Kahlil Cole

For my project I used a data set that had the career statistics for Michael Jordan and Kobe Bryant and Lebron James career from 2003 to 2018 because he is still an active player. I attempted to see who was statistically better between Michael Jordan Kobe Bryant, and Lebron James. In the data set I mainly focused on a few variables. Mainly points(PTS), field goal percentage (FG.), assists(AST), defensive rebounds(DRB), steals(STL), and blocks(BLK). I broke the data set into parts. Regular season stats for each player and playoff stats for each player.

head(LBJreg,10)

Season Age Tm Lg Pos G GS MP FG FGA FG. X3P X3PA X3P. X2P X2PA X2P. eFG.

1 2003-04 19 CLE NBA SG 79 79 3122 622 1492 0.417 63 217 0.290 559 1275 0.438 0.438

2 2004-05 20 CLE NBA SF 80 80 3388 795 1684 0.472 108 308 0.351 687 1376 0.499 0.504

3 2005-06 21 CLE NBA SF 79 79 3361 875 1823 0.480 127 379 0.335 748 1444 0.518 0.515

4 2006-07 22 CLE NBA SF 78 78 3190 772 1621 0.476 99 310 0.319 673 1311 0.513 0.507

5 2007-08 23 CLE NBA SF 75 74 3027 794 1642 0.484 113 359 0.315 681 1283 0.531 0.518

6 2008-09 24 CLE NBA SF 81 81 3054 789 1613 0.489 132 384 0.344 657 1229 0.535 0.530

7 2009-10 25 CLE NBA SF 76 76 2966 768 1528 0.503 129 387 0.333 639 1141 0.560 0.545

8 2010-11 26 MIA NBA SF 79 79 3063 758 1485 0.510 92 279 0.330 666 1206 0.552 0.541

9 2011-12 27 MIA NBA SF 62 62 2326 621 1169 0.531 54 149 0.362 567 1020 0.556 0.554

10 2012-13 28 MIA NBA PF 76 76 2877 765 1354 0.565 103 254 0.406 662 1100 0.602 0.603

FT FTA FT. ORB DRB TRB AST STL BLK TOV PF PTS Player RSorPO

1 347 460 0.754 99 333 432 465 130 58 273 149 1654 Lebron James Regular Season

2 477 636 0.750 111 477 588 577 177 52 262 146 2175 Lebron James Regular Season

3 601 814 0.738 75 481 556 521 123 66 260 181 2478 Lebron James Regular Season

4 489 701 0.698 83 443 526 470 125 55 250 171 2132 Lebron James Regular Season

5 549 771 0.712 133 459 592 539 138 81 255 165 2250 Lebron James Regular Season

6 594 762 0.780 106 507 613 587 137 93 241 139 2304 Lebron James Regular Season

7 593 773 0.767 71 483 554 651 125 77 261 119 2258 Lebron James Regular Season

8 503 663 0.759 80 510 590 554 124 50 284 163 2111 Lebron James Regular Season

9 387 502 0.771 94 398 492 387 115 50 213 96 1683 Lebron James Regular Season

10 403 535 0.753 97 513 610 551 129 67 226 110 2036 Lebron James Regular Season

After I separated the data set by player and regular season or playoff. I did a multiple linear regression for player.

> summary(mlrLBJreg)

Call:

lm(formula = PTS ~ FG. + AST + DRB + STL + BLK, data = LBJreg)

Residuals:

Min 1Q Median 3Q Max

-158.60 -118.75 -42.68 72.59 309.80

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 797.501 902.531 0.884 0.3976

FG. -1181.850 1884.627 -0.627 0.5446

AST -0.683 1.004 -0.680 0.5119

DRB 3.274 1.578 2.075 0.0648 .

STL 4.042 2.390 1.691 0.1217

BLK 3.359 3.465 0.969 0.3552

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 174.4 on 10 degrees of freedom

Multiple R-squared: 0.7244, Adjusted R-squared: 0.5865

F-statistic: 5.256 on 5 and 10 DF, p-value: 0.01263

> summary(mlrMJreg)

Call:

lm(formula = PTS ~ FG. + AST + DRB + STL + BLK, data = MJreg)

Residuals:

Min 1Q Median 3Q Max

-446.03 -116.51 -16.77 97.98 453.91

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 501.612 1523.366 0.329 0.7495

FG. -981.199 3650.704 -0.269 0.7942

AST -2.041 1.500 -1.361 0.2065

DRB 3.025 1.455 2.079 0.0673 .

STL 9.884 3.927 2.517 0.0329 \*

BLK 3.880 5.121 0.758 0.4681

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 265 on 9 degrees of freedom

Multiple R-squared: 0.9323, Adjusted R-squared: 0.8947

F-statistic: 24.79 on 5 and 9 DF, p-value: 5.187e-05

> summary(mlrKBreg)

Call:

lm(formula = PTS ~ FG. + AST + DRB + STL + BLK, data = KBreg)

Residuals:

Min 1Q Median 3Q Max

-264.59 -152.22 -62.48 129.78 507.42

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -612.2519 992.7134 -0.617 0.5473

FG. 1618.3100 2508.5626 0.645 0.5293

AST 0.9078 1.7134 0.530 0.6045

DRB 3.2370 2.5538 1.268 0.2256

STL 6.5274 4.0847 1.598 0.1324

BLK -7.4552 4.2068 -1.772 0.0981 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 242.5 on 14 degrees of freedom

Multiple R-squared: 0.9133, Adjusted R-squared: 0.8823

F-statistic: 29.49 on 5 and 14 DF, p-value: 5.827e-07

I do not think mlr set up is a good indicator for sports stats or it just might be this data set. My response variable was points (PTS) and my indicator variables were field goal percentage (FG.), assists(AST), defensive rebounds(DRB), steals(STL), and blocks(BLK). The only indicator that was significant was steals (STL) for Michael Jordan. That indicator variable had a pvalue of 0.0329. That was the only significant p-value from regular season stats between the players. When I did mlr to the playoff stats I revived some significant variables. My Michael Jordan playoff mlr had no significant variables. The only two that where close to being significant were defensive rebounds and steals.

DRB 4.439 2.133 2.081 0.076 .

STL 13.121 5.972 2.197 0.064 .

My Kobe Bryant playoff mlr had no significant values at all. My LeBron playoff mlr had two significant variables field goal percentage and defensive rebounds.

FG. 828.3529 273.5668 3.028 0.0192 \*

AST 0.8468 0.7304 1.159 0.2843

DRB 2.9852 0.9415 3.171 0.0157 \*

Next, I did some k-fold cross validation to my data set. For the k-fold cross validation and the regression trees I created a new dataframe. In this data frame I had to make all of the stats the same length. Since Michael Jordan on played 15 seasons and Kobe played 20 season and Lebron is on season 17 I made the data frame only for 15 seasons.

> qda.error

[1] 0.2444444

> lda.error

[1] 0.2666667

> lda.fit

Call:

lda(Player ~ PTS + FG. + AST + DRB + STL + BLK, data = reg\_train)

Prior probabilities of groups:

Kobe Bryant Lebron James Michael Jordan

0.3055556 0.3055556 0.3888889

Group means:

PTS FG. AST DRB STL BLK

Kobe Bryant 1783.636 0.4504545 330.1818 290.2727 105.9091 39.00000

Lebron James 2060.091 0.5084545 555.7273 472.5455 123.9091 55.72727

Michael Jordan 2128.643 0.4877857 377.2143 329.2143 166.7143 60.78571

Coefficients of linear discriminants:

LD1 LD2

PTS 0.0001811236 0.002279947

FG. 0.7571695113 -15.071933196

AST -0.0111898826 0.005232203

DRB -0.0015869095 -0.009054509

STL 0.0310778352 -0.025250883

BLK -0.0196421394 -0.009803509

Proportion of trace:

LD1 LD2

0.8103 0.1897

> qda.fit

Call:

qda(Player ~ PTS + FG. + AST + DRB + STL + BLK, data = reg\_train)

Prior probabilities of groups:

Kobe Bryant Lebron James Michael Jordan

0.3055556 0.3055556 0.3888889

Group means:

PTS FG. AST DRB STL BLK

Kobe Bryant 1783.636 0.4504545 330.1818 290.2727 105.9091 39.00000

Lebron James 2060.091 0.5084545 555.7273 472.5455 123.9091 55.72727

Michael Jordan 2128.643 0.4877857 377.2143 329.2143 166.7143 60.78571

Based on the means from the qda and lda fit we can say that Michael Jordan was the better scorer and defender because he has the highest points, steals, and block average. Lebron has the better field goal percentage, assist, and defensive rebound average. Kobe had the lowest average of the three in all categories.

Lastly, I did a tree-based method on my data set. The variable used in the tree were free throw (FT), field goal (FG), and free throw percentage (FT.). The first mean square error before I pruned my tree was 126081.8.

> mean((yhat - test.PTS)^2)

[1] 126081.8

After I pruned my tree my mean square error was 131512.6

> mean((yhat - test.PTS)^2)

[1] 131512.6

My original regression tree is more precise because it has the lowest mse.

