NBA Basketball Analysis

By: Eoin Gawronski

Table of Contents

1. Introduction/Objective
2. Dataset/Features
3. Correlation Matrix
4. Models: Linear Regression,Elastic Net, Lasso, or Random Forest
5. Feature Importance
6. Best Model
7. Conclusions

Introduction

Objective

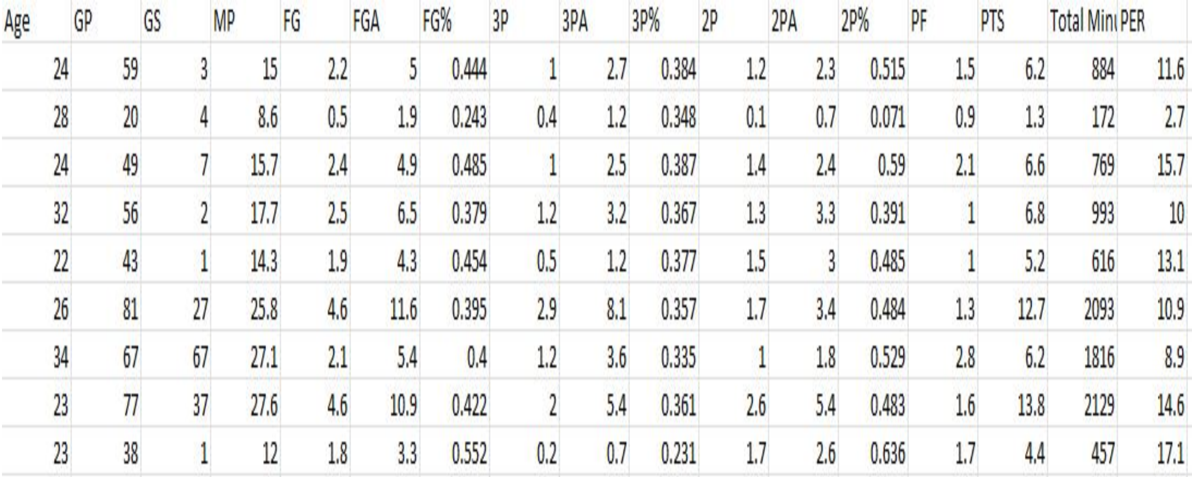
- Provide a deeper understanding of what drives scoring in basketball through models and data analysis

Goal

- Understand and identify which specific game statistics are highly correlated with Points

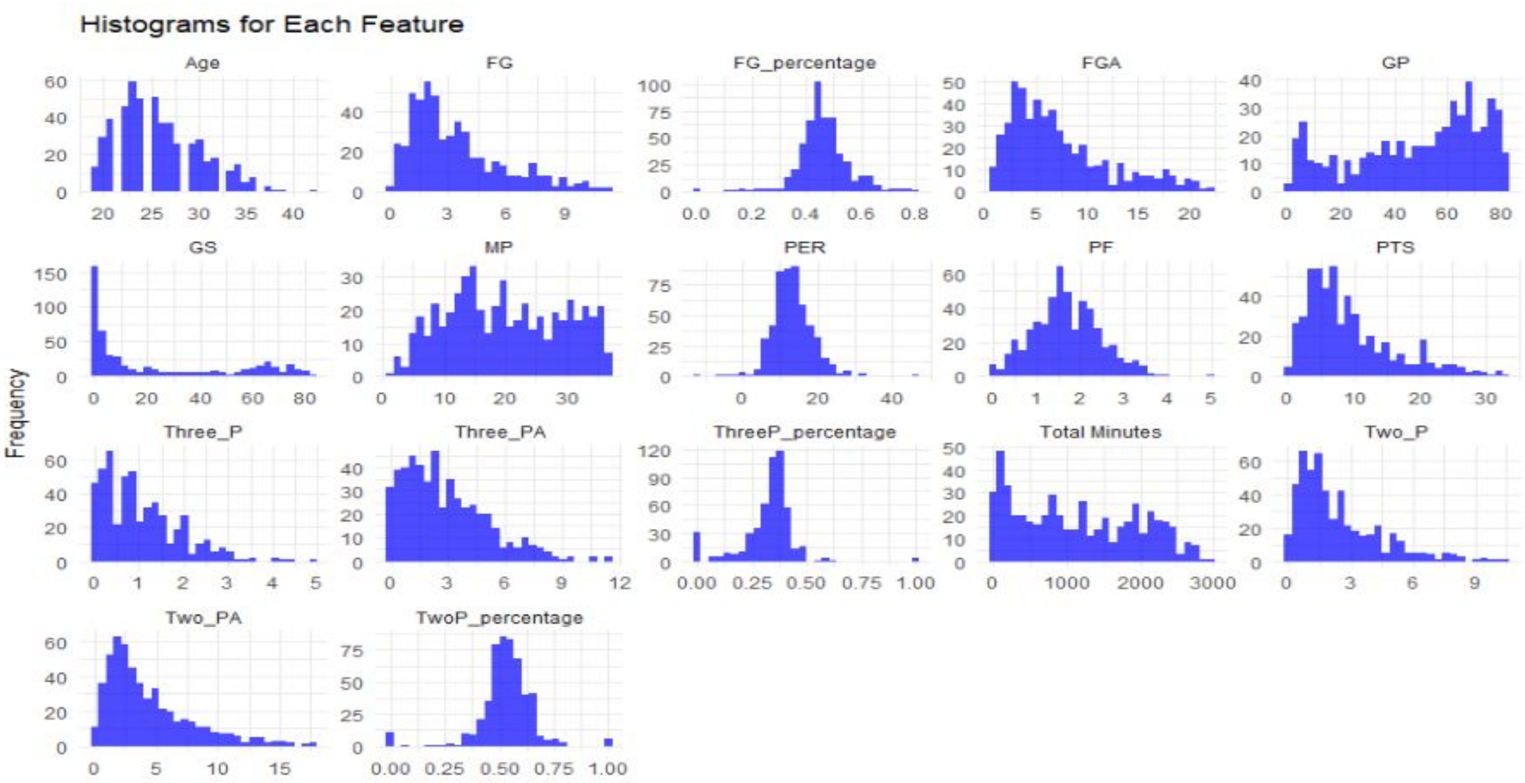
Dataset/Features

- NBA dataset containing per game statistics for the 2022-2023 season
- Contains 540 observations and 17 features
  -MP
  -FG
  -FGA
  -2P
  -2PA
  -GS
  -PER
- Regression problem
- All features are continuous



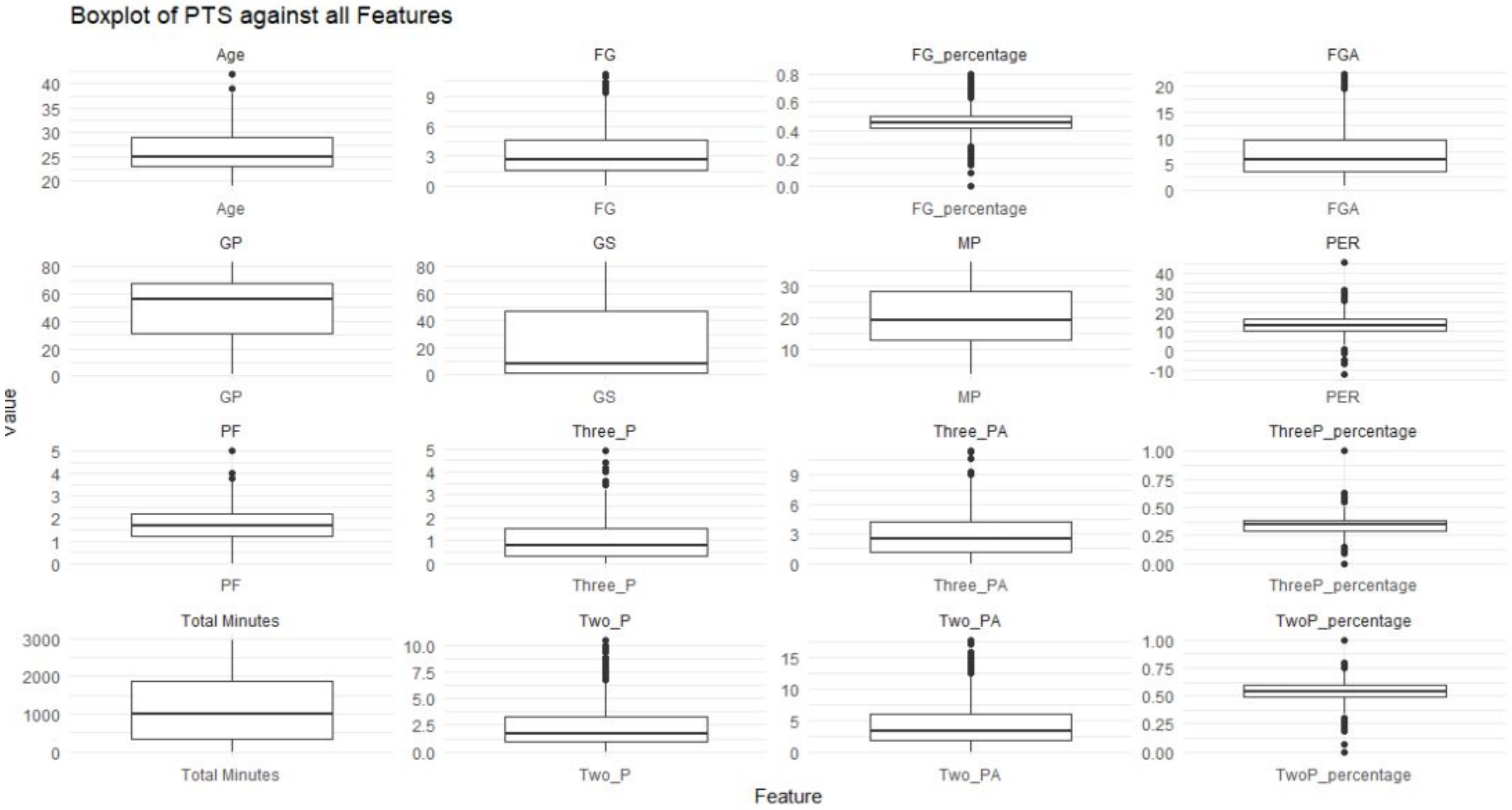
Age	GP	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	PF	PTS	Total Min	PER
24	59	3	15	2.2	5	0.444	1	2.7	0.384	1.2	2.3	0.515	1.5	6.2	884	11.6
28	20	4	8.6	0.5	1.9	0.243	0.4	1.2	0.348	0.1	0.7	0.071	0.9	1.3	172	2.7
24	49	7	15.7	2.4	4.9	0.485	1	2.5	0.387	1.4	2.4	0.59	2.1	6.6	769	15.7
32	56	2	17.7	2.5	6.5	0.379	1.2	3.2	0.367	1.3	3.3	0.391	1	6.8	993	10
22	43	1	14.3	1.9	4.3	0.454	0.5	1.2	0.377	1.5	3	0.485	1	5.2	616	13.1
26	81	27	25.8	4.6	11.6	0.395	2.9	8.1	0.357	1.7	3.4	0.484	1.3	12.7	2093	10.9
34	67	67	27.1	2.1	5.4	0.4	1.2	3.6	0.335	1	1.8	0.529	2.8	6.2	1816	8.9
23	77	37	27.6	4.6	10.9	0.422	2	5.4	0.361	2.6	5.4	0.483	1.6	13.8	2129	14.6
23	38	1	12	1.8	3.3	0.552	0.2	0.7	0.231	1.7	2.6	0.636	1.7	4.4	457	17.1

Histograms



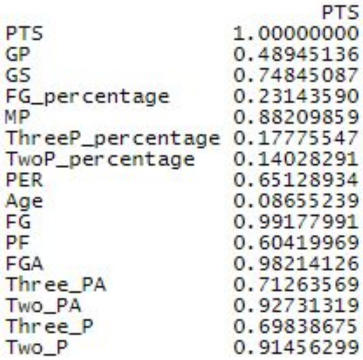
Histograms for Each Feature
Frequency
Age
FG
FG_percentage
FGA
GP
GS
MP
PER
PF
PTS
Three_P
Three_PA
ThreeP_percentage
Total Minutes
Two_P
Two_PA
TwoP_percentage

Box Plots

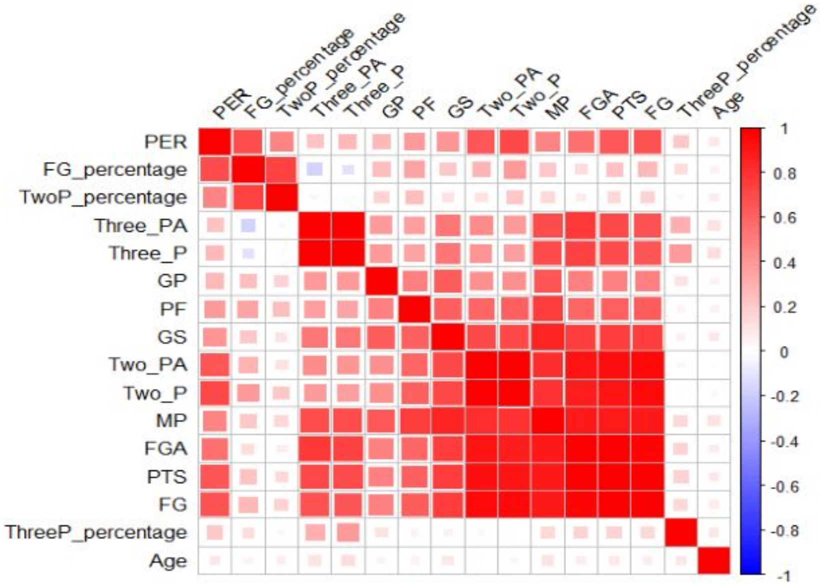

Boxplot of PTS against all Features
Age
40
35
30
25
20
Age
FG
9
6
3
0
FG
FG_percentage
0.8
0.6
0.4
0.2
0.0
FG_percentage
FGA
20
15
10
5
0
FGA
GP
80
60
40
20
0
GP
GS
80
60
40
20
0
GS
MP
30
20
10
MP
PER
40
30
20
10
0
-10
PER
PF
5
4
3
2
1
0
PF
Three_P
5
4
3
2
1
0
Three_P
Three_PA
9
6
3
0
Three_PA
ThreeP_percentage
1.00
0.75
0.50
0.25
0.00
ThreeP_percentage
Total Minutes
3000
2000
1000
0
Total Minutes
Two_P
10.0
7.5
5.0
2.5
0.0
Two_P
Two_PA
15
10
5
0
Two_PA
TwoP_percentage
1.00
0.75
0.50
0.25
0.00
TwoP_percentage
value
Feature

Correlation Analysis

- Correlation matrix heatmap gives an overview of every feature and how they correlate with each other
- Points and its dependents
  -Points & GS = .74
  -Points & MP = .88
  -Points & FG = .99
  -Points & FGA = .98
  -Points & Two_PA = .93
  -Points & Two_P = .91



```
                          PTS
PTS              1.00000000
GP               0.48945136
GS               0.74845087
FG_percentage    0.23143590
MP               0.88209859
ThreeP_percentage 0.17775547
TwoP_percentage  0.14028291
PER              0.65128934
Age              0.08655239
FG               0.99177991
PF               0.60419969
FGA              0.98214126
Three_PA         0.71263569
Two_PA           0.92731319
Three_P          0.69838675
Two_P            0.91456299
```

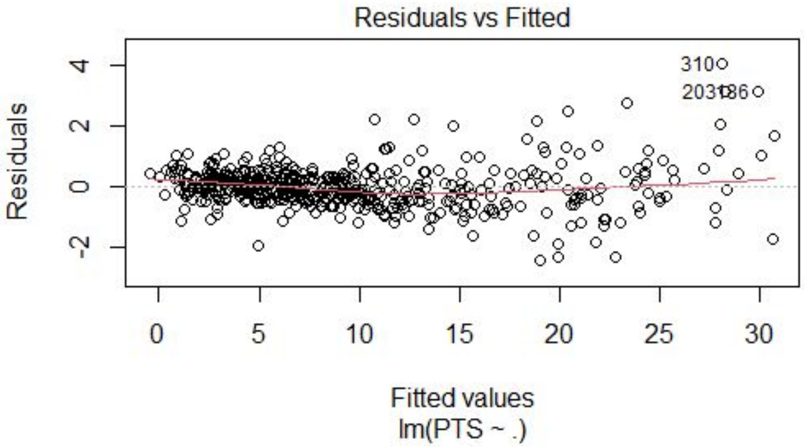


Linear Regression

- Residuals vs Fitted plot to look at linear regression
- X-Axis(Fitted Values): Represents the predicted values of the dependent variable, which in this case is PTS.
- Y-axis(Residuals): the differences between the observed values and the fitted values from the model.
- Key aspects of the plot

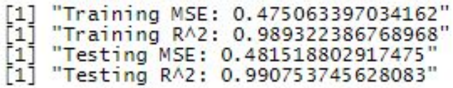
  -Random scatter

  -Outliers

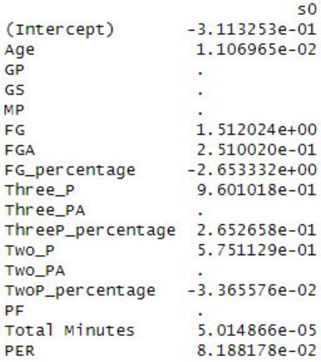


Elastic Net Regression

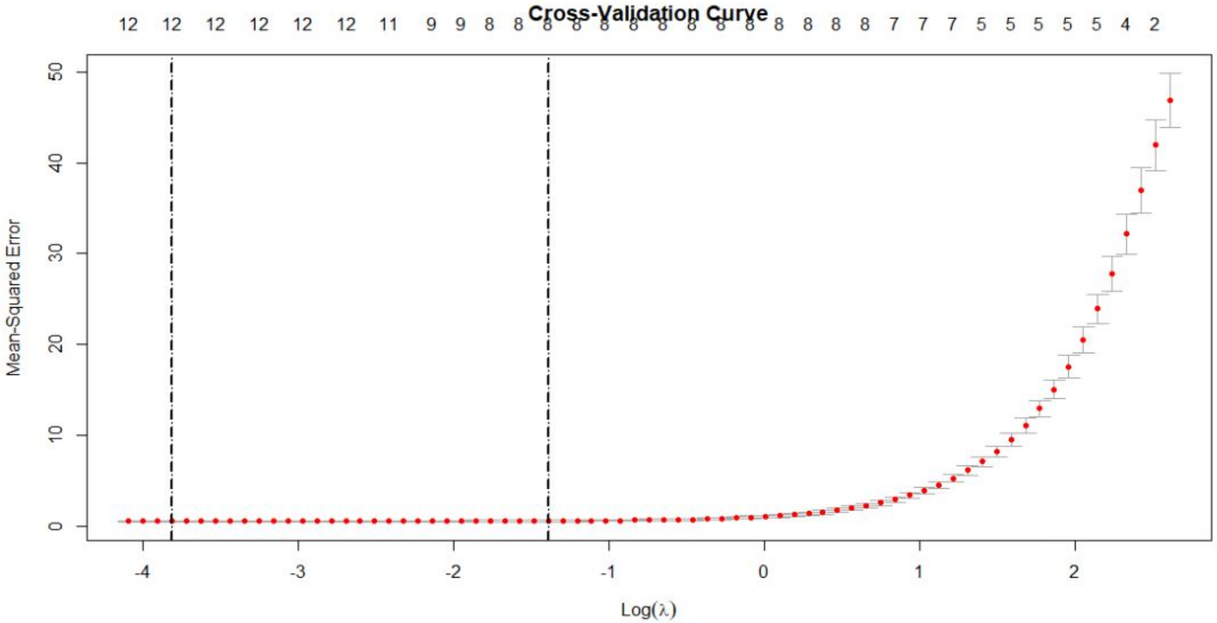
- Optimal lambda is just to the left of -1
- The increase to the right of this point shows potential overfitting
- All the different variables shows whether the coefficient is positive or negative
- MSE training= .48
- MSE testing= .48



```
[1] "Training MSE: 0.475063397034162"
[1] "Training R^2: 0.989322386768968"
[1] "Testing MSE: 0.481518802917475"
[1] "Testing R^2: 0.990753745628083"
```

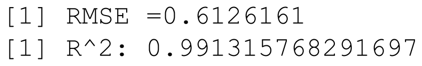


```
                             s0
(Intercept)       -3.113253e-01
Age                1.106965e-02
GP                 .
GS                 .
MP                 .
FG                 1.512024e+00
FGA                2.510020e-01
FG_percentage     -2.653332e+00
Three_P            9.601018e-01
Three_PA           .
ThreeP_percentage  2.652658e-01
Two_P              5.751129e-01
Two_PA             .
TwoP_percentage   -3.365576e-02
PF                 .
Total Minutes      5.014866e-05
PER                8.188178e-02
```

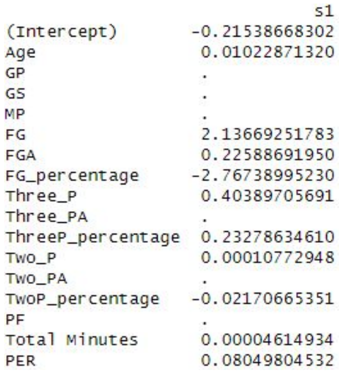


Lasso Regression

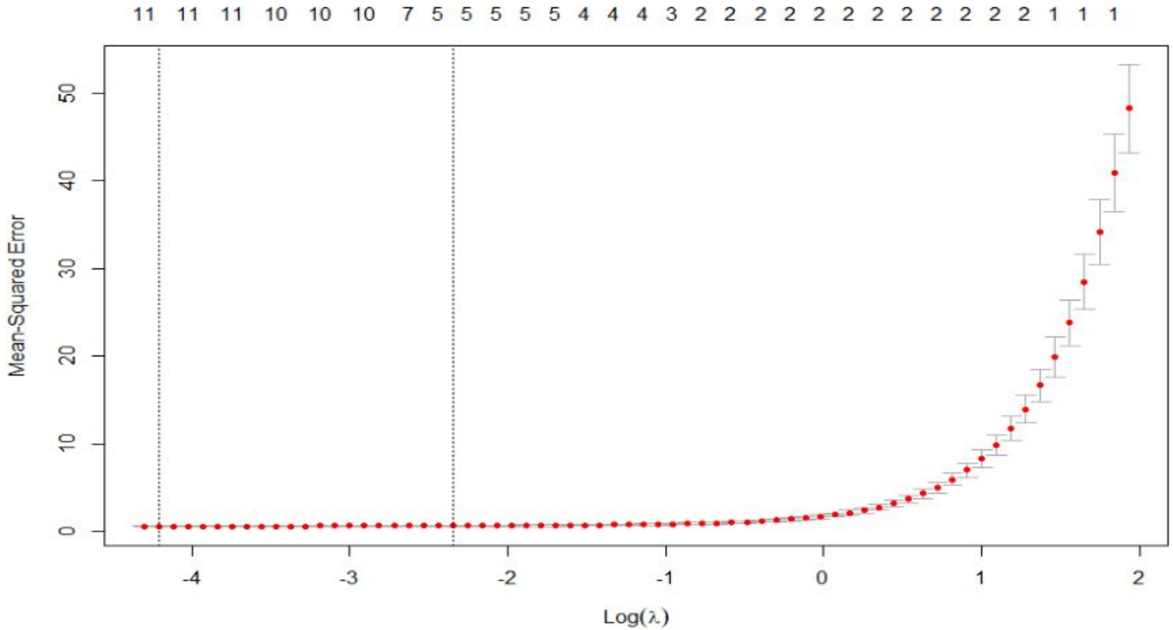
- Optimal lambda to the left of -2
- MSE training=.47
- MSE testing= .50
- High R-squared at .99, similar to the Elastic net model



```
[1] RMSE =0.6126161
[1] R^2: 0.991315768291697
```



```
                           s1
(Intercept)     -0.21538668302
Age              0.01022871320
GP               .
GS               .
MP               .
FG               2.13669251783
FGA              0.22588691950
FG_percentage   -2.76738995230
Three_P          0.40389705691
Three_PA         .
ThreeP_percentage  0.23278634610
Two_P            0.00010772948
Two_PA           .
TwoP_percentage -0.02170665351
PF               .
Total Minutes    0.00004614934
PER              0.08049804532
```

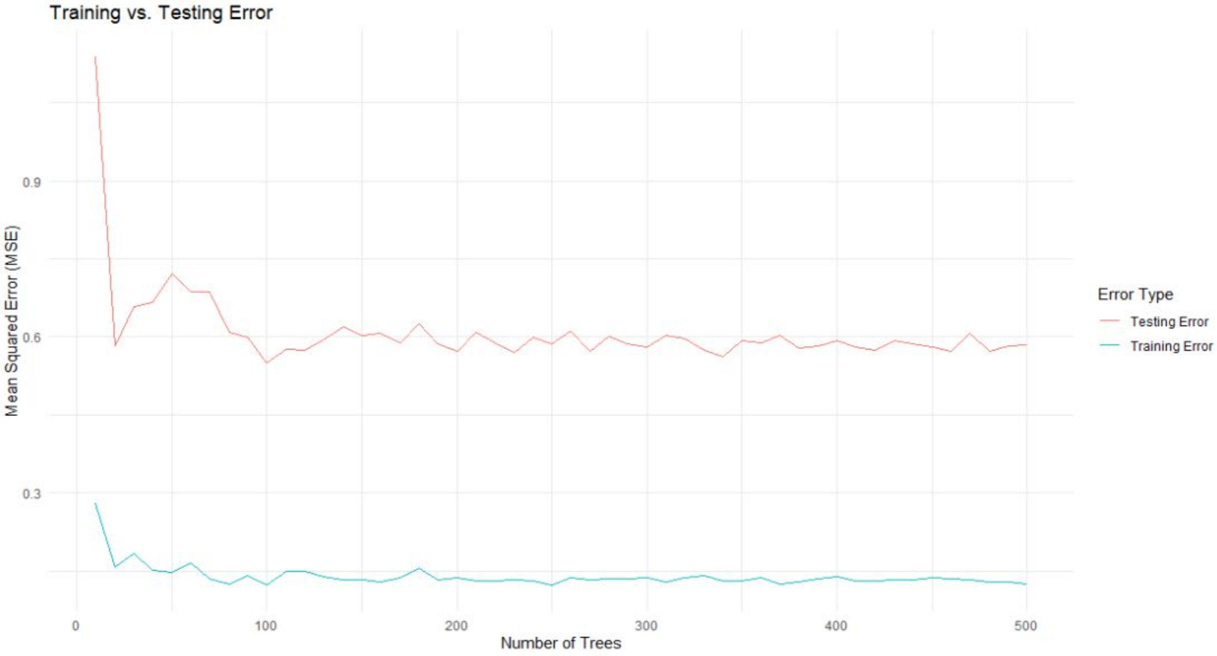


Random Forest

- Shows a training vs testing model of random forest
- Both errors decrease as more trees are added, but they level out, which shows that adding more does not significantly improve the model's performance
- MSE Training/Testing:

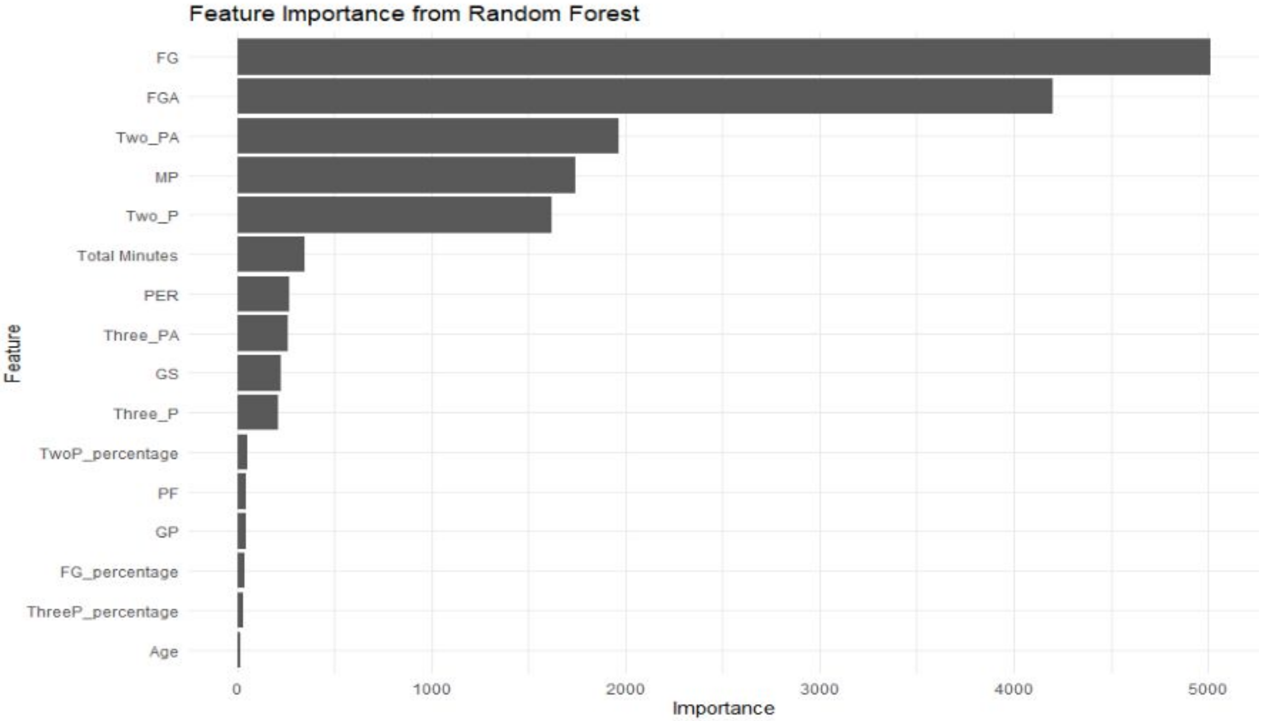
  -Training: .13

  -Testing: .59



Feature Importance

- Show the feature importance from a Random forest model
- Model shows highest importance values starting at FG at the top and goes in descending order to Age
- Narrows down the key features that impact PTS



Model Selection

- Elastic net regression showed to be a good fit for my analysis
- Models shows relatively low MSE at .48 for both the training and testing data
- Both the R squared for training and testing were high, .98 and .99

Conclusions

- The best model for analyzing the dataset was the Elastic Net regression

- The Feature importance suggested there may be more then one variable that affects PTS per game
  - FG, FGA, Two_PA, and MP

- Deeper analysis could be done on a variety of different models to see relationships between different variables and its impact on PTS.