Analyzing NBA Player Salaries: A Statistical Modeling Approach

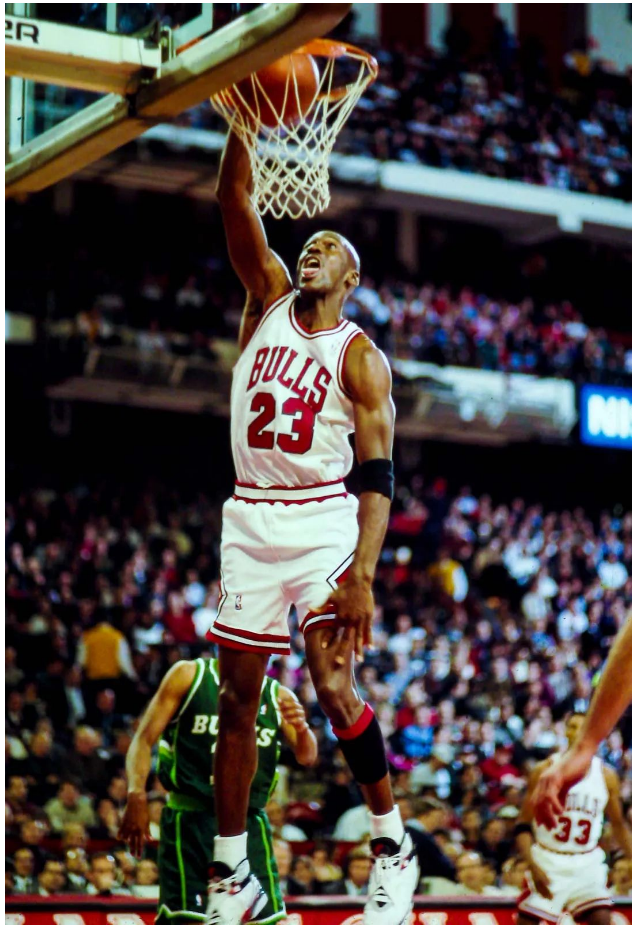
Team 8: Colin Dowd (ccd4xc), Casey Pridgen (ccp2uz), Xin He (ubn4am), Sami Adam (zeh4tv)

Question of Interest

Quantitative Question:

· *What statistics lead to NBA players receiving higher salaries?*

- We are fans of Basketball
- We aim to provide valuable information to players and team managers
- Our question serves multiple parties, such as teams and players themselves



University of Virginia

Data Set Content and Source

- Our data consists of individual player statistics from the 2022-2023 NBA season.(*)
- This data was found through Kaggle
- Figure 1 shows a small part of our data set.

*(Including counting statistics as well as advanced statistics)

B	C	D	E	F	G	H	
Player Name	Salary	Position	Age	Team	GP	GS	MP
Stephen Curry	48070014	PG	34	GSW	56	56	
John Wall	47345760	PG	32	LAC	34	3	
Russell Westbro	47080179	PG	34	LAL/LAC	73	24	
LeBron James	44474988	PF	38	LAL	55	54	
Kevin Durant	44119845	PF	34	BRK/PHO	47	47	
Bradley Beal	43279250	SG	29	WAS	50	50	
Kawhi Leonard	42492492	SF	31	LAC	52	50	
Paul George	42492492	SF	32	LAC	56	56	
Giannis Antetokc	42492492	PF	28	MIL	63	63	
Damian Lillard	42492492	PG	32	POR	58	58	
Klay Thompson	40600080	SF	32	GSW	69	69	
Kyrie Irving	38917057	PG-SG	30	BRK/DAL	60	60	
Rudy Gobert	38172414	C	30	MIN	70	70	
Khris Middleton	37984276	SF	31	MIL	33	19	
Anthony Davis	37980720	C	29	LAL	56	54	
Jimmy Butler	37653300	PF	33	MIA	64	64	
Tobias Harris	37633050	SF	30	PHI	74	74	
Kemba Walker	37281261	PG	32	DAL	9	1	
Trae Young	37096500	PG	24	ATL	73	73	
Zach LaVine	37096500	SG	27	CHI	77	77	
Ben Simmons	35448672	PG	26	BRK	42	33	
Pascal Siakam	35448672	PF	28	TOR	71	71	
Myles Turner	35096500	C	26	IND	62	62	
Jrue Holiday	34319520	PG	32	MIL	67	65	
Karl-Anthony Tov	33833400	PF	27	MIN	29	29	
Devin Booker	33833400	SG	26	PHO	53	53	
Andrew Wiggins	33616770	SF	27	GSW	37	37	

Figure 1 (part)

Important Variables

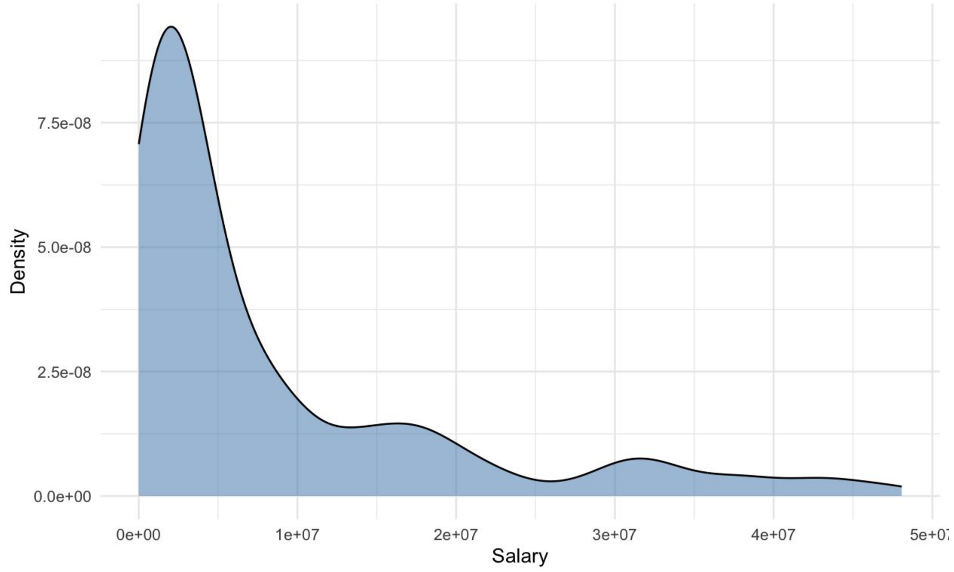
- **Ordinary Least Squares model, Lasso Regression and Random Forest Response Variable:**

  log(Salary) - The log transformation of the salary a player is paid.*

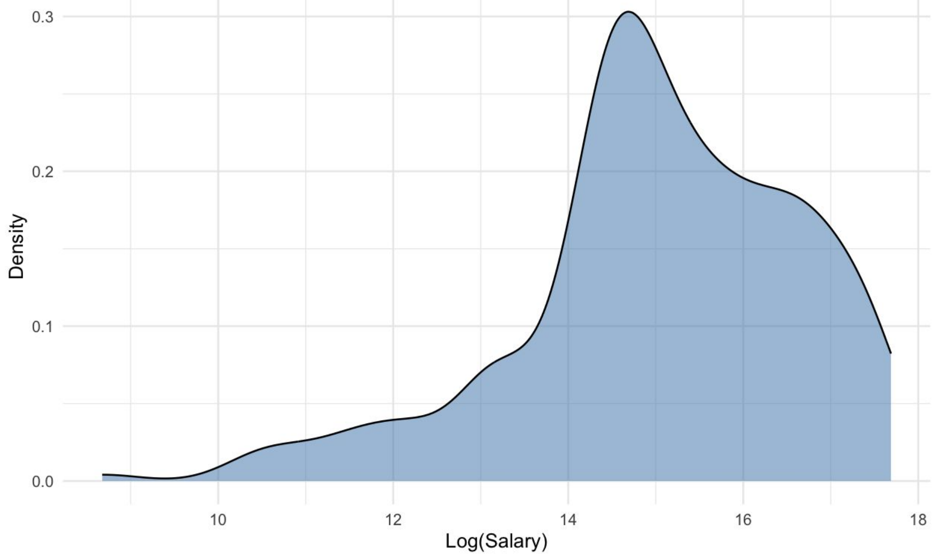
*Since salaries are normally skewed right, we decided to take the logarithm of the variable

Variable Transformations

Raw Salary Distribution



Transformed Salary Distribution



University of Virginia

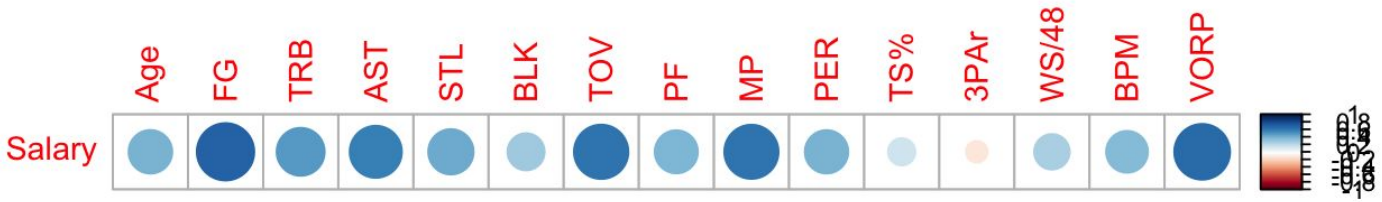
Important Variables

- **Predictor Variables:**
  - _Age_ (Age of NBA player)
  - _FG_ (Field goals made)
  - _TRB_ (Total rebounds per game)
  - _AST_ (Assists per game)
  - _STL_ (Steals per game)
  - _BLK_ (Blocks per game)
  - _TOV_ (Turnovers per Game)
  - _PF_ (Personal Fouls)
  - _MP_ (Minutes Played)
  - _PER_ (Player Efficiency Rating)
  - _TS%_ (True Shooting Percentage)
  - _3PAr_ (Three point attempt rate)
  - _WS/48_ (Win Shares per 48 minutes)
  - _BPM_ (Box Plus Minus)
  - _VORP_ (Value Over Replacement Player)
  - _Position_Group_* (overall positions a player plays)
  - _Starter_* (a player started more than half the games they played)

*created using existing _variables_.

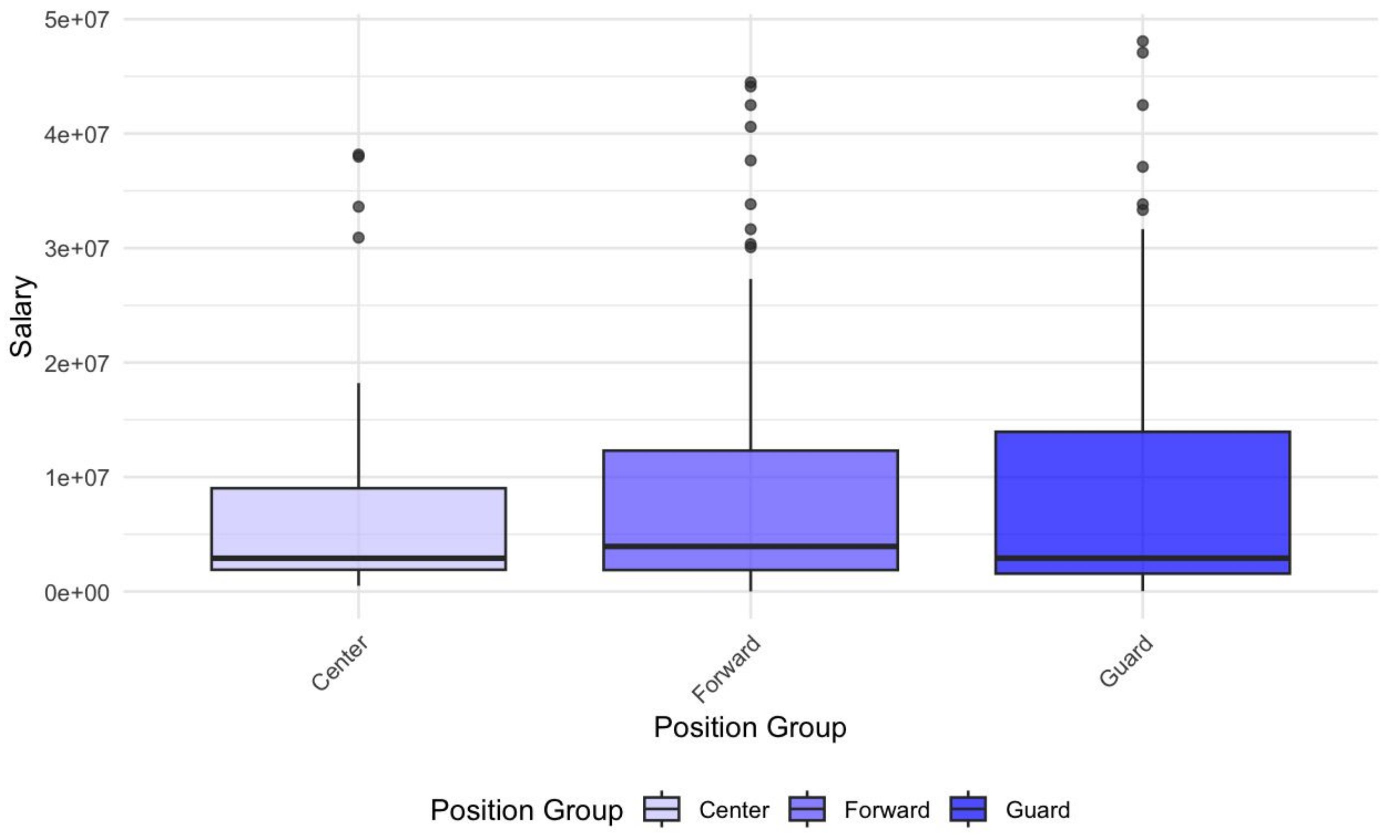
UNIVERSITY of VIRGINIA

Predictor Correlations to Response Variable



All the predictors, except 3PAr, are positively correlated to Salary

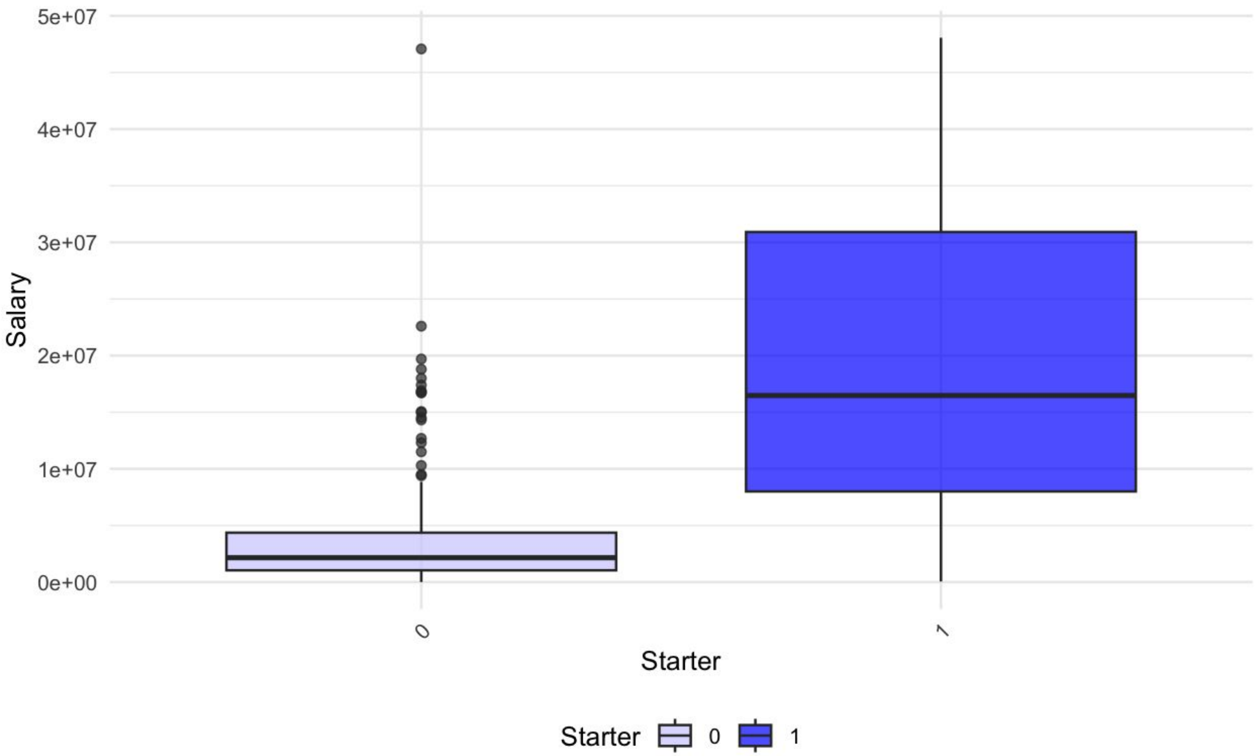
Salary Distribution Among Position Group



Player salaries are relatively consistent among player positions

There does not appear to be a strong relationship between between the two variables

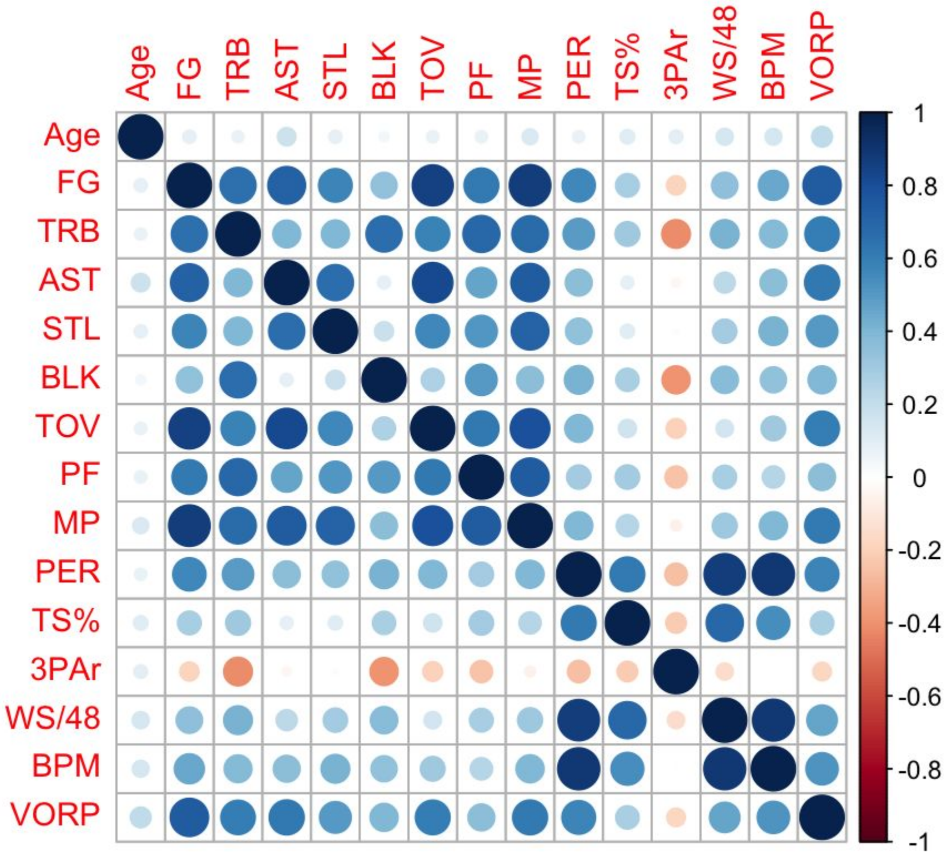
Salary Distribution Among Starters/Non Starters



There is a significant difference between Starter and Non-starter salaries

We expected this, as starters tend to be better players, meaning that they have a bigger impact on the game

Predictor Correlation



There are high levels of multicollinearity among predictors

The data would benefit from variable selection or method with increased bias

Ordinary Least Squares Results

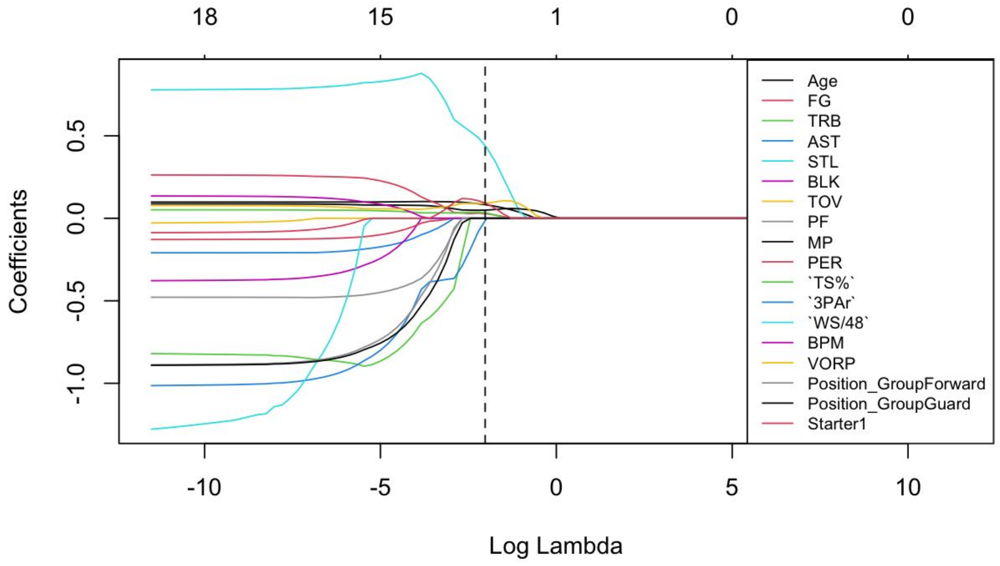
```
Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         14.24189    1.00588  14.159  < 2e-16 ***
x_train_shrinkAge                    0.09811    0.01770   5.544 8.65e-08 ***
x_train_shrinkFG                     0.26400    0.12935   2.041 0.042483 *  
x_train_shrinkTRB                    0.05176    0.06971   0.743 0.458593    
x_train_shrinkAST                   -0.20829    0.10451  -1.993 0.047531 *  
x_train_shrinkSTL                    0.77541    0.36759   2.109 0.036069 *  
x_train_shrinkBLK                   -0.38018    0.30399  -1.251 0.212429    
x_train_shrinkTOV                   -0.03342    0.30456  -0.110 0.912729    
x_train_shrinkPF                    -0.47707    0.18311  -2.605 0.009820 ** 
x_train_shrinkMP                     0.08724    0.03364   2.593 0.010167 *  
x_train_shrinkPER                   -0.12936    0.04256  -3.039 0.002666 ** 
x_train_shrink`TS%`                 -0.80976    0.97460  -0.831 0.406978    
x_train_shrink`3PAr`                -1.01893    0.48192  -2.114 0.035646 *  
x_train_shrink`WS/48`               -1.34192    2.97493  -0.451 0.652391    
x_train_shrinkBPM                    0.13686    0.05606   2.442 0.015438 *  
x_train_shrinkVORP                   0.07779    0.12714   0.612 0.541296    
x_train_shrinkPosition_GroupForward -0.89014    0.25556  -3.483 0.000601 ***
x_train_shrinkPosition_GroupGuard   -0.89317    0.29923  -2.985 0.003167 ** 
x_train_shrinkStarter1              -0.08774    0.26952  -0.326 0.745079    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.082 on 214 degrees of freedom
Multiple R-squared:  0.5918,	Adjusted R-squared:  0.5574 
F-statistic: 17.23 on 18 and 214 DF,  p-value: < 2.2e-16
```

The model explains ~56% of the variability in the log transformation of Salary and has a TMSE of 1.22

There are many insignificant predictors

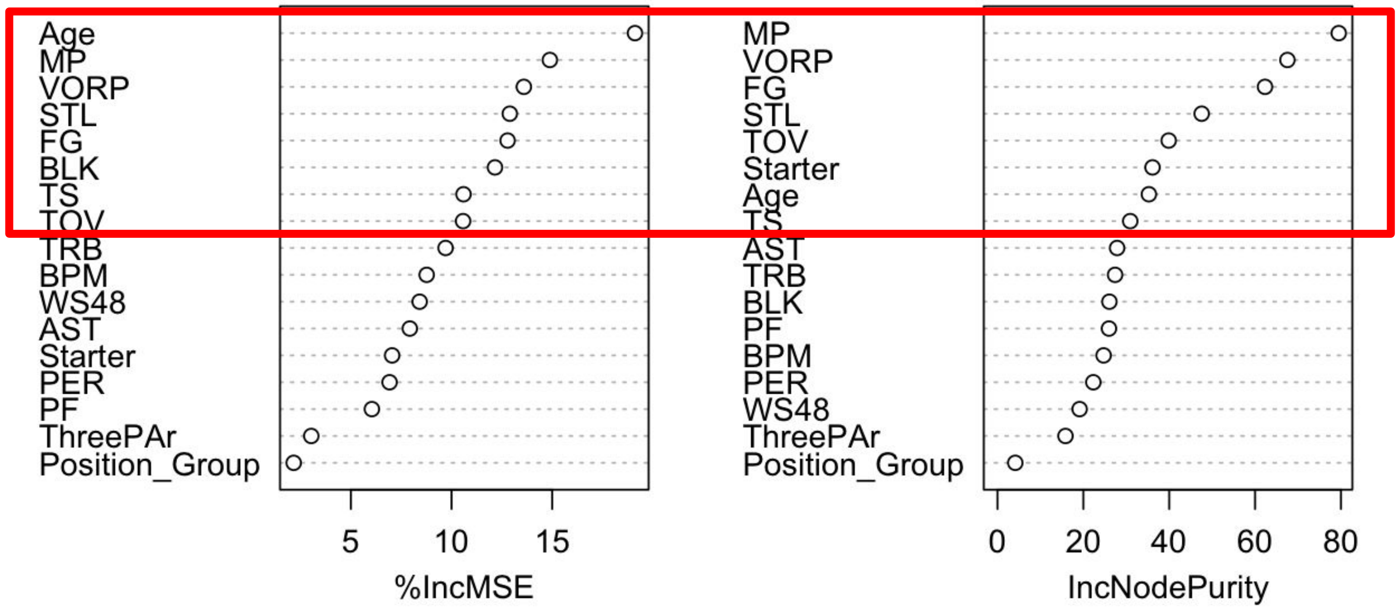
Lasso Shrinkage Method Results

(Intercept)	11.37384208
Age	0.08235278
FG	0.02938858
TRB	0.03400768
AST	.
STL	0.44172025
BLK	.
TOV	.
PF	.
MP	0.04915239
PER	.
`TS%`	.
`3PAr`	.
`WS/48`	.
BPM	.
VORP	0.09074432
Position_GroupForward	.
Position_GroupGuard	.
Starter1	0.09283299



Lasso Regression is able to set coefficient estimates to 0, effectively performing variable selection and has a TMSE of 1.21

Random Forests Results



Random Forests allow us to reduce variance by taking the average predicted response of many trees. This allowed us to get a TMSE of 0.95

Comparison

Model <chr>	Test_MSE <dbl>
OLS	1.22
Lasso	1.21
Random Forest	0.95

Random Forests performed the best and is the most applicable for our question, as the goal is model interpretability