Contents

Motivation and introduction of the problem

For our project, we wondered if it is possible to predict an NBA player's salary based on his performance and his personal profile. We also wanted to know what effects player salary the most. We employed three different techniques to try and predict salary, which are smoothing splines, random forests, and boosting. For splines, we divided the explanatory variates into offensive stats, defensive stats, and miscellaneous profile. We explored the relationship and interactions within each group, and then combined the three group of variates for a final model. For each of random forests and boosting, we used cross validation to find the most appropriate complexity, and fitted a final model on those complexities. Finally, we compared the cross validation score of all three models to pick the best model. We also have a baseline model which we compared our final model against(we will talk more about how we constructed this baseline model and why it is an appriopriate baseline model).

Solving this problem is useful for two reasons. If players understand what affects their salary the most, they can work on those areas in the future and develope their value. Teams can also use the prediction to evaluate if they are overpaying or underpaying a player, which helps them plan their budget.

Data and Preprocessing

We initially started looking at the data of NBA players at https://www.basketball-reference.com/contracts/players.html (updated constantly), which had 582 records of player salaries for year 2017-2018 at the time. We had to remove some duplicated records for players with different salaries on different teams. This is because some players could get cut by teams half way through the season, and sometimes they would get picked up by another team, which resulted in having multiple player contracts in a year. An example for this is Rajon Rondo, who was waived by the Chicago Bulls, signed a contract with New Orleans Pelicans right after.

During the process of matching players' stats/profile with their salaries, we encountered some player are missing a profile. An example is Walt Lemon, Jr., who is listed in our salary data, but we couldn't find all the profile information needed. We were not able to find his player information on https://stats.nba.com/players/bio/, which contains data that we thought could be important in our analysis. Therefore, we removed these records.

For the "Position" categorical variable in our dataset, we stated in our proposal that we would be using 5 values, PG (Point Guard), SG (Shooting Guard), SF (Small Forward), PF (Power Forward), and C (Center). It turns out that many guards in the NBA today are "combo guards", which means they can both play at the Point Guard and Shooting Guard position (e.g. James Harden). There are also many forwards in the NBA who can both play at the Small Forward and Power Forward position (e.g. Lebron James). We reduced the number of values to 3, grouping PG and SG as G (Guard), SF and PF as F (Forward). In addition, there are some players who are "swingman", meaning they can both play at the SG and SF position (e.g. Jimmy Butler). Since this is not a frequent case, we chose a position for each of them based on which position they had mostly been playing at this season (2017-2018) and our understanding of their strength. Because of such swing players, we went through all the data by hand and assigned the most appriorate role to our knowlege.

Our final data set contains 515 records and does not contain any N/A's. Our major explanatory variates are: PTS (Points), REB (Rebounds), AST (Assists), TOV (Turn Overs), STL (Steals), BLK (Blocks), Team, WR (Win Rate), AGE, FGM (Field Goals Made), FGPER (Field Goals Percentage), TPM (Three Points Made), TPPER (Three Point Percentage), FTM (Free Throw Made), FTPER (Free Throw Percentage), PF (Personal Fouls), PLUSMINUS (Plus or Minus), Position, Country, MIN (Minutes), and Draft.Round. Some of these variates are per game statistics, including PTS, REB, AST, TOV, STL, BLK, FGM, TPM, FTM, PF, and MIN.

We then realized a couple outliers in our dataset. For example, Gordon Hayward was horribly injured during his very first game at the beginning of the year. He was not able to return for the rest of the season. With the 4th highest salary on our list, he would be an extreme outlier in our models with minimal statistical contribution. However, this does not mean that he is not worth the salary, since he was only able to play for

about 5 minutes before the injury. However, there are only a few exceptions we think the techniques we use should be robust enough to handle them.

```
##   X Unnamed..0          Name  Salary Team AGE GP  W  L  MIN  PTS  FGM
## 1 0            0 Stephen Curry 34682550  GSW  30 51 41 10 32.0 26.4  8.4
## 2 1            1  LeBron James 33285709  CLE  33 76 46 30 37.1 27.6 10.6
## 3 2            2  Paul Millsap 31269231  DEN  33 32 17 15 29.3 14.8  5.4
##    FGA FGPER TPM TPA TPPER FTM FTA FTPER OREB DREB REB AST TOV STL BLK  PF
## 1 16.9  49.5 4.2 9.8  42.3 5.5 5.9  92.1  0.7  4.4 5.1 6.1 3.0 1.6 0.2 2.2
## 2 19.4  54.7 1.8 4.9  36.1 4.6 6.3  73.0  1.2  7.4 8.6 9.1 4.2 1.5 0.9 1.7
## 3 11.2  48.2 1.1 2.9  36.6 2.9 4.2  70.7  1.4  4.8 6.3 2.8 1.9 1.2 1.1 2.6
##   PLUSMINUS Position Country Draft.Round        WR
## 1       9.5        G     USA           1 0.8039216
## 2       0.6        F     USA           1 0.6052632
## 3       2.3        F     USA           2 0.5312500
```

We first wanted to explore the distribution of Salary, since that is our preditive variable and arguably the most important variable. By graphing the variable, we realized it is heavily skewed, and we tried to fix this by applying a power transformation. After trying multiple different α, we determined $\alpha = 0$ (i.e. log transformation) gave the best result.

Histogram of data$Salary

Histogram of log(data$Salary)
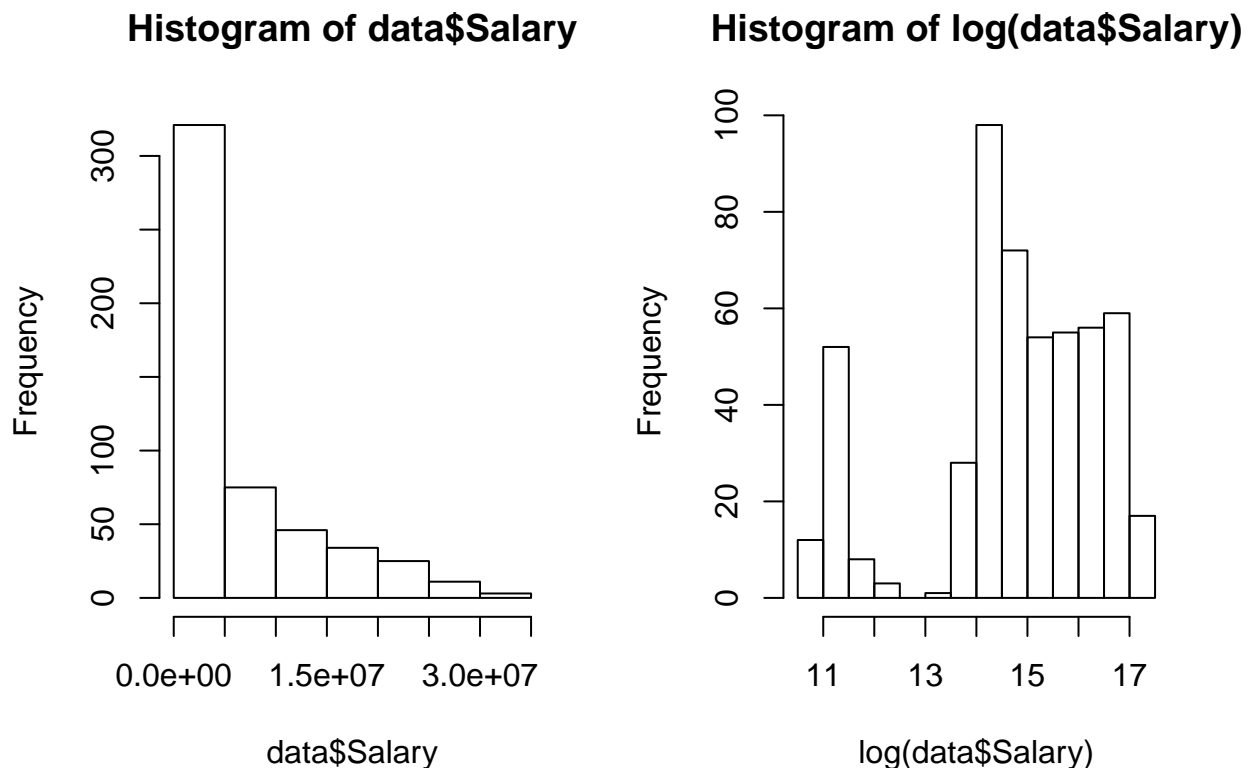
However, after the log transformation, the distribution of salary becomes bimodal. Intuitively, this is because there is a clear seperation of salary between players just entering NBA on their first contract, and players on their subsequent contracts This salary gap is expected because NBAis known to have a mostly standard salary for new players. We thought age might be a predictor of whether players are above or below the gap, but after plotting we see this is not the case. An possible explanation is certain players are entering NBA from the development league instead of out of college or highschool. Such players would be older, but their

salary varies drastically. If they are pulled in to replace an injured player, they might still be in the middle of a development contract, in which case the salary would be very low. However, if they have proved themselves in the development league and is now signing a new contract with NBA, they might be paid above the salary gap. So development players would be older, but could be above or below the wage gap, and this stops us from using age to identify the gap.
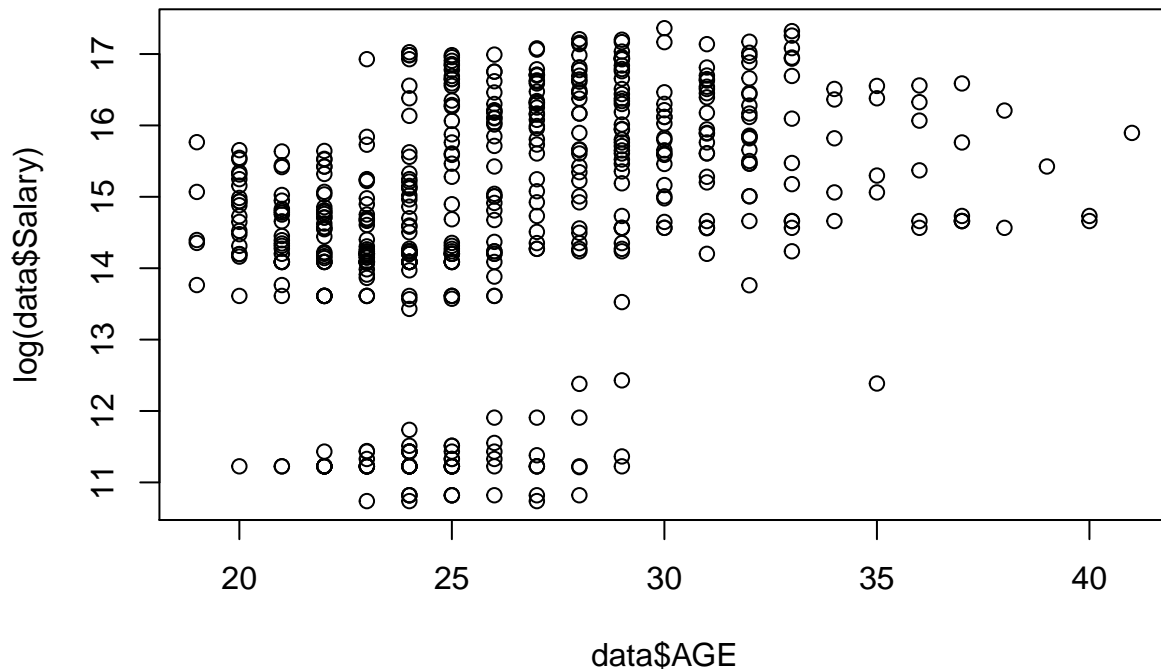
There are couple groups of variables in our data that we know are each internally correlated. For **each** of the three ways to score - field goals, three points, and free throws - there are three corresponding variates - number of attempted shots, number of made shots, and success percentage(number made/number attempted). For each of the ways to score, we only want to use one variate. Intuitively, the number of attempt tells us the least about a player's offensive abilities, so we will only compare between number of made shots and success percentage. For each way to score, we fit a spline model for salary against shots made, and another spline model for salary against success percentage, and pick the variate corresponding to the model with lower BIC score.

It turns out number of shots made is always the preferred, and always has quite lower BIC score. This could be because some players take very few shots, so the variance in success percentage for these players is quite high and does not accurately reflect their skill level.

```
library(splines)
fgm <- lm(log(Salary) ~ bs(FGM, degree=4), data=data )
fgper <- lm(log(Salary) ~ bs(FGPER, degree=4), data=data )
BIC(fgm, fgper)

##        df       BIC
## fgm     6 1803.453
## fgper   6 1923.919
```

```
tpm <- lm(log(Salary) ~ bs(TPM, degree=4), data=data )
tpper <- lm(log(Salary) ~ bs(TPPER, degree=4), data=data )
BIC(tpm, tpper)
```

```
##        df     BIC
## tpm     6 1955.945
## tpper   6 1988.426
```

```
ftm <- lm(log(Salary) ~ bs(FTM, degree=4), data=data )
ftper <- lm(log(Salary) ~ bs(FTPER, degree=4), data=data )
BIC(ftm, ftper)
```

```
##        df     BIC
## ftm     6 1883.222
## ftper   6 1913.473
```

We chose to use degree of freedom of 4 with Bezier by testing different degrees of freedoms and comparing them with BIC. Let's graph salary vs. field goal made to make sure 4 is a reasonable degree of freedom.

Bezier Spline(df=4)
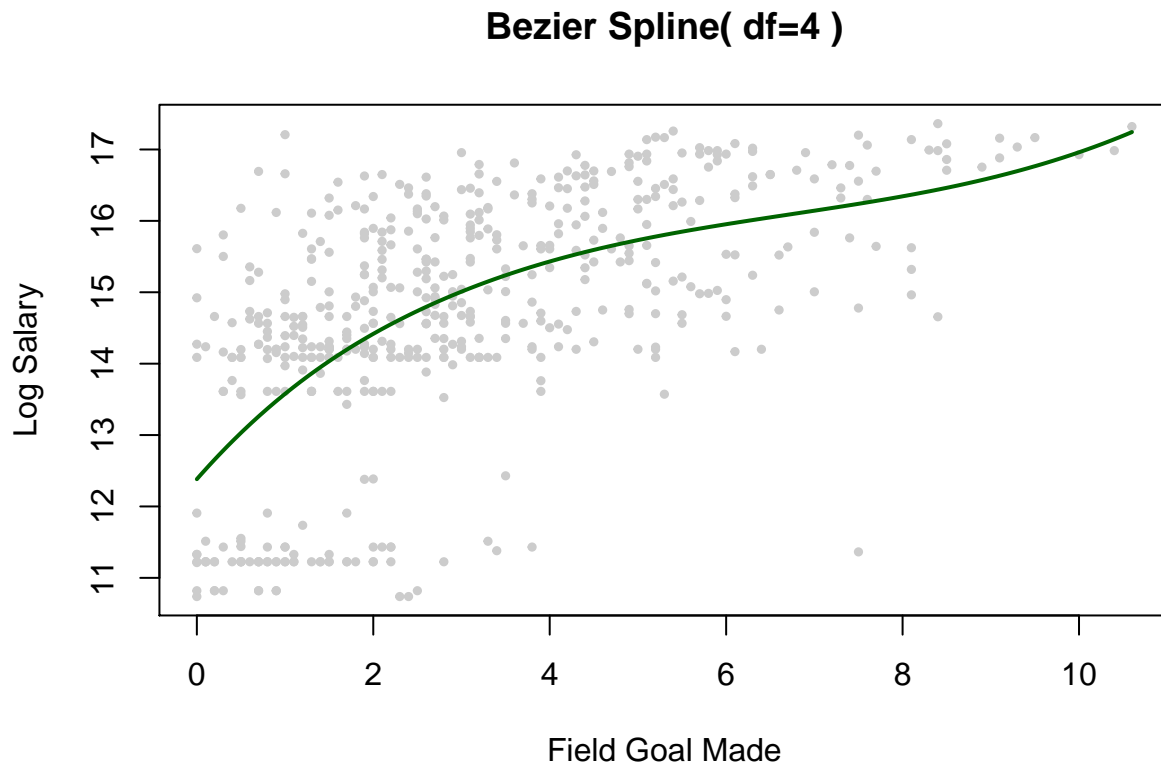
Splines

The variates we still have can be arranged into three groups: Offensive Stats, Defensive Stats, and Miscellaneous Stats. offensive: (Total Points, Field Goals Made, Three Points Made, Free Throws Made, Assists, Turnovers), defensive:(Total Rebound, Offesive Rebound, Defensive Rebound, Steals, Blocks), miscellaneous: (Team, Age, Games Played, Win Rate, Minutes Played, Plus or Minus, Position, Country, Draft Round). We

suspected the offensive abilities are correlated with each other, and the defensive abilities are correlated with each other. Let's explore each of these groups separately.

Offensive Stats

We first wanted to explore the relationships between offensive stats. We started by looking at scoring stats, which includes Field Goals Made, Three Points Made, and Free Throws Made. Since we are using the number of shots made, the total Minutes Played could be a correlating factor, so we also consider correlation with Minutes Played.

```r
library(mgcv)
```

```
## Warning: package 'mgcv' was built under R version 3.4.4

## Loading required package: nlme

## Warning: package 'nlme' was built under R version 3.4.4

## This is mgcv 1.8-23. For overview type 'help("mgcv-package")'.
```

```r
off1 <- gam( log(Salary)~s(FGM) + s(TPM) + s(FTM)
             + ti(FGM,MIN) + ti(TPM, MIN) + ti(FTM,MIN), data=data )
summary(off1)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(Salary) ~ s(FGM) + s(TPM) + s(FTM) + ti(FGM, MIN) + ti(TPM,
##     MIN) + ti(FTM, MIN)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.6482     0.1714   85.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df     F  p-value
## s(FGM)      2.189  2.814 12.469 1.96e-07 ***
## s(TPM)      4.036  5.013  1.927  0.08876 .
## s(FTM)      1.000  1.000  0.581  0.44638
## ti(FGM,MIN) 7.275  8.473  2.847  0.00506 **
## ti(TPM,MIN) 1.000  1.000  1.047  0.30667
## ti(FTM,MIN) 5.494  7.147  0.894  0.55345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.413   Deviance explained = 43.7%
## GCV = 1.7951  Scale est. = 1.7185     n = 515
```

Field Goal and Field Goal with minutes are the only significant variates, which makes sense because majority of points in every game are field goals. We also want to know if a model with total points would be better

```
off2 <- gam( log(Salary)~s(FGM) + ti(FGM, MIN), data=data)
off3 <- gam( log(Salary)~s(PTS) + ti(PTS, MIN), data=data)
BIC(off2, off3)
```

```
##              df      BIC
## off2 10.59219 1810.348
## off3 17.03626 1840.831
```

It seems using field goal made is better than using total points made. Let's look at the summary of the model with field goals.

```
summary(off2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(Salary) ~ s(FGM) + ti(FGM, MIN)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.6327     0.1674    87.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##               edf Ref.df      F  p-value
## s(FGM)      2.567  3.342 24.562 4.12e-16 ***
## ti(FGM,MIN) 6.025  7.022  2.746  0.00921 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.397   Deviance explained = 40.7%
## GCV = 1.7978  Scale est. = 1.7643     n = 515
```

It would appear we should keep both field goals made, and the interaction between field goals and minutes played, (although field goals is much more significant than the interaction)

The other offensive stats include assits and turnover. We are considering turnover as part of offensive stats because it indicates a lack of ability to score. We are now looking at Points, Assits, and Turnovers, and fit different models to explore their relationships. Again, we should take minutes played into consideration.

```
off1 <- gam( log(Salary)~s(FGM) + s(AST) + s(TOV)
            + ti(FGM,MIN) + ti(AST,MIN) + ti(TOV,MIN), data=data)
summary(off1)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(Salary) ~ s(FGM) + s(AST) + s(TOV) + ti(FGM, MIN) + ti(AST,
##     MIN) + ti(TOV, MIN)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  14.6305     0.1445   101.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##               edf Ref.df      F  p-value
## s(FGM)       1.000  1.000 15.180 0.000111 ***
## s(AST)       1.000  1.000 13.844 0.000221 ***
## s(TOV)       4.314  5.590  7.281 6.18e-07 ***
## ti(FGM,MIN) 2.851  3.366  8.279 1.18e-05 ***
## ti(AST,MIN) 1.000  1.000 15.000 0.000121 ***
## ti(TOV,MIN) 3.978  3.989  5.160 0.000752 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.423   Deviance explained = 43.9%
## GCV = 1.7394  Scale est. = 1.6882    n = 515
```

It seems that at $\alpha = 0.05$, field goal, assists, turnovers and their interactions with minutes played are all significant. Let's also explore if we should add the pair-wise interaction between field goal, assists, and turnovers.

```r
off <- gam( log(Salary)~ s(FGM) + s(AST) + s(TOV)
          + ti(FGM, MIN) + ti(AST, MIN) + ti(TOV, MIN)
          + ti(FGM, AST) + ti(FGM, TOV) + ti(AST, TOV)
          + ti(FGM, AST, MIN) + ti(FGM, TOV, MIN) + ti(AST, TOV, MIN), data=data)
summary(off)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(Salary) ~ s(FGM) + s(AST) + s(TOV) + ti(FGM, MIN) + ti(AST,
##     MIN) + ti(TOV, MIN) + ti(FGM, AST) + ti(FGM, TOV) + ti(AST,
##     TOV) + ti(FGM, AST, MIN) + ti(FGM, TOV, MIN) + ti(AST, TOV,
##     MIN)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.2579     0.2532    56.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                  edf Ref.df      F  p-value
## s(FGM)        2.718e+00  3.601  2.801 0.029582 *
## s(AST)        1.000e+00  1.000 14.048 0.000199 ***
## s(TOV)        5.286e+00  6.455  5.545 8.09e-06 ***
## ti(FGM,MIN)  2.620e+00  3.064  4.706 0.002887 **
## ti(AST,MIN)  1.000e+00  1.000  5.612 0.018219 *
## ti(TOV,MIN)  3.243e+00  3.446  4.562 0.013211 *
## ti(FGM,AST)  1.000e+00  1.000  0.592 0.441959
## ti(FGM,TOV)  1.000e+00  1.000  5.691 0.017422 *
## ti(AST,TOV)  1.000e+00  1.000  0.149 0.699238
```

```
## ti(FGM,AST,MIN) 8.298e-07 58.000  0.000 0.999918
## ti(FGM,TOV,MIN) 2.360e+00 51.000  0.061 0.113679
## ti(AST,TOV,MIN) 1.170e+00 44.000  0.044 0.122827
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.437   Deviance explained = 46.2%
## GCV = 1.7251  Scale est. = 1.6467     n = 515
```

It seems the only additional interaction we might care about is between field goal made and turnovers. Field goals made and turnovers might have correlation because they might both be positively correlated with amount of ball possesion the player has. Let's compare a model with the interaction between field goal and turnovers, and a model without the interaction.

```
off1 <- gam( log(Salary)~ s(FGM) + s(AST) + s(TOV)
              + ti(FGM, MIN) + ti(AST, MIN) + ti(TOV, MIN), data=data )
off2 <- gam( log(Salary)~ s(FGM) + s(AST) + s(TOV)
              + ti(FGM, MIN) + ti(AST, MIN) + ti(TOV, MIN) + ti(FGM, TOV), data=data )
BIC(off1, off2)
```

```
##            df      BIC
## off1 16.14357 1816.627
## off2 23.12585 1839.242
```

It seems the interaction doesn't actually create a better model. We will just stick to field goal made, assists, turnovers, and their interaction with minutes for our offensive stats.

Defensive Stats

Next we want to look at the defensive stats. We first wondered if offensive and defensive rebound can just be replaced with total number of rebound. Here, we fit a model against total number of rebound, and another model against offensive and defensive rebound, and a third model that include offensive-defensive interaction. We should also have their interaction with minutes played.

```
reb1 <- gam( log(Salary) ~ s(REB), data=data )
reb2 <- gam( log(Salary) ~ s(REB) + ti(REB, MIN), data=data )

reb3 <- gam( log(Salary) ~ s(OREB) + s(DREB), data=data )
reb4 <- gam( log(Salary) ~ s(OREB) + s(DREB) + ti(OREB, MIN) + ti(DREB, MIN), data=data )

reb5 <- gam( log(Salary) ~ s(OREB) + s(DREB) + ti(OREB, DREB), data=data )
reb6 <- gam( log(Salary) ~ s(OREB) + s(DREB) + ti(OREB, DREB)
              + ti(OREB, MIN) + ti(DREB, MIN) + ti(OREB, DREB, MIN), data=data )

BIC(reb1, reb2, reb3, reb4, reb5, reb6)
```

```
##             df      BIC
## reb1  5.330851 1838.985
## reb2 11.637655 1812.457
## reb3  8.972840 1844.285
## reb4 12.700596 1815.000
## reb5  9.082310 1844.510
## reb6 15.015157 1823.185
```

It seems the model with only the total number of rebound, and its interaction with minutes played is preferred Let's look at the summary for this model.

```
summary(reb2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(Salary) ~ s(REB) + ti(REB, MIN)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.97145    0.07915   189.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##               edf Ref.df      F p-value
## s(REB)      1.000   1.00 191.28  <2e-16 ***
## ti(REB,MIN) 8.638  10.47  11.16  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.401   Deviance explained = 41.2%
## GCV = 1.7898  Scale est. = 1.7528     n = 515
```

It seems we should keep both the total nubmer of rebound, and its interaction with minutes.

We now want to explore the relationship between the remaining defensive stats: Total Rebound, Steals, and Blocks. Let's first look at a model fitted without any pairwise interaction between the three, but with their individual interaction with minutes played.

```
def <- gam( log(Salary)~s(REB)+s(STL)+s(BLK)+ti(REB, MIN)+ti(STL,MIN)+ti(BLK,MIN), data=data)
summary(def)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(Salary) ~ s(REB) + s(STL) + s(BLK) + ti(REB, MIN) + ti(STL,
##     MIN) + ti(BLK, MIN)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.917      0.107   139.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                edf Ref.df      F  p-value
## s(REB)       1.039  1.067 49.677 1.47e-12 ***
## s(STL)       2.413  3.104  0.951 0.440962
## s(BLK)       1.000  1.000  0.145 0.703697
## ti(REB,MIN)  8.293  9.999  3.536 0.000154 ***
## ti(STL,MIN)  3.627  4.840  0.973 0.398642
```

```
## ti(BLK,MIN) 6.061  7.804  1.608 0.112849
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.419   Deviance explained = 44.5%
## GCV = 1.7805  Scale est. = 1.6995    n = 515
```

It seems at $\alpha = 0.05$, only rebound and rebound's interaction with minutes matter. This might be because blocks and steals barely happens in games, and has little variation.

range(data$STL)

```
## [1] 0.0 2.3
```

range(data$BLK)

```
## [1] 0.0 2.5
```

On average, even the best defensive players get less than 3 steals and rebound per game. This is just not a lot of variation in these two stats.

Intuitively, rebound and blocks should both be correlated with height and jumping power. Let's see if their correlation should be included.

```
def1 <- gam( log(Salary)~s(REB)+ti(REB,MIN), data=data)
def2 <- gam( log(Salary)~s(REB)+ti(REB,MIN)+ti(REB,BLK)+ti(REB,BLK,MIN), data=data)
BIC(def1, def2)
```

```
##            df      BIC
## def1 11.63765 1812.457
## def2 18.20550 1835.224
```

It seems the model with only rebound and its interaction with minutes is preferred. Again, this is probably because blocks barely happen in game

Miscellaneous Variates

We now want to look at the miscellaneous varaites. The discrete miscellaneous variable include Team, Position, and Country, and Draft Round. The continous miscellaneous variates include Age, Games Played, Win Rate, Minutes Played, and PLUSMINUS. We are going to ignore Win Rate since that is more a team based attribute.

Minutes played should be correlated with games played. Let's see if we need both of them.

```
misc1 <- gam( log(Salary) ~ s(MIN), data=data )
misc2 <- gam( log(Salary) ~ s(MIN) + s(GP), data=data)
misc3 <- gam( log(Salary) ~ s(MIN) + s(GP) + ti(MIN, GP), data=data)
BIC(misc1, misc2, misc3)
```

```
##             df      BIC
## misc1  3.000000 1774.380
## misc2  8.280238 1717.518
## misc3 18.201317 1755.105
```

Interestingly, it does seem that we should have both minutes played and games played, but not their interaction. Let's see the summary of the model with both variates.

summary(misc2)

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(Salary) ~ s(MIN) + s(GP)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.70403    0.05412   271.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##         edf Ref.df     F p-value
## s(MIN) 2.591  3.254 31.09  <2e-16 ***
## s(GP)  3.689  4.546 19.96  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.485   Deviance explained = 49.1%
## GCV = 1.5299  Scale est. = 1.5083     n = 515
```

It seems both of these variables are very significant.

Let's fit a model with all the continuous variates.

```
misc <- gam( log(Salary) ~  s(AGE) + s(GP) + s(MIN) + s(PLUSMINUS), data=data )
summary(misc)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(Salary) ~ s(AGE) + s(GP) + s(MIN) + s(PLUSMINUS)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.70403    0.04835   304.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                edf Ref.df      F  p-value
## s(AGE)        4.527  5.549 21.332  < 2e-16 ***
## s(GP)         4.054  4.976 15.471 2.98e-14 ***
## s(MIN)        2.503  3.156 29.646  < 2e-16 ***
## s(PLUSMINUS) 3.812  4.799  1.453    0.224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.589   Deviance explained = 60.1%
## GCV = 1.2421  Scale est. = 1.2037     n = 515
```

Plus Minus is not important, which makes sense since this score depends highly on the player's team's

performance and the other team's performance. Plus minus is commonly considered not an accurate reflection of a player's performance. Let's explore the interaction between remaining variates.

```
misc1 <- gam( log(Salary) ~ s(AGE) + s(GP) + s(MIN) , data=data )
misc2 <- gam( log(Salary) ~ s(AGE) + s(GP) + s(MIN) + ti(AGE, GP), data=data )
misc3 <- gam( log(Salary) ~ s(AGE) + s(GP) + s(MIN) + ti(AGE, MIN), data=data )
misc4 <- gam( log(Salary) ~ s(AGE) + s(GP) + s(MIN) + ti(GP, MIN), data=data )
BIC(misc1, misc2, misc3, misc4)
```

```
##             df      BIC
## misc1 13.02188 1631.569
## misc2 19.27232 1603.575
## misc3 19.07363 1624.786
## misc4 19.93758 1658.699
```

It seems the model with interaction between Age and Games Played is the best. Let's look at the summary for this model

```
summary(misc2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(Salary) ~ s(AGE) + s(GP) + s(MIN) + ti(AGE, GP)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.75334    0.04674   315.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df      F  p-value
## s(AGE)     4.407  5.386 27.728  < 2e-16 ***
## s(GP)      4.077  4.999 15.490 2.69e-14 ***
## s(MIN)     2.495  3.133 37.639  < 2e-16 ***
## ti(AGE,GP) 6.293  7.957  8.016 3.59e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.63   Deviance explained = 64.3%
## GCV = 1.1212  Scale est. = 1.0815    n = 515
```

All variates appear to be significant.

We also want to use the categorical variates. Unfortunately, we don't have enough data to treat all their interaction as levels. We will try each one of them as levels individually, and see which is the best.

```
misc1 <- gam( log(Salary) ~  s(AGE) + s(GP)  + s(MIN) + ti(AGE, GP), data=data )
misc2 <- gam( log(Salary) ~  Country + s(AGE) + s(GP) + s(MIN) + ti(AGE, GP), data=data )
misc3 <- gam( log(Salary) ~  Position + s(AGE) + s(GP) + s(MIN) + ti(AGE, GP), data=data )
misc4 <- gam( log(Salary) ~  Team + s(AGE) + s(GP) + s(MIN) + ti(AGE, GP), data=data )
misc5 <- gam( log(Salary) ~  Draft.Round + s(AGE) + s(GP) + s(MIN) + ti(AGE, GP), data=data )
BIC(misc1, misc2, misc3, misc4, misc5)
```

```
##             df      BIC
```

```
## misc1 19.27232 1603.575
## misc2 61.12703 1816.771
## misc3 20.60693 1578.466
## misc4 48.70865 1747.347
## misc5 20.31410 1526.681
```

Draft Round seems to be the only variate that improved upon the model with only continous variates. Let's look at the summary of the model with Draft Round as levels.

```
summary(misc5)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(Salary) ~ Draft.Round + s(AGE) + s(GP) + s(MIN) + ti(AGE,
##     GP)
##
## Parametric coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          15.14356    0.06074 249.298  < 2e-16 ***
## Draft.Round2         -0.57944    0.11352  -5.104 4.74e-07 ***
## Draft.RoundUndrafted -1.17897    0.12270  -9.609  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df      F  p-value
## s(AGE)     4.467  5.457 30.298  < 2e-16 ***
## s(GP)      3.172  3.918 14.936 2.72e-11 ***
## s(MIN)     2.391  3.004 25.957 9.50e-16 ***
## ti(AGE,GP) 6.283  8.013  6.296 8.33e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.685   Deviance explained = 69.6%
## GCV = 0.95763  Scale est. = 0.92171   n = 515
```

It seems all variates are still significant.

Combined Model

Our final model should include the offensive, defensive, and the miscellaneous variables we explored, with the interactions between each group that we explored. Offensive and defensive abilities in the NBA are generally thought not to correlate, so we will not add correlation between them. In the misc variate, we considered the interaction between MIN and offensive and defensive variates. We also suspect Position to correlate with offensive and defensive stats, so we will also include a model with position interacting with the offensive and defensive stats using the **by=** keyword.

```
fit1 <- gam( log(Salary) ~ Draft.Round + s(AGE) + s(GP)+ s(MIN) + ti(AGE, GP)
          + s(FGM) + s(AST) + s(TOV) + ti(FGM,MIN) + ti(AST,MIN) + ti(TOV,MIN)
          + s(REB)+ti(REB,MIN), data=data )

fit2 <- gam( log(Salary) ~ Draft.Round + s(AGE) + s(GP)+ s(MIN) + ti(AGE, GP)
```

```
                + s(FGM, by=Position) + s(AST, by=Position) + s(TOV, by=Position) + ti(FGM,MIN, by=Position)
                + s(REB, by=Position)+ti(REB,MIN, by=Position), data=data )
```

```
BIC(fit1, fit2)
```

```
##             df      BIC
## fit1 34.47762 1552.280
## fit2 63.88260 1652.473
```

It seems we should keep the interaction with position doesn't help. Let's look at the summary of the model with just Draft Round as levels.

```
summary(fit1)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(Salary) ~ Draft.Round + s(AGE) + s(GP) + s(MIN) + ti(AGE,
##     GP) + s(FGM) + s(AST) + s(TOV) + ti(FGM, MIN) + ti(AST, MIN) +
##     ti(TOV, MIN) + s(REB) + ti(REB, MIN)
##
## Parametric coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           14.8414     0.1174 126.387  < 2e-16 ***
## Draft.Round2          -0.5428     0.1108  -4.901  1.3e-06 ***
## Draft.RoundUndrafted  -1.1397     0.1205  -9.456  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                edf Ref.df      F  p-value
## s(AGE)       4.658  5.662 32.694  < 2e-16 ***
## s(GP)        2.608  3.243 15.001 1.00e-09 ***
## s(MIN)       1.000  1.000 12.987 0.000345 ***
## ti(AGE,GP)   6.651  8.428  7.448 1.21e-09 ***
## s(FGM)       3.959  5.043  2.987 0.011574 *
## s(AST)       1.248  1.466  9.827 0.000478 ***
## s(TOV)       1.000  1.000 26.717 3.41e-07 ***
## ti(FGM,MIN)  2.378  2.765  2.972 0.054263 .
## ti(AST,MIN)  1.000  1.000 10.796 0.001090 **
## ti(TOV,MIN)  3.975  3.998  4.935 0.000661 ***
## s(REB)       1.000  1.000  5.607 0.018273 *
## ti(REB,MIN)  1.000  1.000  0.889 0.346110
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.713   Deviance explained = 73.1%
## GCV = 0.89822  Scale est. = 0.83983   n = 515
```

Removing the stats with p-value less than 0.05, we fit our final model

```
fit3 <- gam( log(Salary) ~ Draft.Round + s(AGE) + s(GP)+ s(MIN) + ti(AGE, GP)
            + s(FGM) + s(AST) + s(TOV) + ti(AST,MIN) + ti(TOV,MIN)
            + s(REB), data=data )
```

```
summary(fit3)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(Salary) ~ Draft.Round + s(AGE) + s(GP) + s(MIN) + ti(AGE,
##     GP) + s(FGM) + s(AST) + s(TOV) + ti(AST, MIN) + ti(TOV, MIN) +
##     s(REB)
##
## Parametric coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         14.95537    0.09151 163.432  < 2e-16 ***
## Draft.Round2        -0.54472    0.11101  -4.907 1.27e-06 ***
## Draft.RoundUndrafted -1.13859   0.12074  -9.430  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##               edf Ref.df      F  p-value
## s(AGE)      4.592  5.589 32.477  < 2e-16 ***
## s(GP)       2.715  3.371 15.425 3.20e-10 ***
## s(MIN)      1.000  1.000 13.551 0.000258 ***
## ti(AGE,GP)  6.656  8.448  7.324 1.78e-09 ***
## s(FGM)      4.978  6.139  1.788 0.096202 .
## s(AST)      1.763  2.296  9.569 4.36e-05 ***
## s(TOV)      1.000  1.000 24.775 8.88e-07 ***
## ti(AST,MIN) 1.000  1.000 16.401 5.94e-05 ***
## ti(TOV,MIN) 3.914  3.992  4.808 0.000813 ***
## s(REB)      1.000  1.000  5.573 0.018627 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.71   Deviance explained = 72.7%
## GCV = 0.90351  Scale est. = 0.84804   n = 515
```

Finally let's look at the 5-fold cross validation score.

```
library(gamclass)
```

```
## Warning: package 'gamclass' was built under R version 3.4.4
```

```
CVgam( log(Salary) ~ Draft.Round + s(AGE) + s(GP)+ s(MIN) + te(AGE, GP)
          + s(FGM) + s(AST) + s(TOV) + te(AST,MIN) + te(TOV,MIN)
          + s(REB), data=data, nfold=5 )
```

```
##    GAMscale CV-mse-GAM
##      0.8512     0.9607
```

The apse is 0.9607.

Random Forest

We would like to utilize random forest to determine the importance of explanatory variates.

```
## Warning: package 'randomForest' was built under R version 3.4.4
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##              IncNodePurity
## PTS               62.22661
## REB               53.34613
## AST               23.43383
## TOV               66.27728
## STL               29.47236
## BLK               20.27620
## Team             105.08425
## WR                30.81825
## AGE               55.89737
## FGM               80.02194
## FGPER             49.00783
## TPM               21.53979
## TPPER             20.27744
## FTM               41.06172
## FTPER             27.62203
## PF                63.19264
## PLUSMINUS         23.53919
## Position           9.96612
## Country           14.87307
## MIN               84.62544
## Draft.Round       75.41164
```

```
##                %IncMSE
## PTS           11.105887
## REB           11.896329
## AST            6.181612
## TOV            9.891187
## STL            6.904804
## BLK            3.324602
## Team           2.144388
## WR             6.179159
## AGE           19.519539
## FGM           11.336279
## FGPER          9.771315
## TPM            7.432304
## TPPER          2.759934
## FTM            7.165750
## FTPER          6.106386
## PF            11.466293
## PLUSMINUS      3.255276
## Position       4.748315
## Country        1.216483
## MIN           15.405465
## Draft.Round   17.672138
```
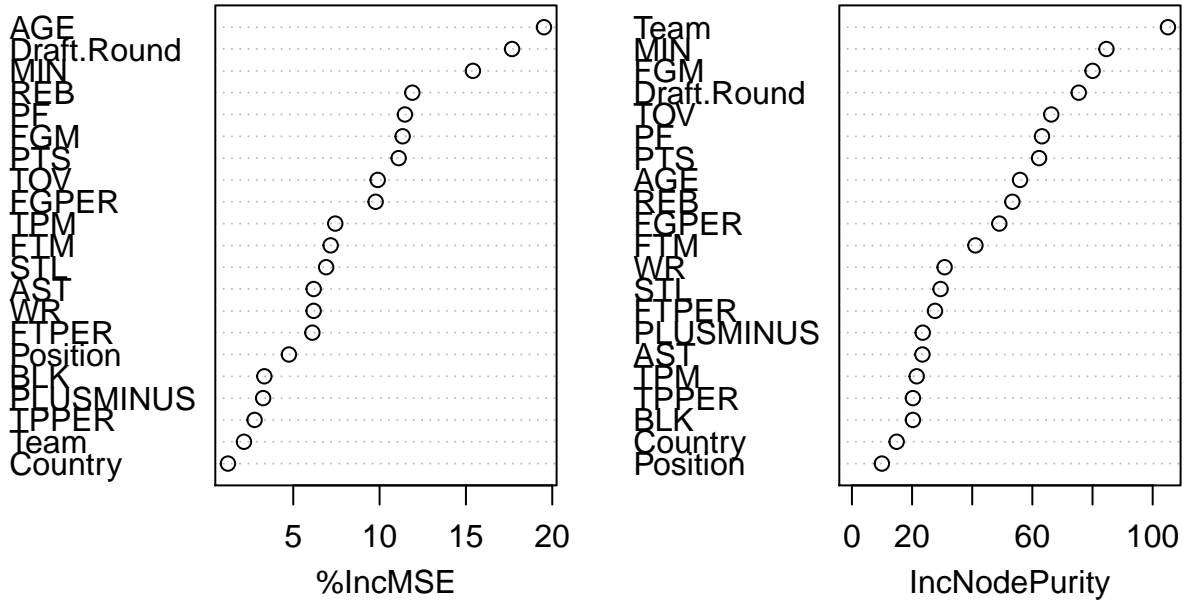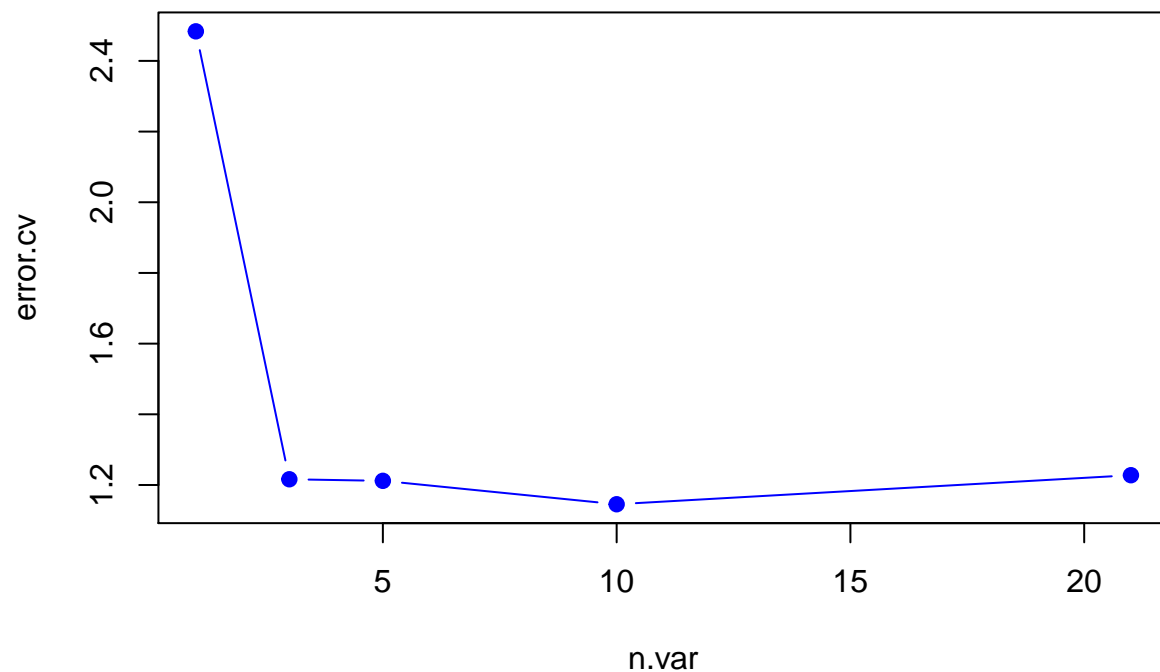
17

data.rf

We used PTS, REB, AST, TOV, STL, BLK, Team, WR, AGE, FGM, FGPER, TPM, TPPER, FTM, FTPER, PF, PLUSMINUS, Position, Country, MIN, and Draft.Round against log(Salary) for the random forest. We did not choose to include Draft.Number because it is a categorical variate with 60 different potential values, but random forest does not accept categorical predictors with more than 53 categories.

The result suggests that the error of cross validation is the lowest for 10 explanatory variates, at about 1.18. We then choose the top 10 most important variates based on RSS, Team, MIN, FGM, Draft.Round, TOV, PF, PTS, AGE, REB, and FGPER, and run the process again.
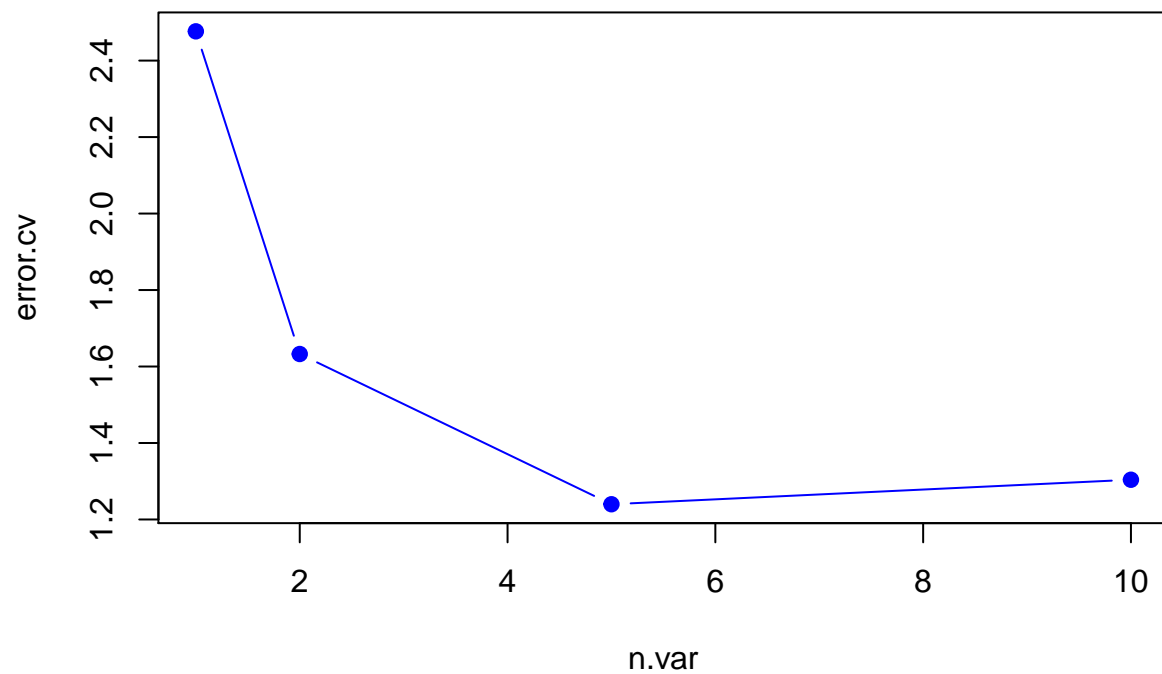
The result from the second run suggests that the error of cross validation is the lowest when there are 5 explanatory variates, at around 1.22. We then choose the top 5 most important variates againbased on RSS, which are Team, MIN, FGM, PTS, and AGE, and run the process again.
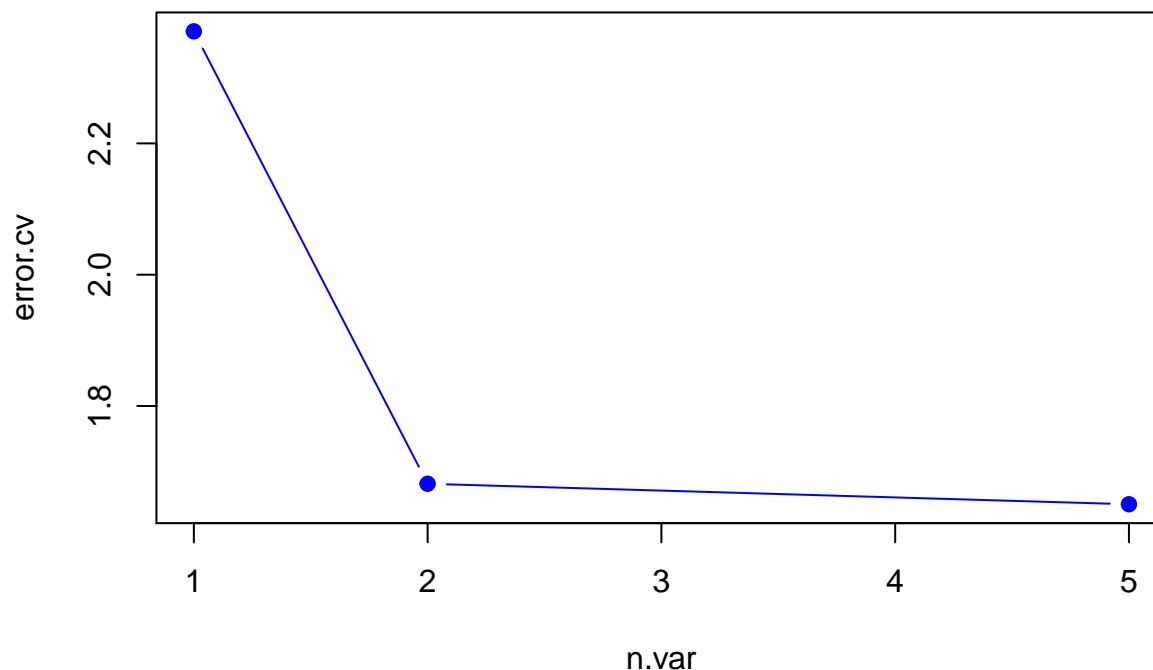
The result from the third run suggests that the error of cross validation is the lowest when there are 5 explanatory variates, at around 1.64. We can say that the Team, MIN, FGM, PTS, and AGE are important variates based on cross validation. Also, AGE seems to be the most important for predictive purposes.

## Player self-evaluate and improvements

For NBA players who would like to self-evaluate and who are trying to see what they can work on to receive a better contract, we can consider how Salary depends on just those explanatory variates that were under the control of the NBA player. Therefore, we removed Team, MIN, WR, AGE, Draft.Round, and PLUSMINUS. Age is obviously an uncontrollable variate. We think players rarely have control for Team, MIN, and Draft.Round since it does not depend on players' previous NBA performance, instead, these would depend on the decisions from coach and the organization Also, WR and PLUSMINUS have a lot to do with the teammates of the NBA player we are trying to analyze, so we decided to take these out of consideration as well. This move left us with 15 explanatory variates.

```
##           IncNodePurity
## PTS           104.02535
## REB            92.10307
## AST            39.37527
## TOV           106.13224
## STL            48.97887
## BLK            30.92172
## FGM           111.32397
## FGPER          78.87601
## TPM            33.60315
## TPPER          41.25709
```
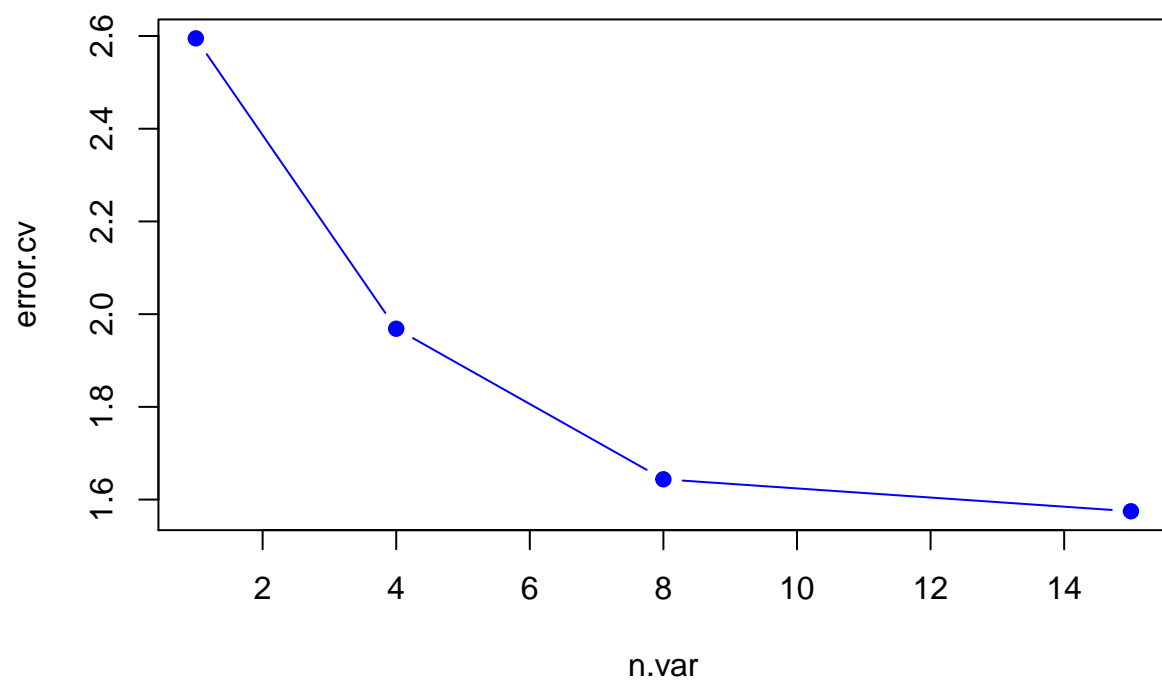
```
## FTM          63.38671
## FTPER        50.17051
## PF           91.36678
## Position     20.87296
## Country      29.39918

##             %IncMSE
## PTS        15.293625
## REB        13.995687
## AST         5.889595
## TOV        14.865521
## STL         9.409757
## BLK         5.070758
## FGM        15.475682
## FGPER      11.682218
## TPM         9.311112
## TPPER       2.879952
## FTM         8.987660
## FTPER       8.374388
## PF         15.156821
## Position    7.256454
## Country     3.896222
```

data.rf2
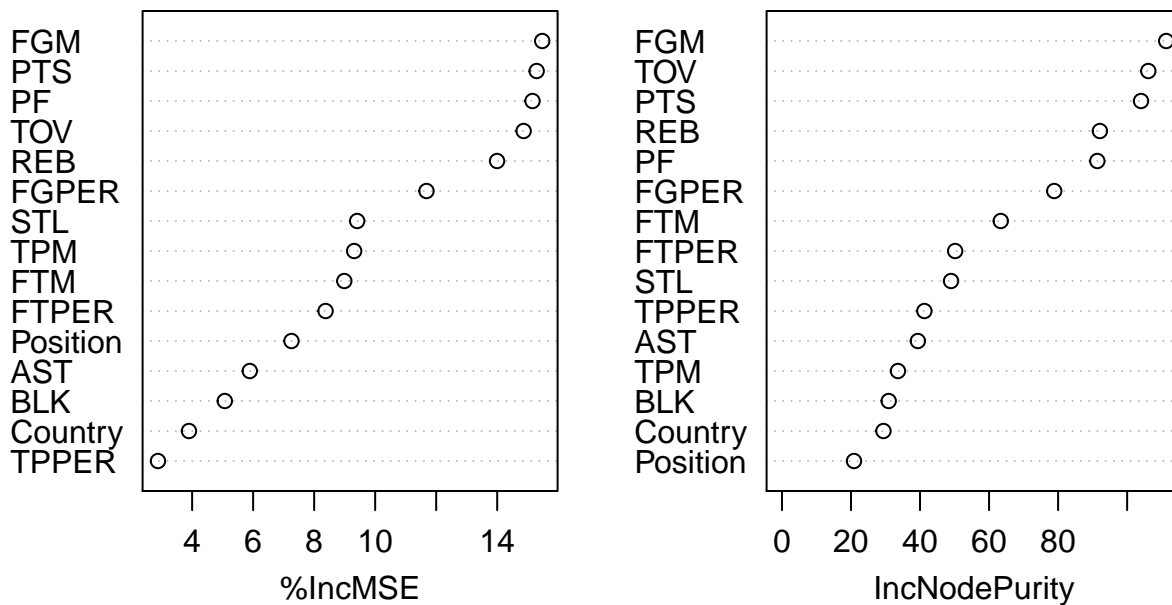
The cross validation suggests that all 15 explanatory variates are important, at an error of around 1.58. The result also suggests that FGM, TOV, REB, and PTS are the most important.

Boosting

We then used the Gradient Boosting method to determine the importance of explanatory variates, and see if it shows a different result compared to Random Forest.

```
## Warning: package 'gbm' was built under R version 3.4.4
```

```
## Loading required package: survival
```

```
## Loading required package: lattice
```

```
## Loading required package: parallel
```

```
## Loaded gbm 2.1.3
```

Effect of each attribute

Relative influence
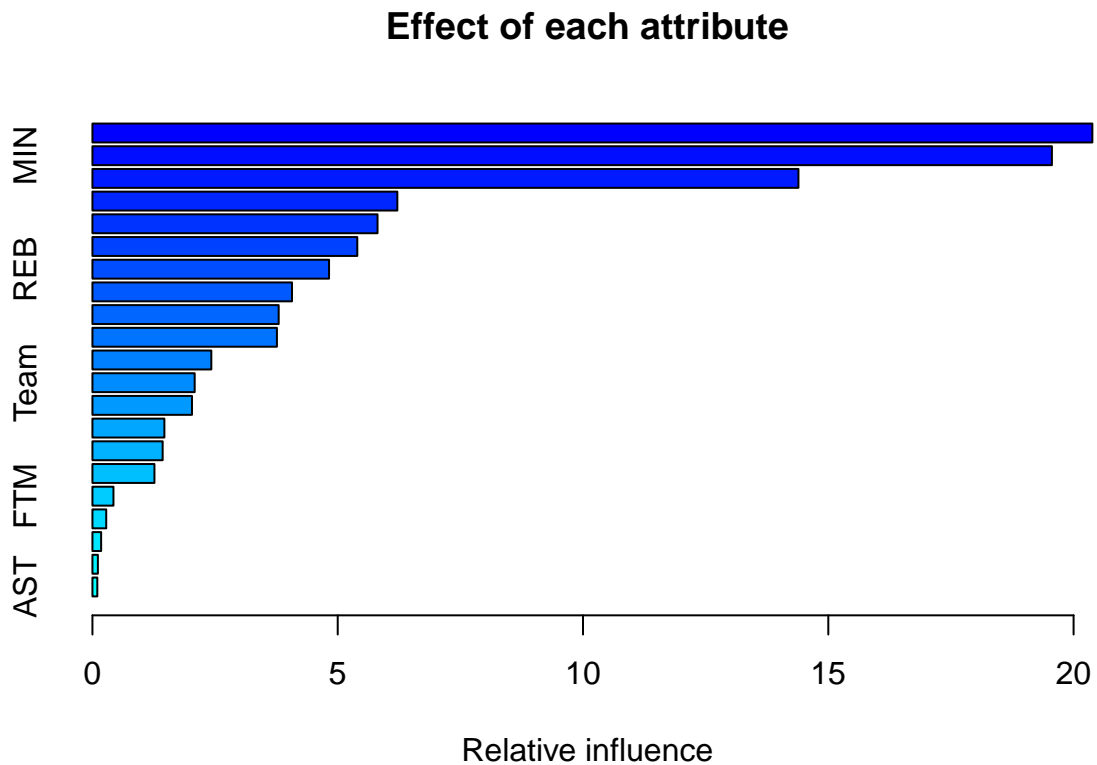
```
##                      var     rel.inf
## Draft.Round Draft.Round 20.38274616
## MIN                 MIN 19.55791333
## AGE                 AGE 14.39111747
## PF                   PF  6.21525093
## TOV                 TOV  5.81122466
## FTPER             FTPER  5.39890328
## REB                 REB  4.82325645
## PTS                 PTS  4.06743224
## Country         Country  3.79626734
## FGPER             FGPER  3.76037604
## FGM                 FGM  2.42323161
## Team               Team  2.08433641
## PLUSMINUS     PLUSMINUS  2.02944182
```

```
## Position        Position  1.46659631
## TPPER              TPPER  1.43128869
## WR                    WR  1.26356778
## FTM                  FTM  0.42841298
## TPM                  TPM  0.28111832
## STL                  STL  0.17711781
## BLK                  BLK  0.11080457
## AST                  AST  0.09959581
```

We used PTS, REB, AST, TOV, STL, BLK, Team, WR, AGE, FGM, FGPER, TPM, TPPER, FTM, FTPER, PF, PLUSMINUS, Position, Country, MIN, and Draft.Round against log(Salary), which is the same as what we used for Random Forest.

The result shows that Draft Round, Minutes played, and Age are the 3 major variates, with much higher influece over others.

cross−validated error
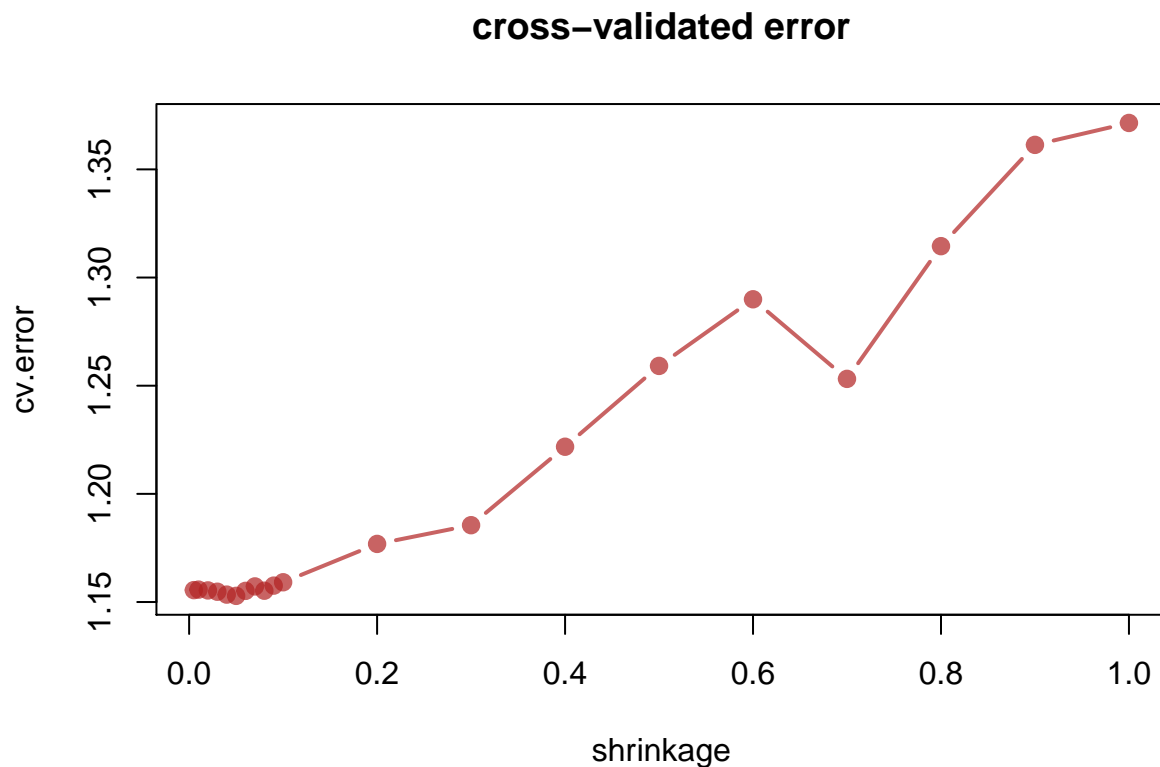
We see that the best learning rate for these 21 explainatory variates is at around 0.06, with cv error around 1.15.

cross-validated error
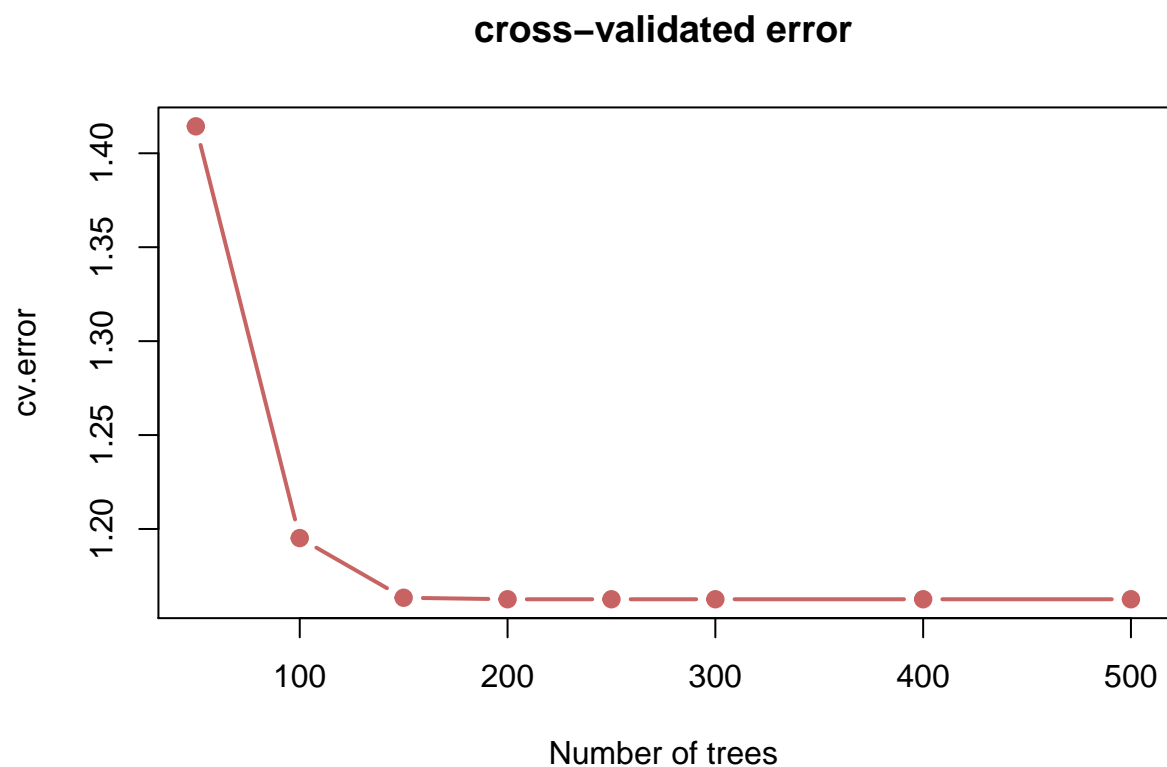
We see that the best value for M for more explainatory variables is at about 150, with cv error around 1.16. We can see that the return is very small when M is bigger than 150.

Player self-evaluate and improvements

We want to do the same thing in boosting for what we did in random tree, allowing players to see what they need to improve on the most to get a higher salary. Again, we removed Team, MIN, WR, AGE, Draft.Round, and PLUSMINUS, and repeated the process.

Effect of each attribute
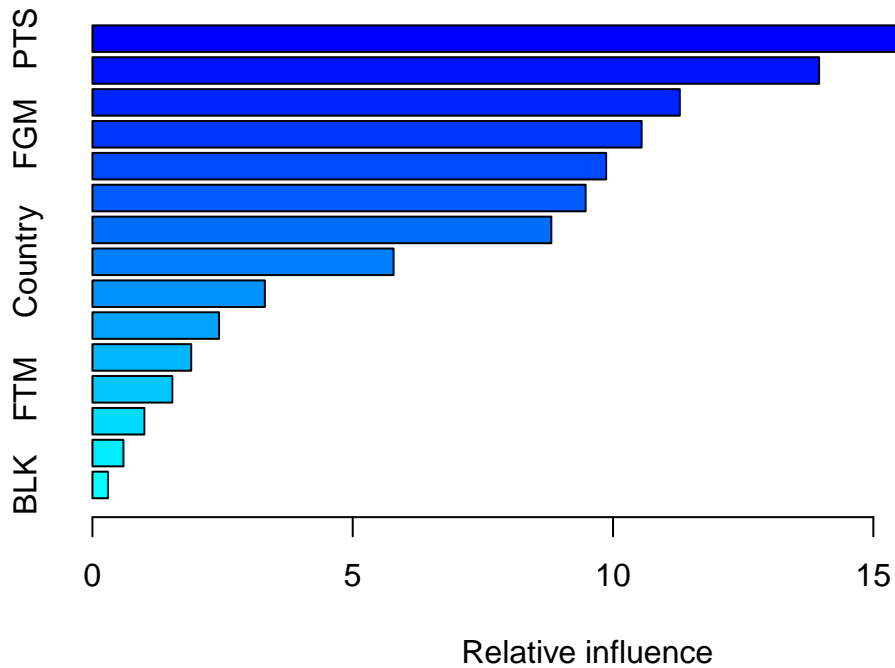

```
##                 var    rel.inf
## PTS           PTS  19.2066075
## REB           REB  13.9579523
## TOV           TOV  11.2821537
## FGM           FGM  10.5487276
## FTPER       FTPER   9.8673637
## PF             PF   9.4727863
## FGPER       FGPER   8.8143765
## Country   Country   5.7839050
## Position Position   3.3121806
## TPM           TPM   2.4321169
## TPPER       TPPER   1.8974861
## FTM           FTM   1.5347916
## AST           AST   0.9965880
## STL           STL   0.5951962
## BLK           BLK   0.2977679
```
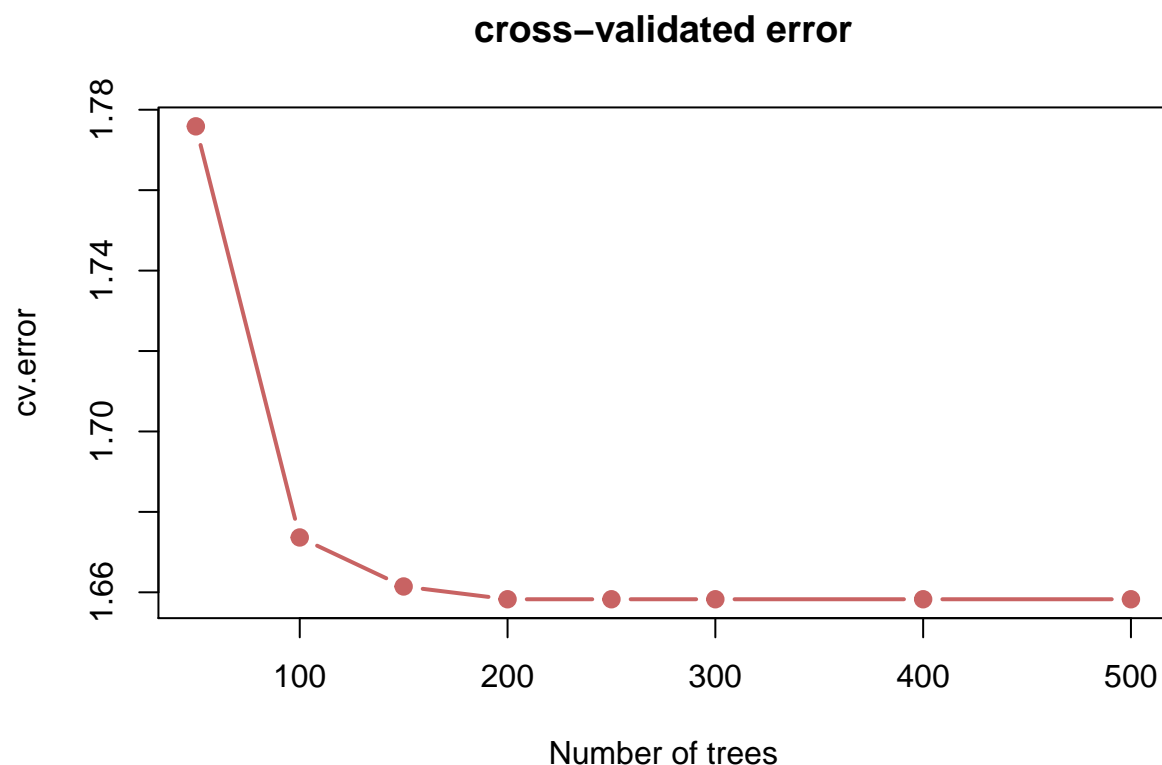
The result shows that PTS is the most important variable here, REB is the second most important, with TOV, FGM, FTPER, PF, and FGPER at 3rd to 7th, which are very close with each other.

cross−validated error
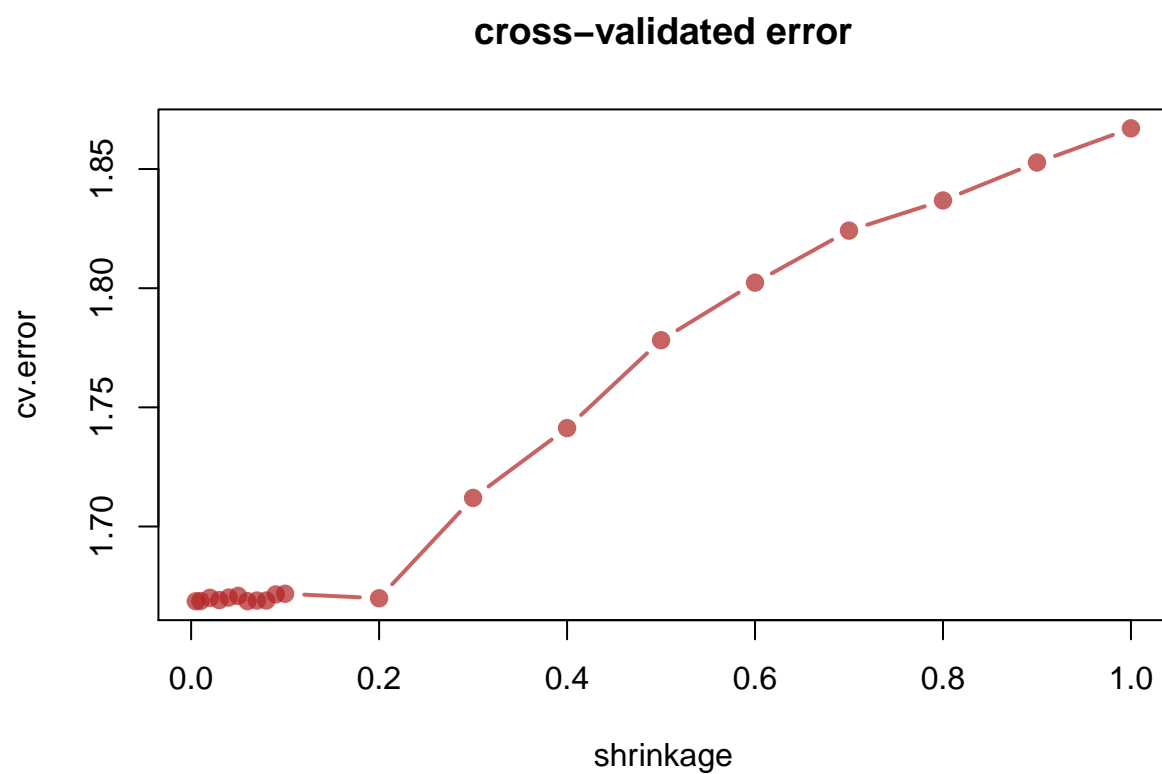
We see that the best learning rate for these 15 explainatory variates is also at around 0.06, with cv error around 1.66.

cross−validated error

We see that the best value for M for more explainatory variables is also at about 200, with cv error around 1.66. We can see that the return is very small when M is bigger than 200.

Statistical Conclusion

A popular way to determine the efficiency of an NBA player is to look at his EFF, calculated by EFF = PTS + REB + AST + STL + BLK - FGM - FTM - TOV, where all variates are averaged per game. The EFF takes Steal (STL), Block (BLK), Field Goal Missed (FGM), Free Throw Missed (FTM), and Turn Over (TOV) into account, which adds the defensive ability (STL and BLK) and inefficiency (FGM, FTM, TO) into the equation. This formula is the most popular metric for evaluating players in NBA, and a good baseline is to use EFF as the sole explanatory variate to predict salary.

Bezier Spline(df=4)
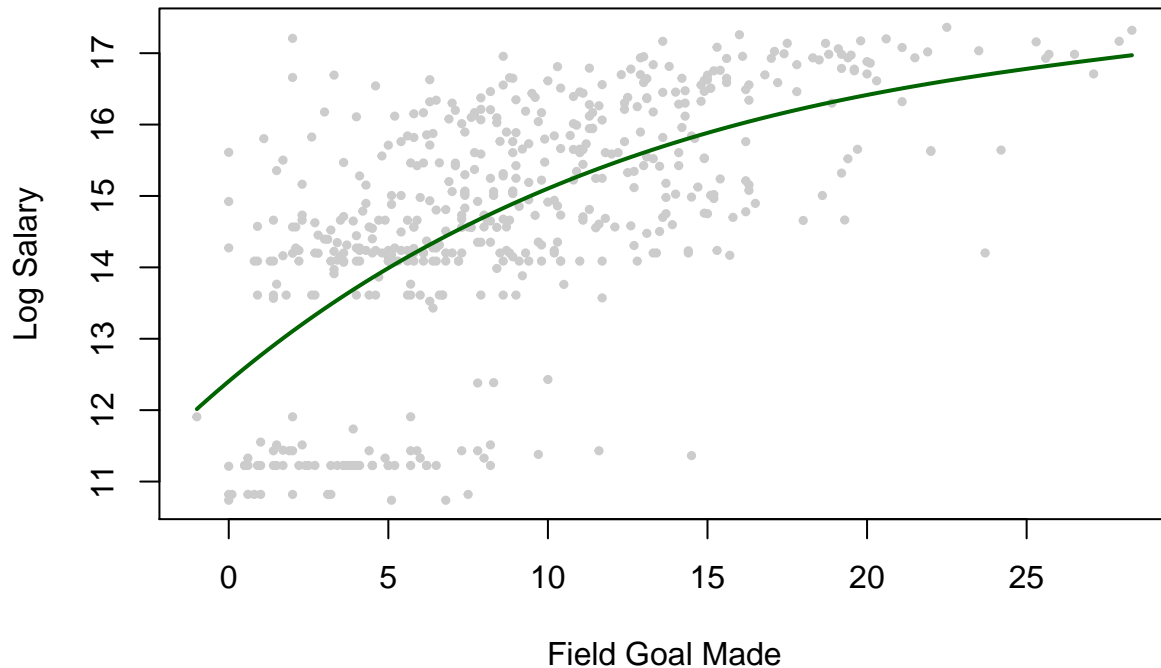

```
##     GAMscale CV-mse-GAM
##     1.7858     3.5582
```

The cross validation apse of this model is 3.5582, which is significantly higher than the cross validation value of all three of our models. This can not be used to evaluate the goodness of our models, since EFF was not developed to predict salary, but should serve as a reasonable sanity test that our fitted models are somewhat reasonable.

The smoothing method we created has the lowest apse of 0.96, while random forest has a cross validation error of at least 1.18, and gradient boosting has a cross validation error of 1.15. This model for smoothing our method is also the fastst to fit of the three method we attempted. We concluded this is the model that should be used for future salary prediction.

As mentioned before, some of the more powerful explantory variates such as age are out of the control of the player. Of the stats players can effect, the most important stats are Points(especially Field Goals), Rebound, and Turnovers.

Conclusion

We have created a good model for predicting player salary, that teams can use to evaluate if they are overpaying or underpaying a player, and help plan future budget efficiently.

For players who want to improve their salary, they should first realize there are certain factor that strongly influence their salary but are outside of their control (such as Age). Of the factor they will be able to influence, they should focus on improving their field goal scoring abilities and rebounding ability.

Future Works

Use Win Shares instead of Plus or Minus.

We initially thought that PLUSMINUS would be a very important factor in determining if the player is top of the league or not. It turns out that it is not very significant for individual players when it is modeled against salary. We think that the reason behind this is because PLUSMINUS measures the difference between the points the team scores and the points the team loses when the player is on the court, which does not only depend on individual ability, but instead the ability of all 5 players on the court. Also, coaches tend to put the best players on the court when the other team has their best players on the court, so on average, players on the court from the two teams would not have a high skills difference. Therefore, we figured that the PLUSMINUS will not have a significant impact on the players' salary.

Win Share, on the other hand, is a player statistic which is designed to estimate a player's contribution in terms of the team's wins. More information about utilizing win share in NBA can be found on https://www.basketball-reference.com/about/ws.html. We believe using Win Share instead of Plus or Minus in our model will make our model more accurate.

We did more research later into our project, and realized that players who have higher Plus or Minus are usually not the players who are the best in the NBA, but players who are on teams that have a high Win Rate (e.g. Eric Gordon, ranked 3rd in Plus or Minus, member of Huston Rockets who has the highest win rate among all NBA teams this year, ranked 90th in salary). In addition, 4 players out of the top 10 in Plus or Minus are from Huston Rockets. On the other hand, according to https://www.basketball-reference.com/leagues/NBA_2018_advanced.html, the top 10 players in Win Share are all from different teams, who are also arguably the best players in their respective teams.

Use Per Minute statistics instead of Per Game statistics

Most of our data collected are stats per game. For example, PTS, REB, and AST are all per game stats. One thing we notice during our model fitting is that MIN seemed to be a major influence. However, we know that players in the NBA never have the same minutes played. Intuitively, Minutes Played as a variable would have a very strong correlation with Salary, since teams need to put better players on the court to win games, and better players usually have higher salary. One thing we could do is to divide all per game stats by MIN to get the per minute stats. This could eliminate the relationship between MIN and per game stats in our model. One problem if we pursue this method is that the fitted values might be too random, and our models might not sugguest a relationship between per minutes stats and salary. One possible reason for this to happen is because there is only so much a player can do in a minute during a basketball game.

Contribution

All three of us worked on collecting the data, cleaning the data togther.

MingHao Lu: Motivation and Introduction, Data, Preprocessing, Smoothing Methods, Statistical Conclusion, Conclusion in Context.

Zizhou Wang: Motivation and Introduction, Data, Preprocessing, Random Forest, Future Work.

Zhaoyang Wang: Data, Preprocessing, Boosting, Appendix.

Appendix

Data Details

PTS (Points)-Average points score per game, scoring from free throws, 2 pointers, 3 pointers all counts in total score, with value 1, 2, 3 respectively. PTS = total score / games played.

REB (Rebounds)-Average rebounds per game, including offensive rebounds and defensive rebounds, each count as 1. REB = (total offensive rebound + total defensive rebounds) / games played.

AST (Assists)-Average assist per game. One assist is gained by player A when player A pass to player B and player B scores by a 2-pointer or 3-pointer right away (no other passes). AST = total assist / games played.

STL (Steals)-Average steals per game. A steal is awarded when the player gets the control of ball from the opponent team. STL = total steals / games played.

BLK (Blocks)-Average blocks per game. A block is awarded when the player successfully blocks the shot from the opponent while defending. BLK = total blocks/ games played.

PF(Personal Fouls)-Average fouls per game. A foul is a slight rule violation that is caught by the referee. Higher fouls is not preferred, but if a player does not defend properly, he can never foul! PF = total fouls / games played.

TOV (Turn Overs)-Average turn overs per game. A turnover is when a player makes a bad move and lost the control of ball. Unlike most other statistics, less turnover is preferred. TOV = total turnover / games played.

WR (Win Rate)-Winning rate is calculated as WR = Games Won/ Games Played.

AGE-Players age as of this year.

FGM (Field Goals Made)-Field goal includes 2-pointers and 3-pointers but not free throws. FGM = total field goals made / games played.

FGPER (Field Goals Percentage)-FGPER = total field goals made / total field goals attempted.

TPM (Three Points Made)-A subset of field goals, including only 3-pointers made. 3-pointers shots are made beyond the 3-point-line which is makes it the longest and (usually) the hardest shots to make. TPM = total 3 pointers made/ games played.

TPPER (Three Point Percentage)-Similarly TPPER = total 3 pointers made/ total 3 pointers attempted. We do expected TPPER to be the lowest among FTPER, FGPER and TPPER.

FTM (Free Throw Made)-Free throws is taken behind the free throw line, without any defenders. Each of the free throws counts as 1 point. Free throw opportunity can be given to a player in many cases, but the most important one is when he is fouled while attempt to shoot a normal field goal. FTM = Free throws made / games played.

FTPER (Free Throw Percentage)-Again, FTPER = total free throws made/ total free throws attempted. Because there are no defenders, FTPER is usually higher than other percentages. However, for some players play as center, because their usual shooting range is closer than the free throw line, they might struggle in free throws.

PLUSMINUS (Plus or Minus)-Plus or minus of one game is the number of points the player's team win by, while the player is on the courter playing. It could be positive or it could be negative. Here PLUSMINUS = total plus minus sum over all games / games played Position-Position can be center (C), power forward (PF), small forward (SF), shooting guard (SG), point guard (PG). Because usually a player can play multiple similar positions such as PF and SF, or C and PF, we simply our data into C, F and G, ordered from closest to basket to further away from the basket. C is more likely to get rebounds, shot 2 pointers. G is more likely to control/handle the ball and shoot from long range. On the other hand, F is a combination of both.

Country-Nationality of the player. In cases a player with multiple citizenship, we select the nation he plays for in the national basketball team. MIN (Minutes)-Average minutes played per game. MIN = total minutes played / games played.

Draft.Round-The draft number of the player. Each year there are 2 round of drafts with 30 teams picking 1 player in each round. This results in players having either 1 or 2 in this attribute, or Undrafted, if the player is not from the draft (e.g. other leagues).