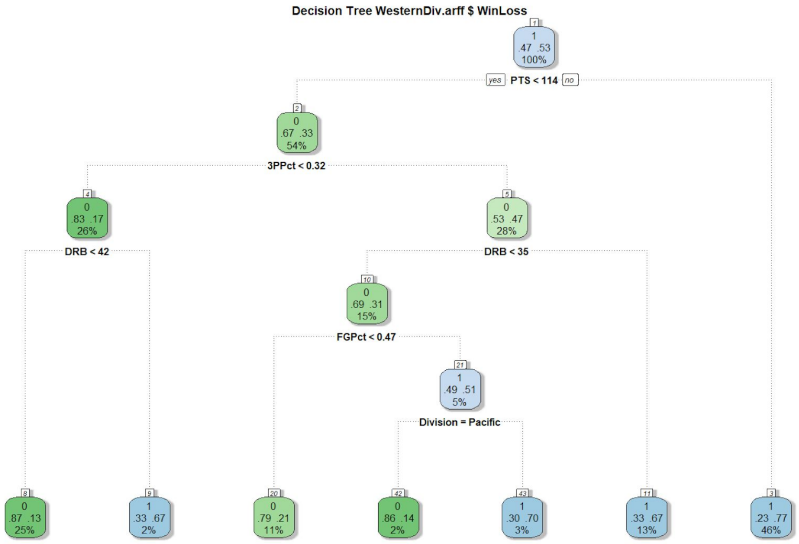
Daniel Trevino

04/17/19

Next the association rule mining model is used to gather more insight into win chance in the eastern conference. The association rules support the key attributes of the decision tree. The top rules that result in a Win are Blocks, Free throws attempted Defensive rebounds and Steals. From these two models we can see that the eastern conference is a defense oriented conference, relying on defense to win games.

Western Conference

Which stats are most valid for win in western conf. (arules) /decision tree



When trying to predict a win in the western conference, we can see through the decision tree that first, is that the ideal point range is greater than 114 points scored in a game. Next the key attribute for the west is 3 point percentage, ideally greater than 32%. The final key is defensive rebounds.

```
Apriori
=======

Minimum support: 0.25 (159 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 13

Size of set of large itemsets L(2): 12

Best rules found:

 1. 3PPct='(0.3744-0.4412]' 198 ==> WinLoss=1 198    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 2. 3PPct='(0.3076-0.3744]' 195 ==> WinLoss=1 195    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 3. BLK='(4.8-6.4]' 194 ==> WinLoss=1 194    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 4. PF='(19.4-22]' 189 ==> WinLoss=1 189    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 5. FGA='(86.8-92.5]' 188 ==> WinLoss=1 188    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 6. 3PA='(24.8-30.2]' 188 ==> WinLoss=1 188    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 7. FT='(16.6-20]' 172 ==> WinLoss=1 172    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 8. FGA='(81.1-86.8]' 171 ==> WinLoss=1 171    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 9. 3P='(7.8-10.2]' 169 ==> WinLoss=1 169    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
10. AST='(22.8-26]' 168 ==> WinLoss=1 168    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
```

Daniel Trevino

04/17/19

The association rules model again supports the attributes presented in the decision tree. It can be seen that 3 point percentage is a top rule that resulted in a win. A few other attributes that are within rules are blocks, personal fouls, and assists. From these two models we can see that the western conference is more offense oriented, specifically depending on shooting threes.

This alone will not be able to predict which conference is better, who will win more, or which style of play is better. We can however support that idea that there are different playing styles. There are different attributes that are significantly correlated to winning in the eastern and western conference. From this we can assume that the eastern conference is a defense first conference and the western conference is an offense first conference, specifically through shooting 3 point shots.

Prediction Model Comparison

SVM and Random Forest Model Comparison (Western Conference)

SVM Model (Weka)

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1022               83.8392 %
Incorrectly Classified Instances       197               16.1608 %
Kappa statistic                          0.6764
Mean absolute error                      0.1616
Root mean squared error                  0.402
Relative absolute error                 32.3826 %
Root relative squared error             80.477  %
Total Number of Instances             1219

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.838    0.161    0.850      0.838    0.844      0.676    0.838     0.797     0
               0.839    0.162    0.826      0.839    0.832      0.676    0.838     0.770     1
Weighted Avg.  0.838    0.162    0.839      0.838    0.838      0.676    0.838     0.784

=== Confusion Matrix ===

   a   b   <-- classified as
 533 103 |   a = 0
  94 489 |   b = 1
```

Random Forest (Weka)

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         963               78.9992 %
Incorrectly Classified Instances       256               21.0008 %
Kappa statistic                          0.5797
Mean absolute error                      0.3329
Root mean squared error                  0.3964
Relative absolute error                 66.7012 %
Root relative squared error             79.3577 %
Total Number of Instances             1219

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.786    0.206    0.806      0.786    0.796      0.580    0.854     0.873     0
               0.794    0.214    0.773      0.794    0.783      0.580    0.854     0.820     1
Weighted Avg.  0.790    0.210    0.790      0.790    0.790      0.580    0.854     0.847

=== Confusion Matrix ===

   a   b   <-- classified as
 500 136 |   a = 0
 120 463 |   b = 1
```

Comparing the Western vs Eastern conference and predict with SVM model for winning

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     WesternDiv-weka.filters.unsupervised.attribute.NumericToNominal-R23-weka.filters.unsupervised.attribute.Remove-R4
Instances:     unknown (yet). Reading incrementally
Attributes:   22

=== Summary ===

Correctly Classified Instances        1007               82.6087 %
Incorrectly Classified Instances       212               17.3913 %
Kappa statistic                          0.6512
Mean absolute error                      0.1739
Root mean squared error                  0.417
Total Number of Instances             1219

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.808    0.157    0.825      0.808    0.816      0.651    0.825     0.758     0
               0.843    0.192    0.827      0.843    0.835      0.651    0.825     0.779     1
Weighted Avg.  0.826    0.175    0.826      0.826    0.826      0.651    0.825     0.769

=== Confusion Matrix ===

   a   b   <-- classified as
 471 112 |   a = 0
 100 536 |   b = 1
```

SVM and Random Forest Model Comparison (Eastern Conference)

Daniel Trevino

04/17/19

SVM Model (Weka)

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1043               85.5619 %
Incorrectly Classified Instances       176               14.4381 %
Kappa statistic                          0.711
Mean absolute error                      0.1444
Root mean squared error                  0.38
Relative absolute error                 28.9307 %
Root relative squared error             76.0668 %
Total Number of Instances             1219

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.861    0.149    0.841      0.861    0.851      0.711    0.856     0.790     0
               0.851    0.139    0.870      0.851    0.860      0.711    0.856     0.818     1
Weighted Avg.  0.856    0.144    0.856      0.856    0.856      0.711    0.856     0.805

=== Confusion Matrix ===

   a   b   <-- classified as
 502  81 |   a = 0
  95 541 |   b = 1
```

Random Forest (Weka)

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         937               76.8663 %
Incorrectly Classified Instances       282               23.1337 %
Kappa statistic                          0.5357
Mean absolute error                      0.3489
Root mean squared error                  0.4041
Relative absolute error                 69.9144 %
Root relative squared error             80.9011 %
Total Number of Instances             1219

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.738    0.203    0.769      0.738    0.753      0.536    0.845     0.832     0
               0.797    0.262    0.768      0.797    0.782      0.536    0.845     0.849     1
Weighted Avg.  0.769    0.234    0.769      0.769    0.768      0.536    0.845     0.841

=== Confusion Matrix ===

   a   b   <-- classified as
 430 153 |   a = 0
 129 507 |   b = 1
```

Using SVM Model for comparison between western vs eastern conference and predict with SVM model for winning

```
=== Summary ===

Correctly Classified Instances         995               81.6243 %
Incorrectly Classified Instances       224               18.3757 %
Kappa statistic                          0.6276
Mean absolute error                      0.1838
Root mean squared error                  0.4287
Relative absolute error                 36.6822 %
Root relative squared error             85.4923 %
Total Number of Instances             1219

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.948    0.328    0.759      0.948    0.843      0.650    0.810     0.747     0
               0.672    0.052    0.922      0.672    0.778      0.650    0.810     0.777     1
Weighted Avg.  0.816    0.196    0.837      0.816    0.812      0.650    0.810     0.761

=== Confusion Matrix ===

   a   b   <-- classified as
 603  33 |   a = 0
 191 392 |   b = 1
```

The SVM model above represents the Eastern NBA conference vs the Western conference. This model illustrates a true positive rate of .672 for winning and a .948 rate for losing.

Actionable Insight:

(1) Assessing the data above and the outputs related to the Eastern and Western conference, the first actional sight is related to the Western conference. As seen above in the decision trees and datatable, the Western conference teams need to keep the Eastern conference teams from defensive rebounds, by crashing the boards and boxing out the Eastern conference team.

(2) Our second actionable insight, based on the results above, in order for a team in the Eastern conference to win against the Western conference they need to play to their strengths which is defense. Specifically they need to stop the Western conference teams from shooting and making 3 point shots to have the best chance at winning the game. We suggest implementing a 2-3 zone and mixing it up with a 3-2 zone. The reason that this is effective is because mixing up the defense keeps the offense guessing and the 3-2 zone specifically defends the 3 point line really well.

Daniel Trevino

04/17/19

Conclusions:

Winning a game in the NBA, with the goal of winning the NBA finals includes multiple variables, some measurable and some not. We focused on the main measurable variables that best predict winning in the NBA and trying to predict the Eastern or Western Conference team that will win the NBA Finals. All our models predicted with varying accuracies, however they produced significant variables that can be used to predict winning. We were able to develop decision trees and other models that produced an accuracy greater than 75%. The data supports that for Western conference teams ideally they need to shoot the 3 very well to have a better chance at winning and the Eastern Conference needs to play good defense and get a lot of defensive rebounds to have the best chance at winning. Each variable reported on had a level of significance that was at least less than .05. Overall, there are many ways to win a game and eventually win the NBA finals but what we have found best supports a significant path to predicting what team will win.

Daniel Trevino

04/17/19

Appendix

Full league Association Rules for Win:

```
Apriori
=======

Minimum support: 0.25 (291 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 14

Size of set of large itemsets L(2): 13

Best rules found:

 1. FGA='(86.8-92.5]' 379 ==> WinLoss=1 379    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 2. 3PA='(29.5-35]' 377 ==> WinLoss=1 377    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 3. 3P='(9.9-12.2]' 372 ==> WinLoss=1 372    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 4. 3PPct='(0.3744-0.4412]' 370 ==> WinLoss=1 370    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 5. 3PPct='(0.3076-0.3744]' 356 ==> WinLoss=1 356    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 6. BLK='(4.8-6.4]' 349 ==> WinLoss=1 349    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 7. PTS='(112.7-120.6]' 343 ==> WinLoss=1 343    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 8. PF='(19.4-22]' 336 ==> WinLoss=1 336    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 9. STL='(6.4-8.2]' 330 ==> WinLoss=1 330    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
10. AST='(22.8-26]' 327 ==> WinLoss=1 327    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
```

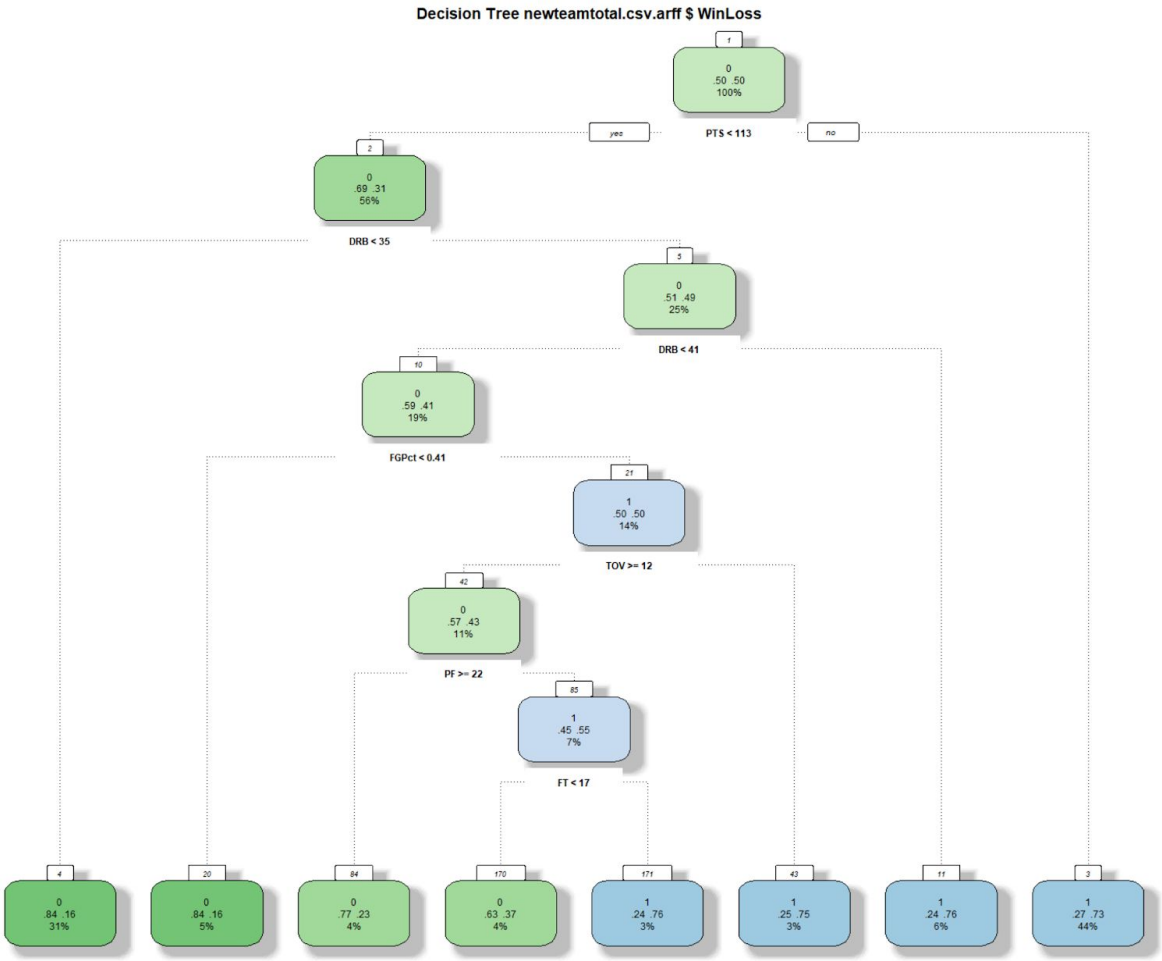
Full league Association Rules for loss:

```
Apriori
=======

Minimum support: 0.25 (291 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 27

Size of set of large itemsets L(2): 26

Best rules found:

 1. FG='(35.8-39.4]' 391 ==> WinLoss=0 391    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 2. BLK='(3.8-5.7]' 385 ==> WinLoss=0 385    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 3. FGA='(88.2-94]' 371 ==> WinLoss=0 371    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 4. 3PA='(29.4-35.2]' 368 ==> WinLoss=0 368    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 5. ORB='(8.5-11]' 367 ==> WinLoss=0 367    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 6. PTS='(95.9-105.2]' 363 ==> WinLoss=0 363    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 7. 3PPct='(0.2486-0.3154]' 358 ==> WinLoss=0 358    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 8. 3PPct='(0.3154-0.3822]' 356 ==> WinLoss=0 356    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 9. FGA='(82.4-88.2]' 354 ==> WinLoss=0 354    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
10. PTS='(105.2-114.5]' 345 ==> WinLoss=0 345    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
```

Daniel Trevino

04/17/19

Full league Decision Tree (Rcmdr):



Full league Decision Tree (Weka):

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1701               73.067  %
Incorrectly Classified Instances       627               26.933  %
Kappa statistic                          0.4613
Mean absolute error                      0.2927
Root mean squared error                  0.5011
Relative absolute error                 58.5347 %
Root relative squared error            100.2208 %
Total Number of Instances             2328

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.748     0.287     0.723       0.748   0.735      0.462    0.706     0.652     0
               0.713     0.252     0.739       0.713   0.726      0.462    0.706     0.658     1
Weighted Avg.  0.731     0.269     0.731       0.731   0.731      0.462    0.706     0.655

=== Confusion Matrix ===

   a   b   <-- classified as
 871 293 |   a = 0
 334 830 |   b = 1
```

Daniel Trevino

04/17/19

Full league SVM (Weka):

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1957               84.0636 %
Incorrectly Classified Instances       371               15.9364 %
Kappa statistic                          0.6813
Mean absolute error                      0.1594
Root mean squared error                  0.3992
Relative absolute error                 31.8729 %
Root relative squared error             79.8409 %
Total Number of Instances             2328

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.845    0.164    0.837      0.845    0.841      0.681    0.841     0.785     0
               0.836    0.155    0.844      0.836    0.840      0.681    0.841     0.787     1
Weighted Avg.  0.841    0.159    0.841      0.841    0.841      0.681    0.841     0.786

=== Confusion Matrix ===

   a   b   <-- classified as
 984 180 |   a = 0
 191 973 |   b = 1
```

Full league Naive Bayes (Weka):

```
=== Summary ===

Correctly Classified Instances        1789               76.8471 %
Incorrectly Classified Instances       539               23.1529 %
Kappa statistic                          0.5369
Mean absolute error                      0.2575
Root mean squared error                  0.4165
Relative absolute error                 51.4959 %
Root relative squared error             83.3086 %
Total Number of Instances             2328

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.768    0.231    0.769      0.768    0.768      0.537    0.849     0.854     0
               0.769    0.232    0.768      0.769    0.769      0.537    0.849     0.838     1
Weighted Avg.  0.768    0.232    0.768      0.768    0.768      0.537    0.849     0.846

=== Confusion Matrix ===

   a   b   <-- classified as
 894 270 |   a = 0
 269 895 |   b = 1
```

Full league Random Forest (Weka):

```
=== Summary ===

Correctly Classified Instances        1795               77.1048 %
Incorrectly Classified Instances       533               22.8952 %
Kappa statistic                          0.5421
Mean absolute error                      0.3497
Root mean squared error                  0.4006
Relative absolute error                 69.9374 %
Root relative squared error             80.1145 %
Total Number of Instances             2328

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.765    0.223    0.774      0.765    0.770      0.542    0.852     0.855     0
               0.777    0.235    0.768      0.777    0.772      0.542    0.852     0.848     1
Weighted Avg.  0.771    0.229    0.771      0.771    0.771      0.542    0.852     0.851

=== Confusion Matrix ===

   a   b   <-- classified as
 891 273 |   a = 0
 260 904 |   b = 1
```

Daniel Trevino

04/17/19

Compare eastern conf vs all nba data prediction model

```
=== Summary ===

Correctly Classified Instances        1023               83.9212 %
Incorrectly Classified Instances       196               16.0788 %
Kappa statistic                          0.6775
Mean absolute error                      0.1608
Root mean squared error                  0.401
Total Number of Instances             1219

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.857    0.180    0.838      0.857    0.848      0.678    0.838     0.793     0
                 0.820    0.143    0.840      0.820    0.830      0.678    0.838     0.775     1
Weighted Avg.    0.839    0.162    0.839      0.839    0.839      0.678    0.838     0.784

=== Confusion Matrix ===

   a   b   <-- classified as
 545  91 |   a = 0
 105 478 |   b = 1
```

Compare western conf vs all nba data prediction model

```
=== Summary ===

Correctly Classified Instances        1021               83.7572 %
Incorrectly Classified Instances       198               16.2428 %
Kappa statistic                          0.6747
Mean absolute error                      0.1624
Root mean squared error                  0.403
Relative absolute error                 32.4856 %
Root relative squared error             80.6048 %
Total Number of Instances             1219

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.835    0.160    0.827      0.835    0.831      0.675    0.837     0.769     0
                 0.840    0.165    0.848      0.840    0.844      0.675    0.837     0.795     1
Weighted Avg.    0.838    0.163    0.838      0.838    0.838      0.675    0.837     0.783

=== Confusion Matrix ===

   a   b   <-- classified as
 487  96 |   a = 0
 102 534 |   b = 1
```

Daniel Trevino

04/17/19

References

"2018-19 NBA Season Summary." *Basketball*, www.basketball-reference.com/leagues/NBA_2019.html

Johnwmillr. "Johnwmillr/SportradarAPIs." *GitHub*, https://github.com/johnwmillr/SportradarAPIs/blob/master/tests/test_NBA.py