Predicting NBA Player Salaries

ECON590 Research Paper

Authors: Ethan DellaMaestra, Chase Ellinger, Alec Wright

University of North Carolina at Chapel Hill

April 22, 2021

Introduction

The introduction of analytics into the NBA has been nothing short of revolutionary in its influence on the pace and playstyle of the game. This is seen most clearly in the explosion of three-point shooting. Once a niche shot that was only seen sparingly, analytics has shown the efficiency of both the three-point shot and the layup compared to the midrange game seen in the 2000's and early 2010's, and has even spawned teams such as the Houston Rockets under Mike D'antoni.

While the influence on play has been direct and observable, we wanted to look more into the effect it has had on player contracts and what box score/advanced stats are being valued these days. Our thinking is that, with the emphasis on efficiency, stats other than traditional box score stats(ppg, rbg, bpg, apg) have become less important in determining the salaries among players in the league and newer advanced stats that show efficiency(efg%,tov%,rb%,per) are becoming increasingly important in determining who gets larger contracts. This is a hot topic of conversation within the league as basketball purists see max contracts given to players like Rudy Gobert, Tobias Harris, and others who demonstrate elite efficiency, but don't stand out as superstars when looking at their traditional box score stats. We wanted to see if the league is really changing as much as people say, or if some old truths such as volume scorers still making good money still hold true.

We hope by including more obscure metrics into the mix and looking holistically at how they model into fitting salaries, we could more clearly define what combination of

skills, if any in particular, would be useful knowledge for players attempting to get the larger contracts.

Our results backed up our thinking, 3p%, win shares, value over replacement players(VORP), and player efficiency rating all played a large role in determining player salary. One difference we observed compared to others is that power forward and small forwards were the most valuable positions, which is contrary to the NBA talking point that it is a guard driven league where your shooting guard and point guard are the most important and therefore costly positions. Another sparsely discussed but intuitive result was that teams had a significant impact on salary. This makes sense as bad teams may give out large contracts to mid-tier players in an attempt to use the salary cap fully and try to become more competitive.

Literature Review

There have been a number of different analyses in regards to NBA player salaries. These studies have grown in recent years as player salary has become a very hot topic following the countless number of max contracts prevalent in the NBA today. The papers almost exclusively focus on linear regression analysis and utilize basic player statistics. These studies suggest statistics such as points scored, field goals made, 3 point percentage, and free throw percentage are most influential when it comes to determining player salary. The goal for our study is to build upon this previous analysis and incorporate more advanced player statistics and higher level regression prediction methods to construct a more effective model to predict NBA player salary.

Data

Data Collection

The data for this project was collected from two different internet sources. We scraped statistics and salary information for NBA players for the 2018-2019 season from the Basketball Reference and Hoops Hype websites. The 2018-2019 season seemed to be the best data fit for predicting NBA player salaries because the current season and last season were altered and shortened as a result of the COVID-19 pandemic. Thus, scraping data from the 2018-2019 season provided an entire season's worth of data for all of the players which would help us best predict player salaries. The Hoops Hype website provided each player's salary information while Basketball Reference provided the predictor variables utilized in our analysis. Some of these measures were descriptive such as the player's Position, Age, and Team. The remaining statistics were measures of player performance ranging from basic stats such as Points, Rebounds, Assists, Steals, and Blocks to more advanced measures such as Player Efficiency Rating(PER), True Shooting Percentage, Usage Percentage, Win Shares, Box Plus/Minus, and Value over Replacement Player(VORP). Nearly all the variables are numeric with the exception of Player Name, Team, and Position.

Data Cleaning

Because the data was collected from multiple sources, these two datasets had to be merged together. This data cleaning process used Excel to match the common variable of the two datasets, Player Name, in order to combine them. While this procedure was

fairly simple, there were a few slight issues. There were several instances in which players were present in the salary dataset but nonexistent in the player statistics dataset. After some research, this was largely attributed to a couple of reasons. The player was either recently retired or injured for the entirety of the 2018-2019 season. This meant that their salary information still existed but they did not play any games to record any statistics. With this newly attained knowledge, we decided to omit these players from the final dataset because their presence would ultimately skew our predictions. Players were still getting paid despite not actually performing on the court. Our goal was to analyze on the court performance and its impacts on player salary so removing non active players was justified. There were also several instances where players in the dataset had values of NA for certain statistics. For example, some players did not attempt a three pointer all season so their 3 point field goal percentage was NA. We decided to remove any player with missing data from our dataset as well. Ultimately, our final dataset had 475 observations (players), and 48 predictor variables. There were three additional variables not used as predictor variables. These being the response variable, salary, as well as the Player Name variable. It should also be noted that the salary variable was divided by 1 million in order to make the results more understandable.

Methods

As stated earlier, this project is attempting to figure out which player performance statistics have an impact on the player's salary in the NBA. The data was split into a

training set used for prediction and a test set used for validation. 80% of the data was used in the training set while the remaining 20% was used for validation in the test set. We used a number of different model selection methods for our prediction. These methods are Linear Regression, LASSO and Ridge regression, Elastic Net, Regression Trees and Random Forests, Boosting, Support Vector Machines (SVM), and K-Nearest Neighbors (kNN). We then were able to compare MSE and R-squared values to determine which models were performing the best.

Results

Linear Regression

Multiple Linear Regression

To begin building a predictive linear regression model, we started with a model containing all the predictor variables with the response variable being Salary. After running this model we discovered that there were several significant predictors. The predictors significant at a 0.05 level were PosPG, Age, TmMEM, TmTOT, G, TOV%, FT, and FTA. The Adjusted R Squared value was 0.5465. To attempt to improve on this model, we decided to run selection methods to only incorporate specific predictor variables. We utilized the forwards, backwards, and stepwise selection methods. The backwards and stepwise selection methods resulted in the same model so we went forward with this. The resulting model contained 15 of the 47 predictor variables. These variables were Age, G, 3PAr, FTr, AST%, TOV%, OWS, BPM, GS, FG, FGA, FG%, 3P, FT, and FTA. Every variable except for FG%,3PAr, FTR, and 3P was significant at the

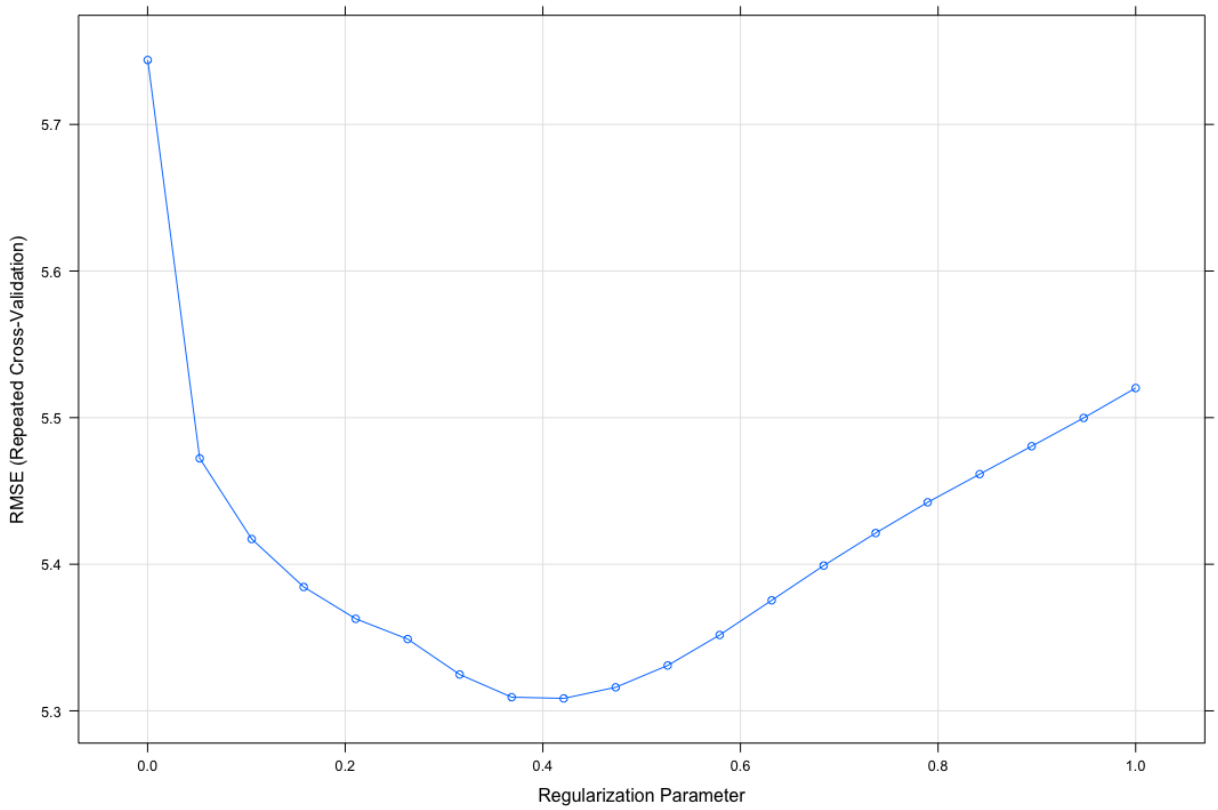
0.05 significance level. The coefficients and significance levels of the variables for this model can be seen in the table below. The Adjusted R Squared value was 0.5567 showing an improvement to the multiple linear regression model containing all predictors. The Mean Squared Error(MSE) for this model was 38.83.

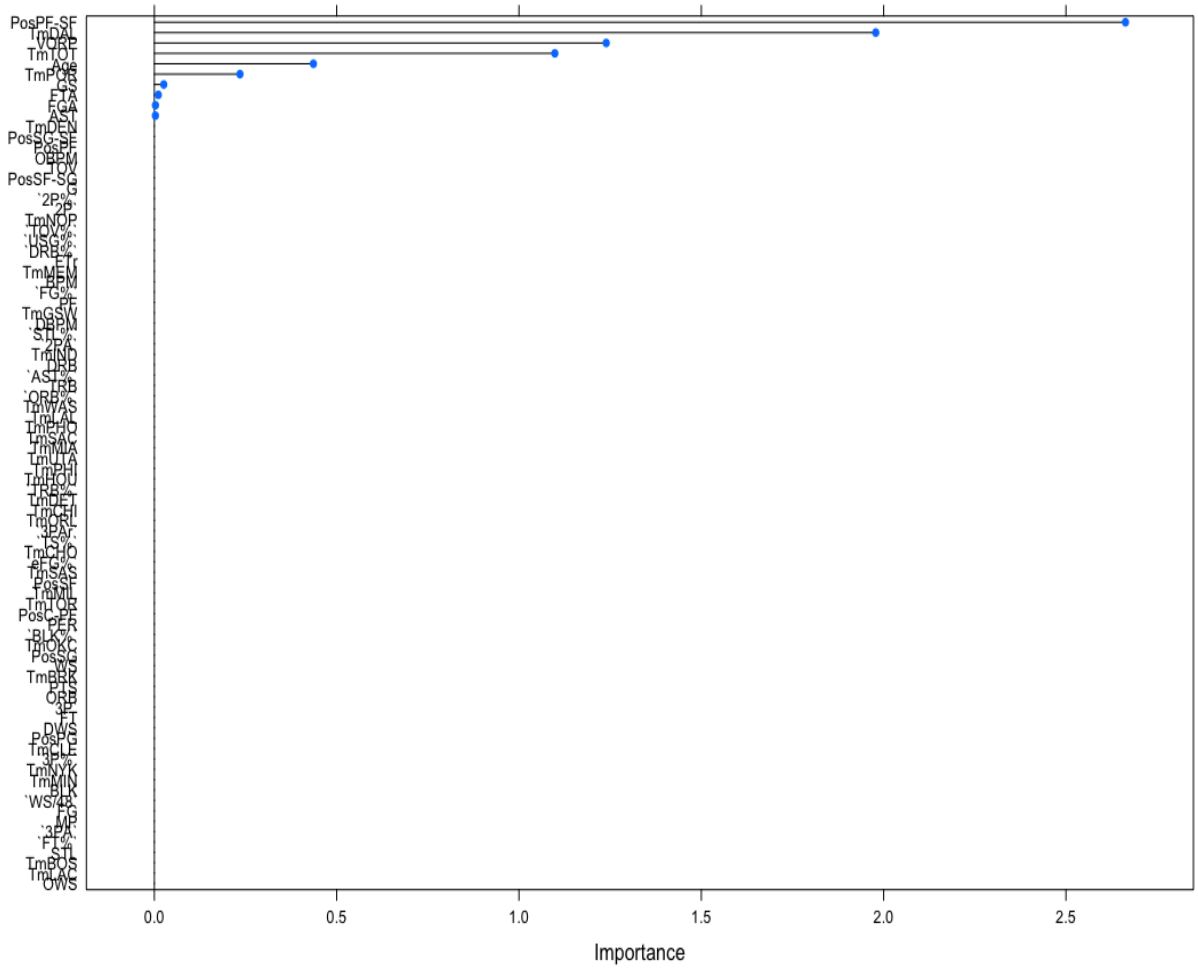
```
Call:
lm(formula = Salary2 ~ Age + G + `3PAr` + FTr + `AST%` + `TOV%` +
    OWS + BPM + GS + FG + FGA + `FG%` + `3P` + FT + FTA, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9635  -3.3120  -0.4318   2.6136  18.3303

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.22155    4.76124  -0.677 0.499077
Age           0.61251    0.06595   9.287  < 2e-16 ***
G            -0.10454    0.01979  -5.283 2.19e-07 ***
`3PAr`       -4.34256    2.91638  -1.489 0.137346
FTr          -6.20290    3.61804  -1.714 0.087300 .
`AST%`       -0.16988    0.04909  -3.461 0.000603 ***
`TOV%`        0.35581    0.08399   4.236 2.88e-05 ***
OWS           2.03946    0.46594   4.377 1.57e-05 ***
BPM           0.49130    0.15527   3.164 0.001687 **
GS            0.03817    0.01606   2.377 0.017960 *
FG           -0.10513    0.02481  -4.238 2.86e-05 ***
FGA           0.05879    0.01309   4.492 9.47e-06 ***
`FG%`       -12.97655    7.40105  -1.753 0.080386 .
`3P`         -0.03321    0.01689  -1.966 0.050108 .
FT           -0.10758    0.02441  -4.407 1.38e-05 ***
FTA           0.08919    0.01936   4.607 5.65e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.084 on 364 degrees of freedom
Multiple R-squared:  0.5743,    Adjusted R-squared:  0.5567
F-statistic: 32.73 on 15 and 364 DF,  p-value: < 2.2e-16
```

Lasso Regression

“LASSO” stands for Least Absolute Shrinkage and Selection Operator. Lasso Regression uses shrinkage to move data points closer to the central point mean and

allow for a feature selection to turn some variables coefficients into zero. Lasso encourages simple and sparse models and will remove variables that just add noise. After running the Lasso Regression with the alpha set at 1, the ideal lambda was found at .421 as shown in the graph below. The train R Square for Lasso was .5059 and the RMSE was 5.316 (MSE=28.26). The test R square for Lasso was 0.4966633 and the RMSE was 6.770760 (MSE= 45.83). It makes sense for the training set to have a higher R squared as 75% of the data was placed in the training set and should account for more of the variance for Salary. The Lasso regression greatly diminished a large portion of the variables, which is by design but could be a little too much as a large amount of importance is placed on a few variables. The Lasso regression focused greatly on variables such as VORP, a few selective teams, and POS: PF-SF. Lasso put a large emphasis on VORP and POS: PF-SF, which is quite interesting as one stat is a deep analytical stat that has only emerged recently and the other is considered the highest scoring position and potentially most important in this age of Basketball. Age was also considered an important variable which also makes sense as when you leave your rookie contract you ask for a higher and long term contract. With medicine and advancements, players are able to play in the NBA until they are 40 like LeBron James which could also be driving that stat. Teams appeared to have a consistent trend with Salary and many were considered larger variables than stats. This could be a point towards how a good percentage of explaining salary goes to teams decision making, timing the salary cap, and other external factors like relationships with fans.

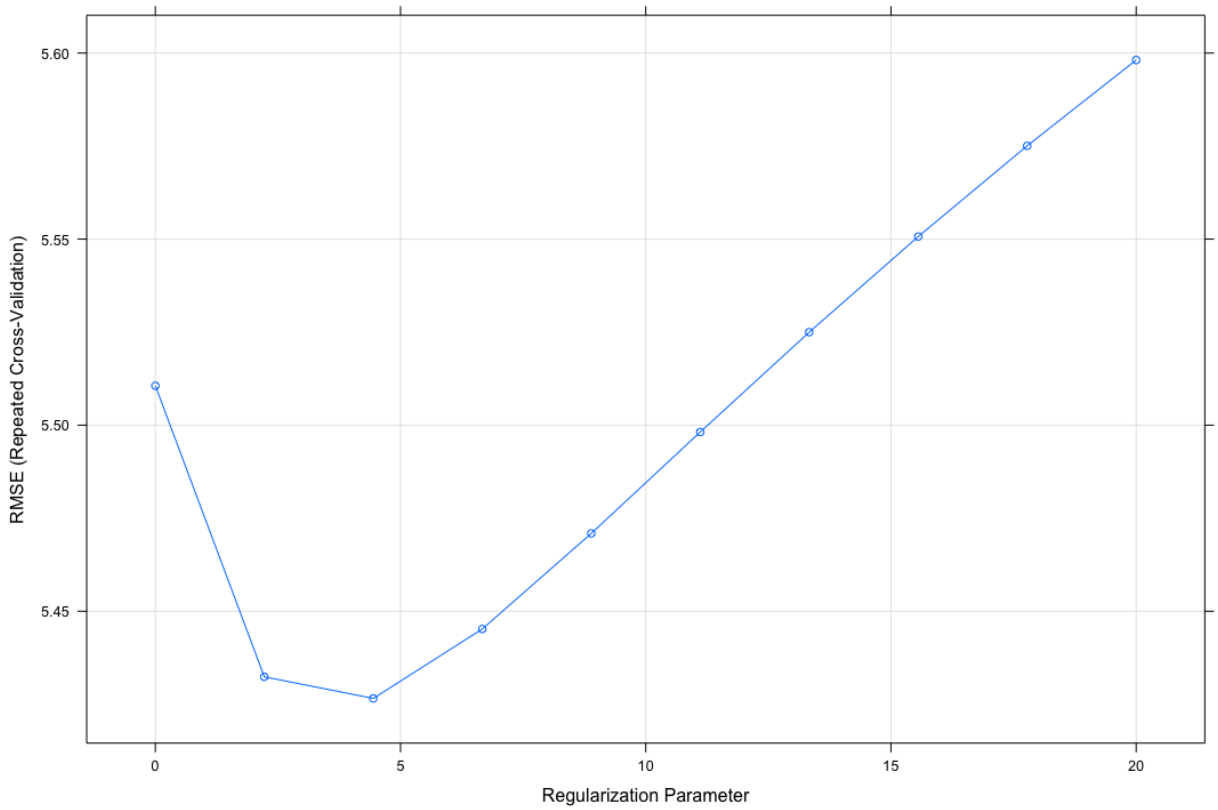

RMSE (Repeated Cross-Validation)
5.7
5.6
5.5
5.4
5.3
0.0
0.2
0.4
0.6
0.8
1.0
Regularization Parameter

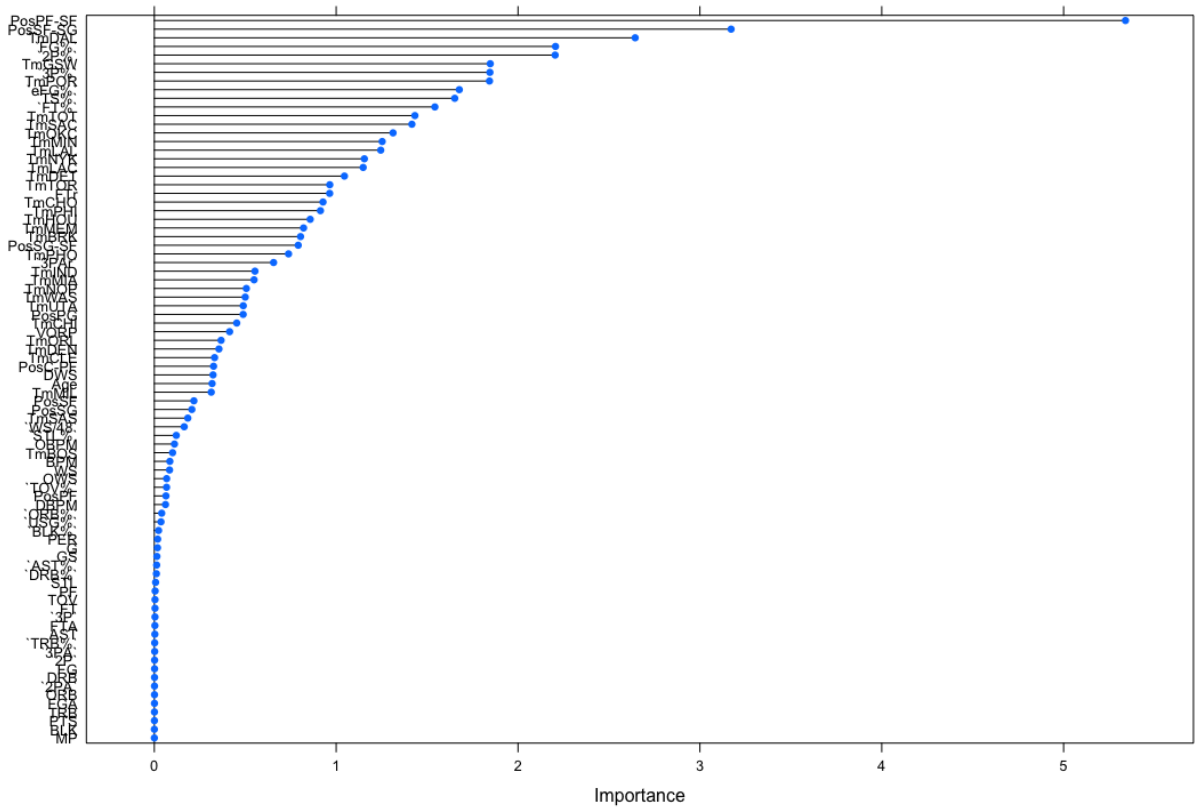


Ridge Regression

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. Ridge Regression shrinks the coefficients to non-zero values to prevent from overfitting but keeps all the values. After running the Ridge Regression with Alpha set equal to 0 it was found that 4.44 was the best lambda as shown below in the graph. The train R squared for the Ridge Regression was .4834 and the RMSE was 5.426 (MSE=29.44). The test R

squared for Ridge Regression was .4715 and the RMSE was 7.16 (MSE=51.27). Again, this makes sense since there was 75% of the data in the training compared to 25% in the test set. Ridge had similar trends as Lasso in terms of what variables were important at predicting Salary but the variables importance level varied. POS:PF-SF was the most important variable in Ridge just as it was in Lasso but Ridge gave a large percentage of stats some share of importance in predicting salary. Variables like POS:SF-SG, 2P%, FT%, 3P% and certain teams had a much higher level of importance for Ridge than in Lasso. This appears to be more accurate when looking at the league at full scale as certain players make all their money based on one skill like shooting, dunking, or just playing defense. It is clear that being a PF-SF or a guard has a much better trend of a higher salary which adds up since most of the best players are forwards or guards. An important note for ridge is that % stats and stats focused on efficiency performed really well and showed the growing trend to paying for efficient teams.

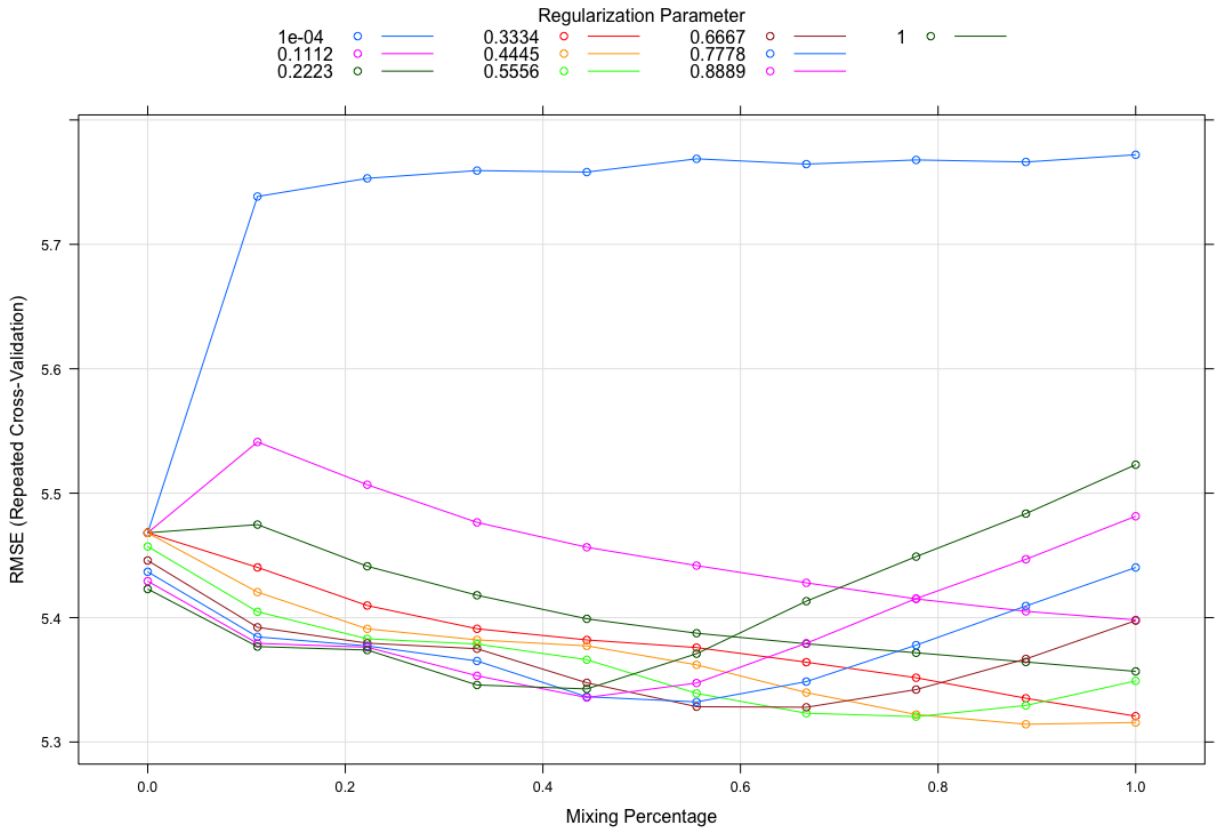

5.60
5.55
5.50
5.45
RMSE (Repeated Cross-Validation)
0
5
10
15
20
Regularization Parameter

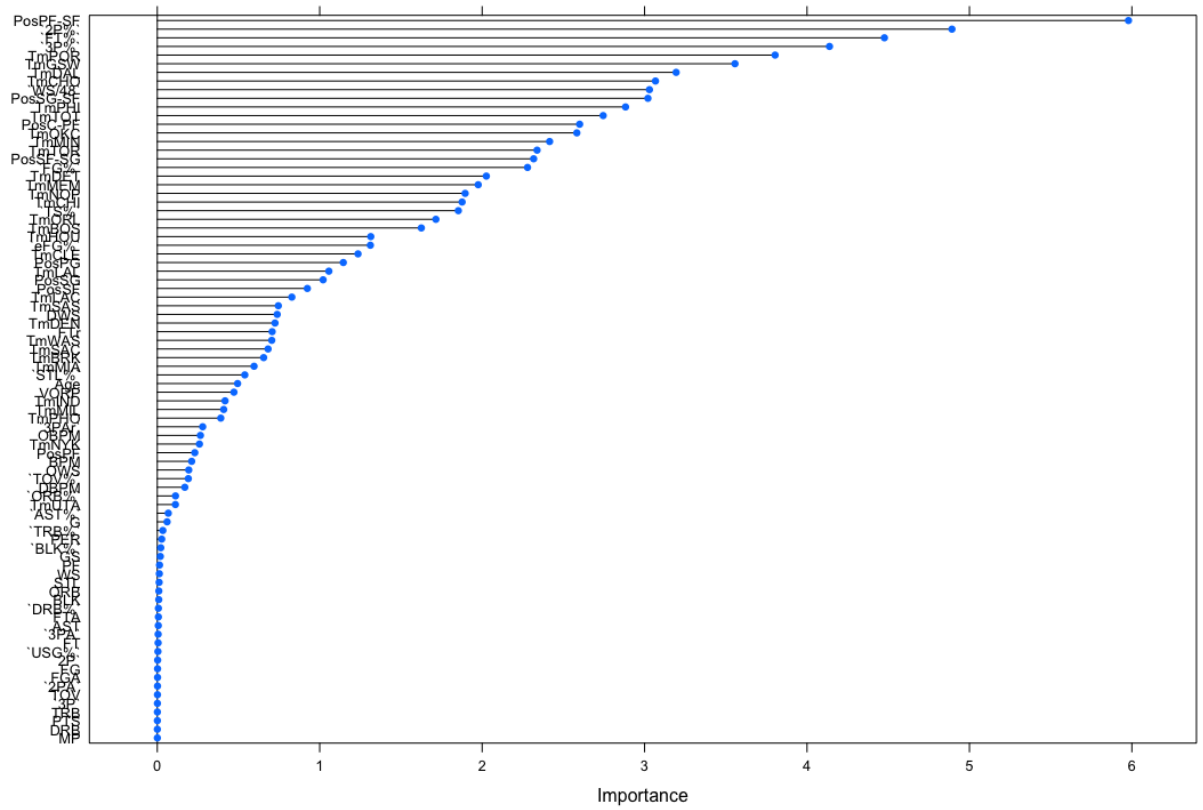


Elastic Net Regression

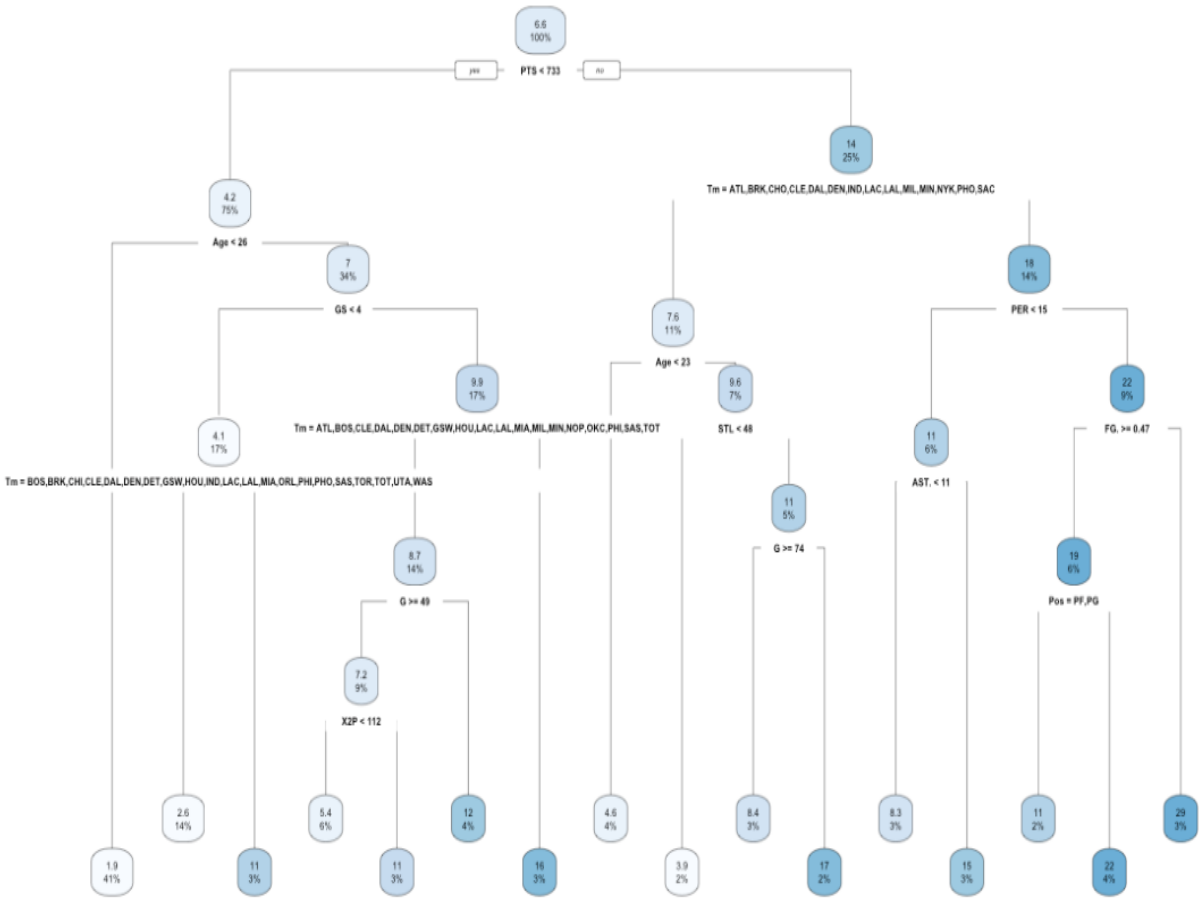
Elastic Net Regression is a combination of Lasso and Ridge regression and its advantage of the elastic net penalty is that it enables effective regularization via the ridge penalty with the feature selection characteristics of the lasso penalty. After running the Elastic Net regression to look for the ideal sequence for alpha between 0 and 1 it found that .60 Alpha was ideal. With the Alpha being higher than .5 there is a heavier Lasso penalty applied compared to the Ridge penalty. The ideal lambda was found at .445. The training R squared for Elastic was .51445 and the RMSE was 5.329 (MSE=28.4). The test set R squared for Elastic was .4884 and the RMSE was 6.91(MSE=47.75). There are two big takeaways from the Elastic regression, the first is that with an Alpha closer to 1 the Elastic model favored the Lasso format which makes

sense as Lasso had a higher R squared. The second point is that this favoring of the Lasso can also be seen when looking at the importance of each variable and how certain variables were completely excluded. It is clear that elastic used a balance of both Ridge and Lasso as many middle variables were given a decent amount of importance. Out of the 3 regressions, Elastic Net had the highest R squared and also showed to value many variables rather than just a few. Elastic did share one major trend with Ridge and Lasso in that all regressions valued POS:PF-SF as one of the most important stats. Another important note is how Elastic valued the 3P% as an extremely important stat and could be evidence of the league willing to pay players more based on their 3 point shooting percentage due to advanced analytics. Regardless of which regression was run between the 3, it is clear that certain variables that favor efficiency in NBA and also notorious positions for scoring in PF and SF were impactful variables in determining salary. The regressions also showed a trend that a large percentage of salary cannot be explained by stats and are affected by external factors such as timing the salary cap, teams poor or wise decision making, market location as smaller teams have to overpay for players.


Regularization Parameter
1e-04
0.1112
0.2223
0.3334
0.4445
0.5556
0.6667
0.7778
0.8889
1
RMSE (Repeated Cross-Validation)
5.3
5.4
5.5
5.6
5.7
0.0
0.2
0.4
0.6
0.8
1.0
Mixing Percentage

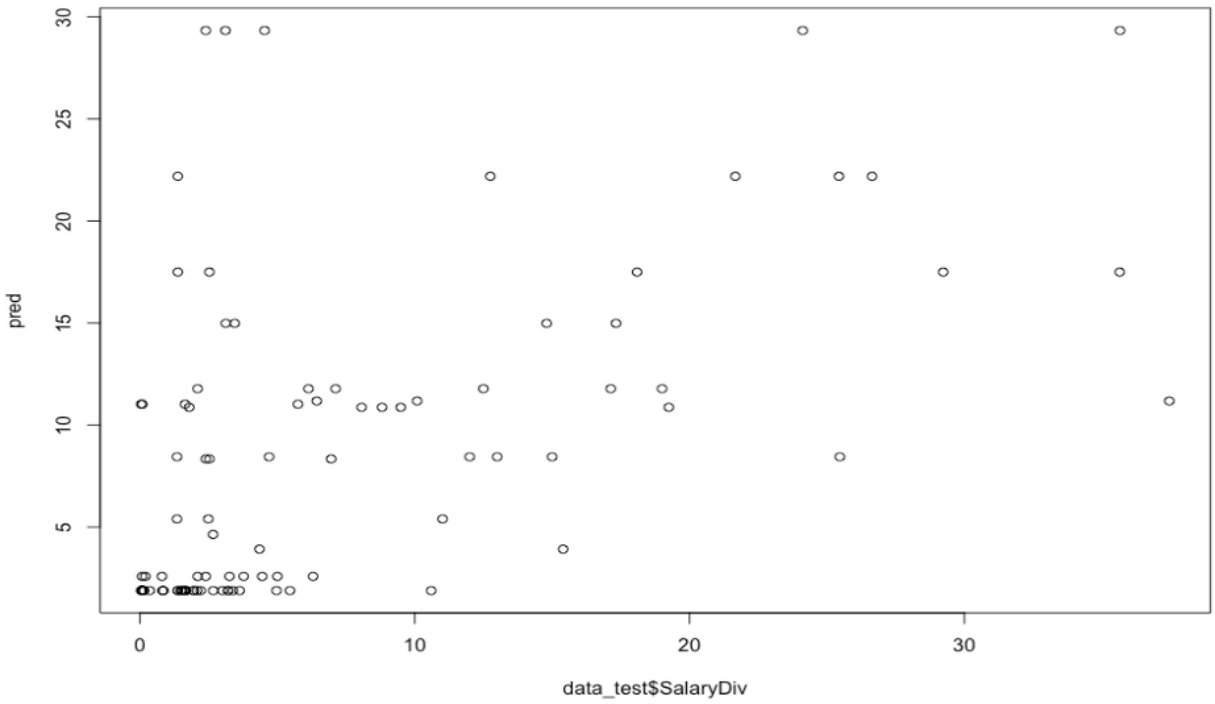

PosPF_SF
2P%
FT%
3P%
TmPOR
TmGSW
TmPTS
TmDRB
WS/48
PosSG_SF
TmPTS
TmTOV
PosC_PF
TmOKC
TmMIN
TmTOR
PosSF_SG
FG%
TmDET
TmMEM
TmNOP
TmCHI
TS%
TmORL
TmBOS
TmPHO
eFG%
TmCLE
PosPG
TmLAL
PosSG
PosSF
TmDAL
TmLAC
DWS
TmDEN
FT
TmWAS
TmSAC
TmBRK
TmMIA
STL%
Age
VORP
TmIND
TmMIL
TmPHO
FT%
OBPM
TmNYK
PosSF
BPM
OWS
TOV%
DBPM
ORB%
TmUTA
AST%
TRB%
PER
BLK%
GS
PF
WS
STL
ORB
BLK
DRB%
FTA
AST
3PA
FT
USG%
2P
FG
FGA
2PA
TOV
3P
TRB
DRB
MP
0
1
2
3
4
5
6
Importance

Regression Trees



For the regression trees the data was split into a standard 80-20 Train-Test set before the model was run. The above results show the splits that the tree decided on when performed on the training data using salary as the dependent variable. The model initially splits the groups into whether a player has scored more or less than 733 points on the season with the over 733 group earning significantly more at 14 million average salary compared to under 733 group of 4.2 million. After that splits, some other metrics came into play. Broadly speaking age, player efficiency rating, team and fg% were subdivision splits. A 23 and 26 year old split makes sense because rookie deals are

significantly less lucrative than future contracts for older players and last 4 years after being drafted into the league, with the youngest rookies being 19 and the oldest being 22 typically. Efficiency as hypothesized is also a very important factor in salary as PER and fg% splits showed big differences in the pay between the splits. The split that we found interesting, and unexpected, was a position where power forwards and point guards made less than other positions. Having been a basketball fan for a while. Those are the two most important roles in basketball today, but it could be explained by superstars in today's game playing outside of their listed roles.

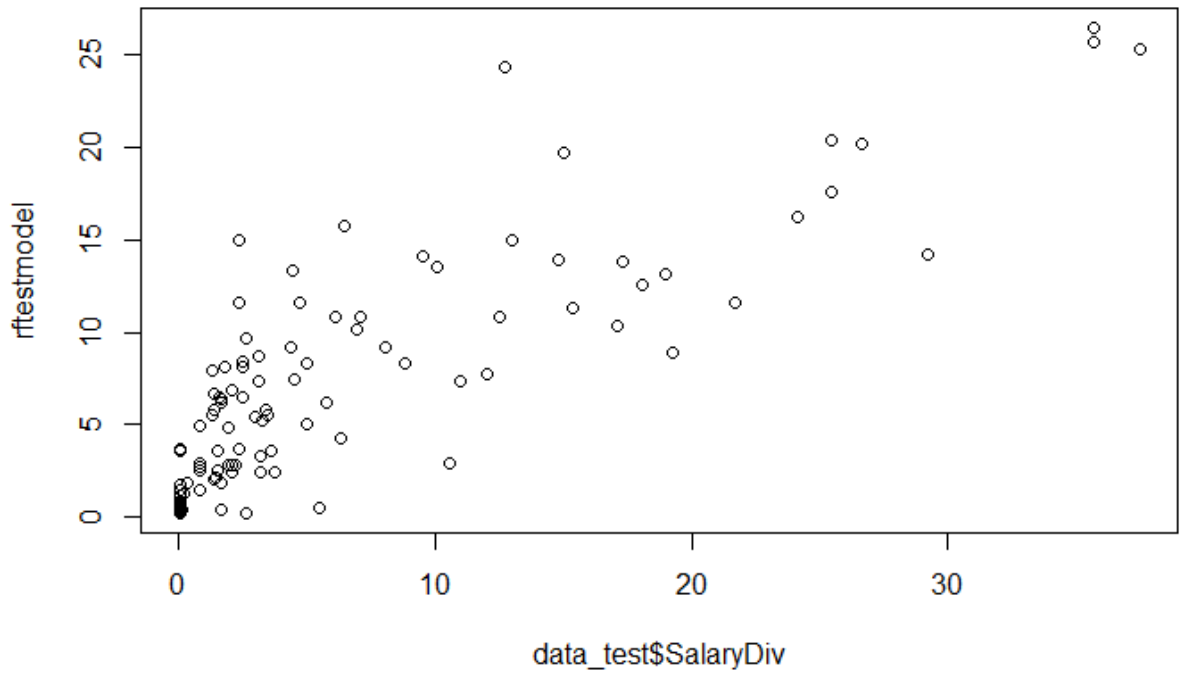


This graph shows the correlation between the actual vs predicted salary values from the decision tree and it came out to around 0.56. The RMSE came out to around 7.7 meaning a MSE of around 60, which is not great. Our opinion was that this meant that the trees did a reasonably good job of modeling salary, which is good because they are

highly explainable meaning that the model has predictive ability and can be understood well.

Random Forests

Above shows the results of a base random forest being run on the train data. We did a default random forest with 5 fold cross validation as to be consistent with earlier train/test splits. It had a great RMSE and solid R^2 values, with the best model having a RMSE of 5.35 or a MSE of around 28.5, and a R^2 of .508. Quite good results on the training set overall.

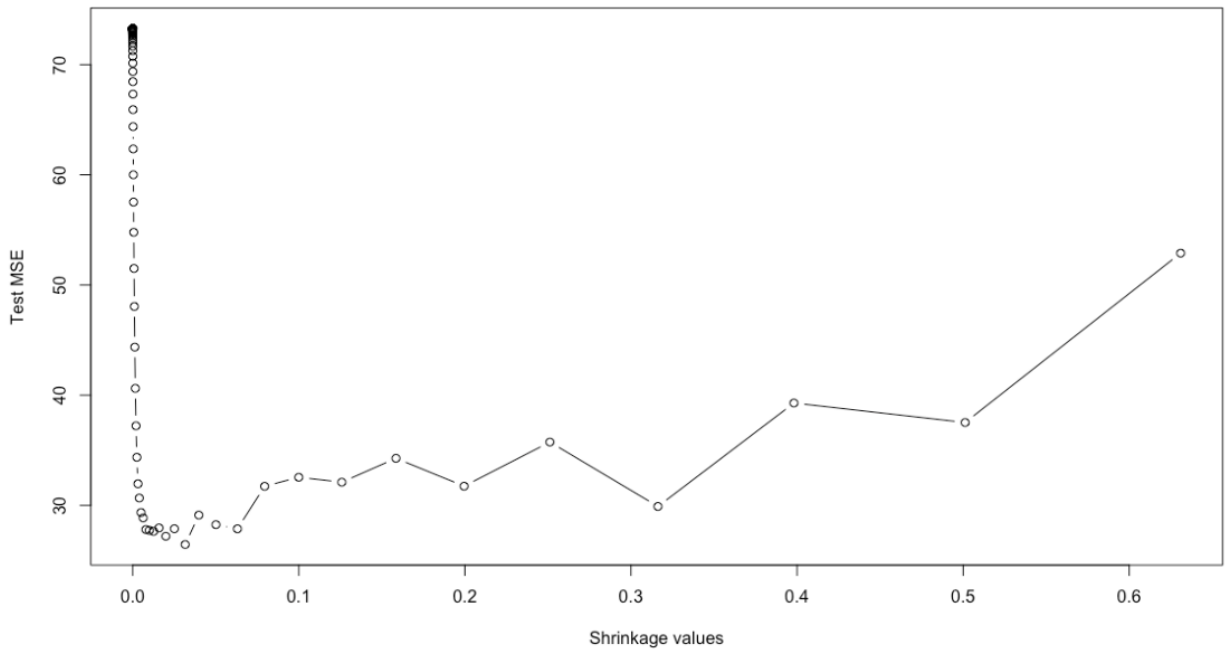


After running the random forest model on the test data set though we found the model to be extremely predictive of salary. The models predictions of salary had around a .835 correlation with actual salary as seen above, and bolstered by a great RMSE of 4.84 or

a MSE of 23.8 which is expected as random forests are a black box in terms of explainability, but most times outperform decision trees. Overall this model did an amazing job at predicting Salary.

Boosting

Boosting is a process in which trees are grown sequentially where each tree is grown using information from previously grown trees. Our boosting model used 1000 trees to predict player salary from all predictor variables. We calculated the shrinkage values in order to see which lambda value minimized the Test MSE. The plot below shows that the optimal lambda value is 0.032 and this yields a test MSE of 26.445. Boosting is also beneficial because it can tell us the relative influence of each variable in the dataset. We found that the variables with the most influence on player salary were Team, Age, Game Started, and Value Over Replacement Player (VORP). The most influential variables can be seen in the table below. In this case it appears that the most important variables in determining a player's salary are somewhat out of their control, Team and Age. This makes sense because certain teams have different salary restrictions and age is extremely important in regards to the player's prime point in their career.



	var <fctr>	rel.inf <dbl>
Tm	Tm	36.99189441
Age	Age	9.74493945
GS	GS	8.21003029
VORP	VORP	6.68165760
WS	WS	4.37785260
AST	AST	4.04727118
FTA	FTA	3.40609448
TOV	TOV	2.30979403
FT	FT	1.93248665
X2P	X2P	1.55082383
X2PA	X2PA	1.51041536
DRB	DRB	1.42740543
PF	PF	1.26786211
FT.	FT.	1.11644372
PTS	PTS	1.03644595
G	G	1.01711273
DWS	DWS	1.01696656
FGA	FGA	0.85676154
BLK.	BLK.	0.77491726
TRB.	TRB.	0.74207818
DRB.	DRB.	0.73161771
X3P.	X3P.	0.70677129
OWS	OWS	0.68636349
MP	MP	0.61752966
FG	FG	0.61240495
X3PAr	X3PAr	0.61233358

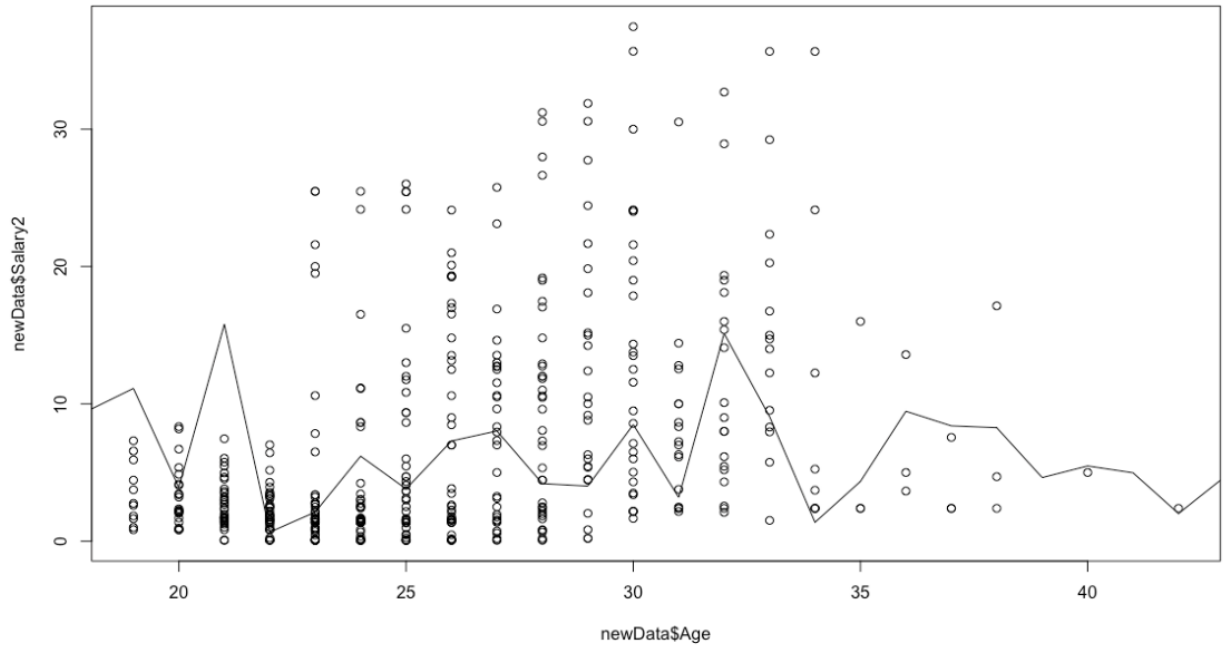
Support Vector Machines (SVM)

The goal of SVM is to establish hyperplanes that divide the subspace in order to develop better predictions for our regression analysis. To begin our SVM analysis, we started with the multiple linear regression model calculated through stepwise selection above. The new SVM model utilized eps-regression and radial kernels with a cost of 1, gamma value of 0.06667, and epsilon value of 0.1. There were 312 support vectors in this model. When the model was applied to the test data set, the Root Mean Squared Error(RMSE) was calculated to be 6.024. Squaring this yielded a Test MSE value of 36.289. In an attempt to improve on this, we utilized 10 fold cross validation to tune the parameters for our SVM model. We found that the best model contained an epsilon value of 0.4 and a cost of 4. This newly tuned model had 196 support vectors. However, this tuned model did not improve upon the original SVM model as the Test MSE value was 41.051, higher than the original MSE.

K-Nearest Neighbor (kNN)

The K Nearest Neighbor regression algorithm takes into account the closest training examples in the data set to estimate the value of the original point. The property value of the object, in this case, the player's salary, is taken as an average of the values of its k nearest neighbor points. For our kNN model, we used a k value of 10 and attempted to predict the players salary based on all of the numeric variables in the dataset. The resulting Test MSE was 50.611. The kNN model with k=10 is demonstrated below using

the Age variable.



Conclusion

The regression models developed in this paper have provided valuable information on the question of what factors impact NBA player salaries. While not the best model, the Multiple Linear Regression model tells us that about 55% of the variation of NBA player salary can be explained by player statistics. Clearly there are outside factors influencing salary, but it is good to know that player performance does have a large impact. The more advanced models developed in this paper also showed which predictors were deemed most important. When running the LASSO, Ridge, and Elastic Net regression models, the Elastic Net model performed the best. This Elastic Net model valued 3P% and the PF and SF positions. The Team variable also had a very large impact on player

salary. The Regression Trees and Boosting model also supported the importance of the player's team on their salary. There were several key splits in the regression tree involving player teams and the Team variable was also by far the most influential predictor in the Boosting regression. The other important variables displayed by boosting as most influential were Age, Games Started (GS), Value Over Replacement Player (VORP) and Win Shares (WS). VORP and WS are both relatively newer and more advanced statistics when it comes to measuring player performance. Overall, in terms of interpretability, the Linear Regression model is superior because its impact can be easily explained. In this case, the linear model also did a good job predicting salary as it explained over half of the variance in player salary. That being said, the random forests model was superior to others in predictive performance, but the black box interpretability makes it less useful in the context of players and teams wanting to find out what stats are important in determining salary. Through our research, it is clear that there are other impacts on player salary beyond player performance on the court. It would be interesting for further studies to investigate off the court variables such as a player's social media status and popularity, as well as economical variables such as the current market and salary cap information that is constantly changing every year.

References

https://hoopshype.com/salaries/players/2018-2019/

https://www.basketball-reference.com/leagues/NBA_2019_advanced.html

https://rstudio-pubs-static.s3.amazonaws.com/371407_e21330910f3c4bd2b6e19440013ea793.html

https://towardsdatascience.com/nba-salary-predictions-4cd09931eb55

https://medium.com/analytics-vidhya/nba-player-salaries-prediction-with-linear-regression-2b90280ff4e8