Estimating Salaries Of NBA Players Using Predictive Modeling Techniques

16th August, 2019

Aishwarya Pawar
Eric Naumann
Brandon Whiteley
Yeong-in Jang
Haritha Maheshkumar

The University of Texas at Austin, McCombs School of Business - MSBA Class of 2020

Agenda

- **Problem statement** *(What are we doing?)*
- **Exploratory Data Analysis** *(Why are we doing it?)*
- **Methodology & Results summary** *(How are we doing it?)*
- **Findings and Insights** *(What did we find?)*
- **Next steps and ideas** *(What can be done better and how?)*

What metrics can be used to predict NBA player salaries?

Key questions: Out of the many assessment metrics, which have been significant in estimating a player's salary? How accurate is the estimation? How are they different from rest of the metrics?

Available information: Dataset containing salary and performance information of 419 NBA players, spanned over 48 variables

Sample variable types / Data snapshot:

- ***Per Game Metrics***: Minutes, Field Goals, 3 pointers, Free Throws, Rebounds etc..
- ***Percentage Metrics:*** FT%, 3PT% etc..
- ***Advanced Metrics***: PER, True Shooting %, TO%, Usage %, Win Shares, Plus Minus, etc..

Cleaning the data and observing the relationships across variables

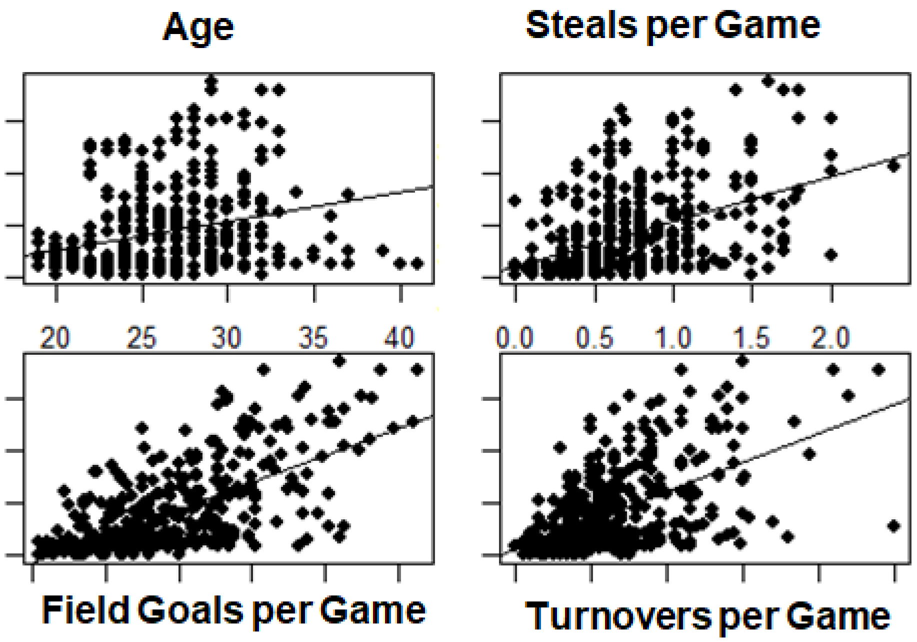
Data cleaning:

- Predictors from 2017 are used to predict salary information in 2018
- 0.09% of players were removed due to missing information in a few metrics
- Averaged stats for players who played for multiple teams

Correlations:

- Correlated variables were removed in some regressions (Corr<Abs(0.5))
  - Ex: 2 PT FGs , FGA per game, & 2 PT FGA
  - Effective Field Goal Percentage ~ True Shooting Percentage ~ Field Goal Percentage

Age, Steals, Field Goals, Turnovers (per game) metrics show a promising relationship with the salaries



Other potential significant variables :

Minutes played per Game
Win Shares
2 Points Field Per Points
Value over replacement player

We tried estimating the salaries of players using the following predictive modeling techniques

Modeling technique	Refinements tried	Test RMSE obtained	Variance Explained
Linear	1. Introduced interaction terms 2. Tested for non-linear relationships	USD 5.14M	58.11%
Ridge	Best lambda selection based on RMSE curve	USD 5.48M	59.23%
Lasso	Best lambda selection based on RMSE curve	USD 5.45M	59.81%
Random Forest	1. Experimentation with mtry (best result at mtry=23) 2. Variable elimination using importance() function	USD 5.46M	58.99%
Bagging	1. 31 variables selected based on importance() function	USD 5.24M	54.31%

Linear Regression with Variable Selections

All Variables OLS
# Significant Variables= 14
RMSE = 5.34 M

Forward Selection OLS
# Significant Variables= 5
RMSE = 5.16 M

Backward Selection OLS
# Significant Variables= 7
RMSE = 5.14 M

Salary = - 7,862,915 + 429,094×Age - 48,917×Games + 66,699×Games started + 547,183×Field goal attempts per game + 1,041,936×Free throw attempts per game - 1,429,831×Turnovers per game + 1,580,038×*Value over replacement player*

Linear Regression with Non-linearity and Interactions

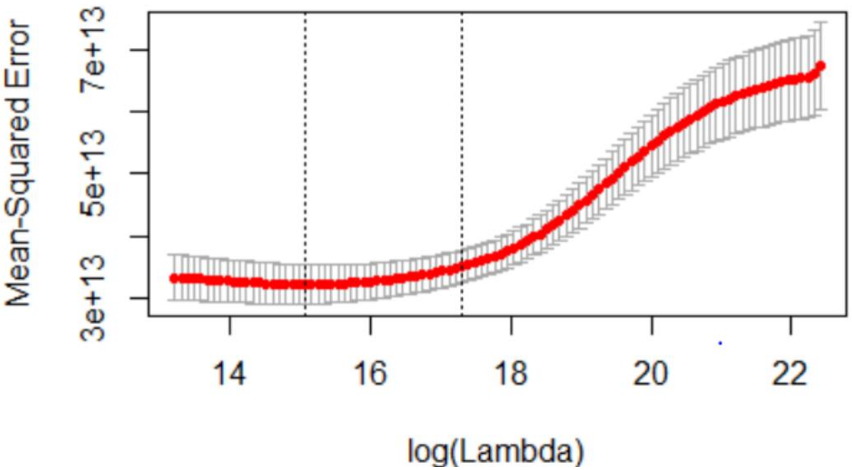
Non- Linearity

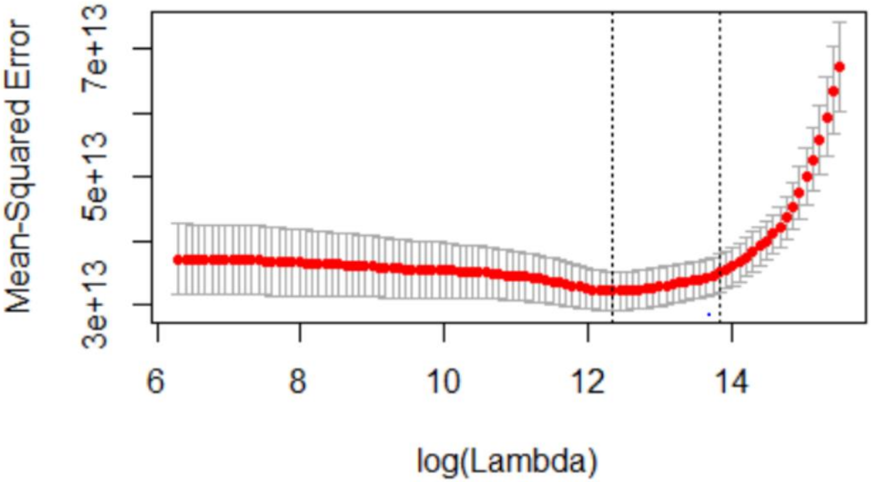
- Non- Linearity was observed for win share (2 degree) and box plus (2 and 3 degree
- RMSE post introducing non-linearity = 5.17 Mn

Interactions

- Interaction between was observed
- RMSE post introducing interaction = 5.17 Mn

Ridge and Lasso





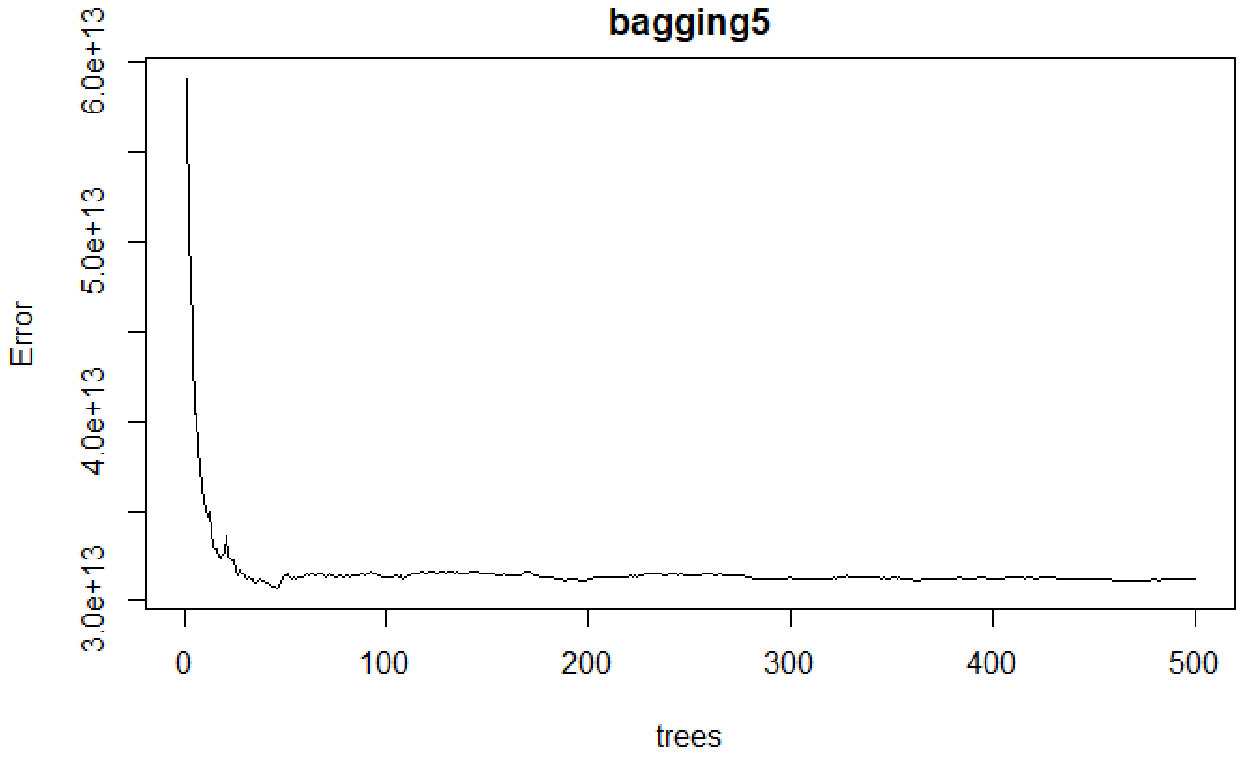
With minimum Lambda, we obtained RMSE of 5.48 Mn

With minimum Lambda, we obtained RMSE of 5.45 Mn

We tried estimating the salaries of players using the following predictive modeling techniques

Modeling technique	Refinements tried	Test RMSE obtained	Variance Explained
Linear	1. Introduced interaction terms 2. Tested for non-linear relationships	USD 5.14M	58.11%
Ridge	Best lambda selection based on RMSE curve	USD 5.48M	59.23%
Lasso	Best lambda selection based on RMSE curve	USD 5.45M	59.81%
Random Forest	1. Experimentation with mtry (best result at mtry=23) 2. Variable elimination using importance() function	USD 5.46M	58.99%
Bagging	1. 31 variables selected based on importance() function	USD 5.24M	54.31%

Bagging :



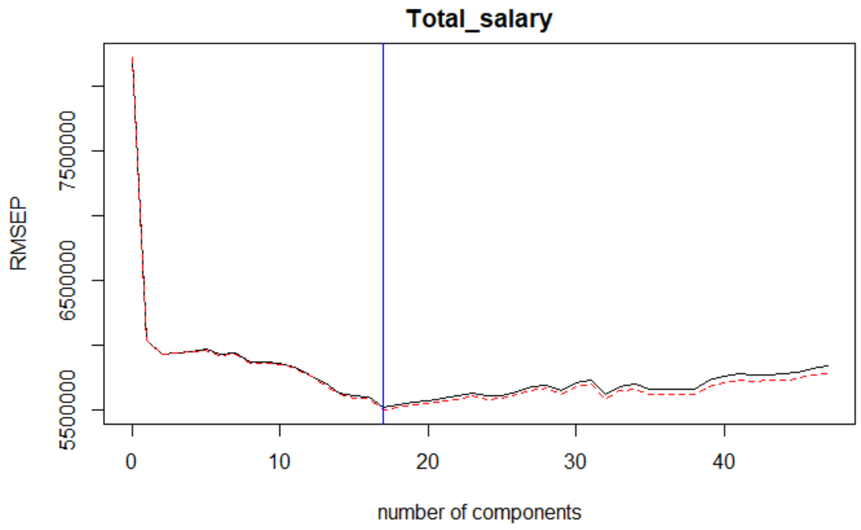
The accuracy just doesn't improve

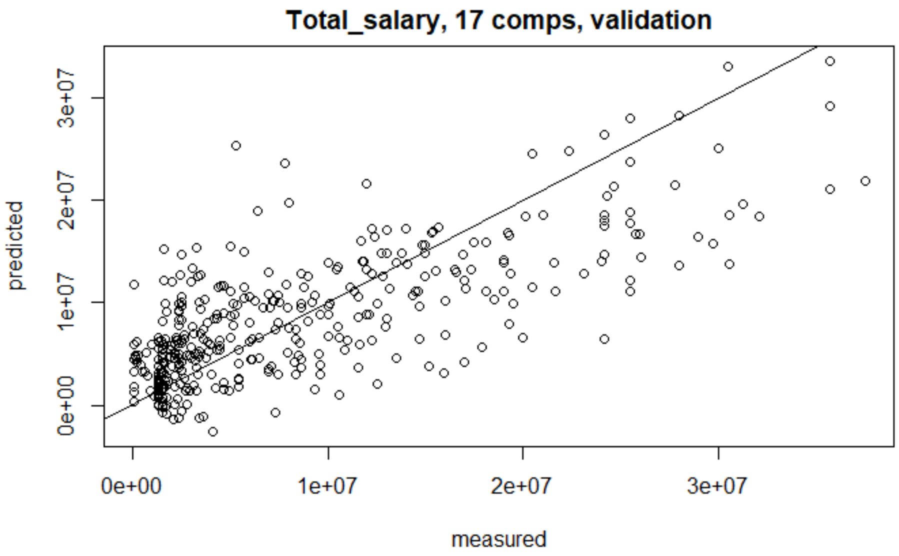
...ctd

Modeling technique	Refinements tried	RMSE obtained	Variance Explained
PCR	1. Chose # principal components model based on min RMSE	USD 5.07M	-
Boosting	1. 31 variables selected based on importance() function	USD 5.24M	54.31%
Neural networks	1. Choose the size and decay factor with minimum RMSE	USD 617K	

Principal Component Regression

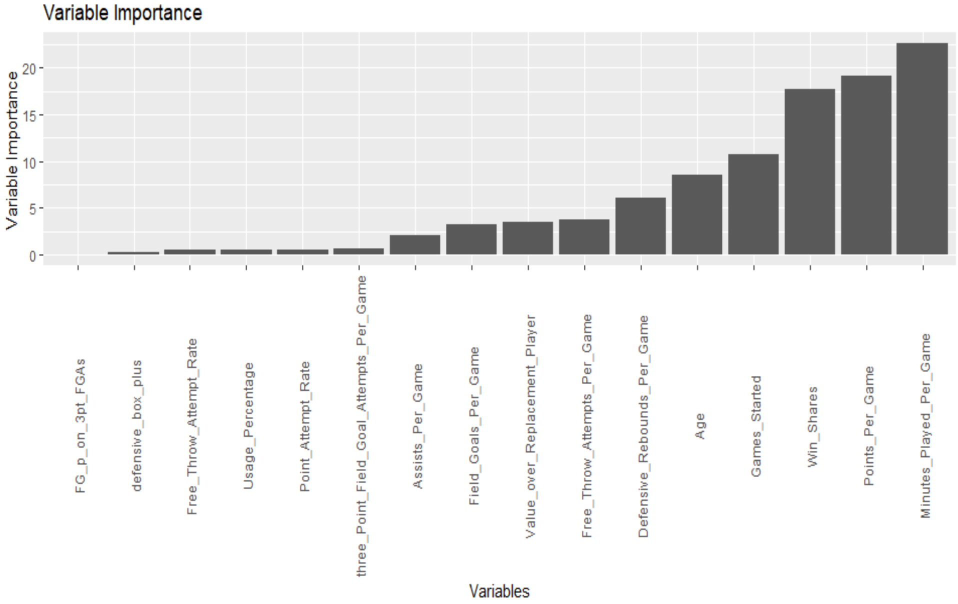
- Number of Components : 17
- Test RMSE : 5.07 Mil
- Adjusted R_squared: 57.34%

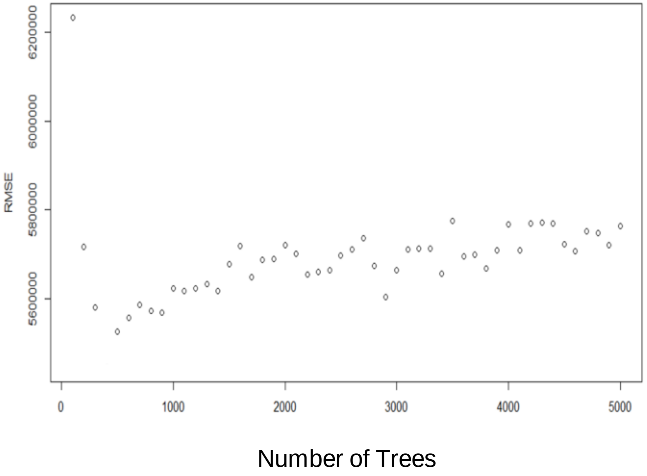




Boosting :

Best Iteration with shrinkage factor=0.01 and #Trees = 500, : RMSE: 5.4Mil



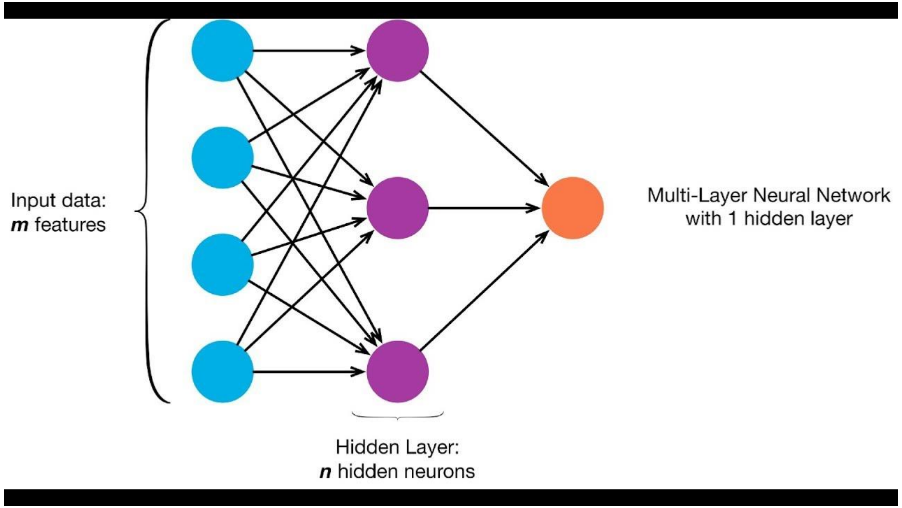


The accuracy just doesn't improve

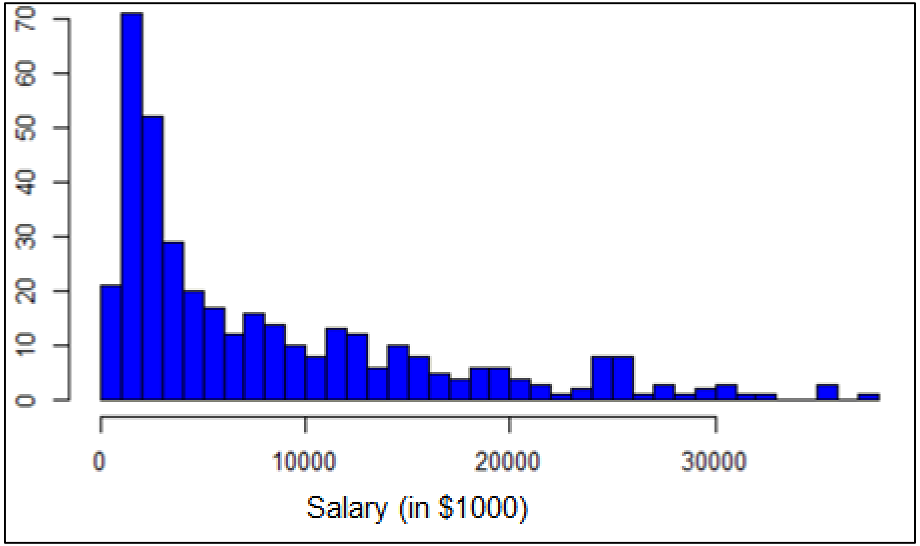
Neural Network :

Best Iteration with decay factor=0.01 and size= 10

RMSE obtained : $ 617K



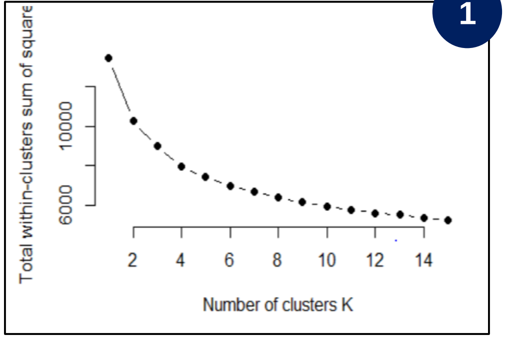
We observed that irrespective of any variations, there was still an average error of USD 5M associated with our model, and we explored why…



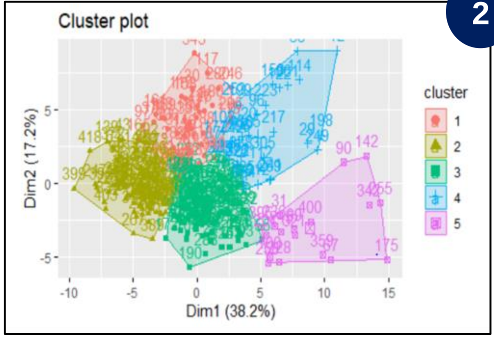
- The skewness of the dependent variable could be the reason behind the high RMSE
- Our hypothesis was that variation in salary could be better explained if similar salaried players are grouped together
- Since salary is what we're trying to predict, we tried using other independent variables to group the players, as a proxy

Grouping the players into different clusters, reduced the variation in salaries and combined similar players together

1



2



3

Cluster <int>	avg_sal <dbl>
1	5689368
2	3357706
3	8709291
4	14069664
5	22619060

Linear regression

- 28% and 4,649,000 - k=5
- 35% and 4,566,000 - k=4
- 48% and 6,016,000 - k=3
- 47.8% and 6,616,000 - k=2

- Optimal number of clusters were selected based on the elbow curve. Well, 5 seemed like a good point (Maybe not?)
- The average salaries across the 5 clusters indicates that the objective of grouping the players based on their salaries is achieved to an extent
- However, increase in number of clusters, decreased the data points in each clusters to ~60 per cluster

Improving Predictions in the future

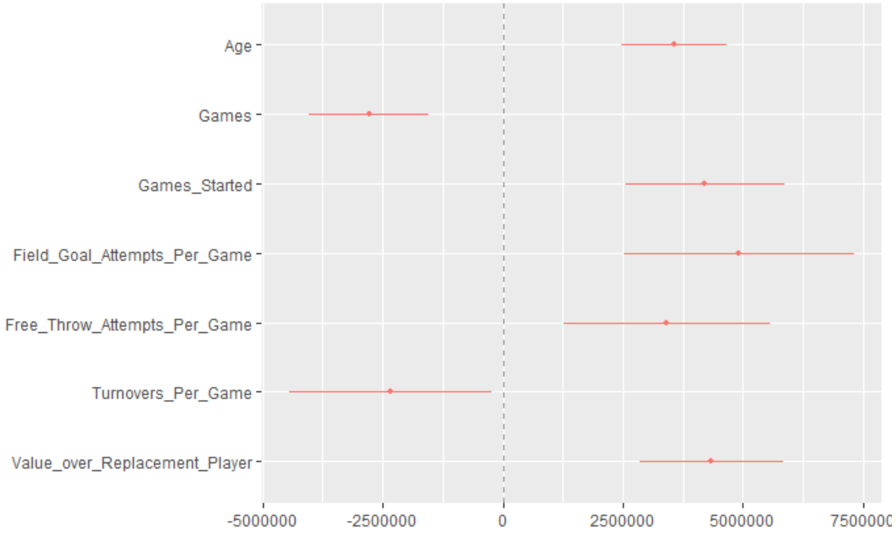
- Examining only contracts signed in the previous season
  - Contracts signed during different years use a different salary cap number

- Examining the contracts in subsets
  - Some players such as LeBron James are undervalued because the cap doesn't allow for him to be paid market value

Questions?

Linear Regression

Test RMSE : 5,140,189
Best subset method : Backward



Salary = - 7,862,915 + 429,094×Age - 48,917×Games + 66,699×Games started + 547,183×Field goal attempts per game + 1,041,936×Free throw attempts per game - 1,429,831×Turnovers per game + 1,580,038×*Value over replacement player*