Predicting NBA Player Salary and Positions

Arsh Irfan Modak
Data Science, M.S.
Khoury College of Computer Sciences
modak.a@husky.neu.edu

Omkar Shivaji Waghmare
Data Science, M.S.
Khoury College of Computer Sciences
waghmare.o@husky.neu.edu

Jay Majithia
Data Analytics Engineering, M.S.
College of Engineering
majithia.j@husky.neu.edu

Sahar Shady
Bioinformatics, M.S,
College of Science,
shady.s@husky.neu.edu

***Abstract* — NBA is the men's professional basketball league in North America, composed of 30 teams that compete with each other to win the Larry O'Brien Championship Trophy – previously known as Walter A. Brown Trophy. Our data set consists of data from 1996 to 2020, for all seasons and all 30 teams. A team has a maximum of 12 players. We have scraped three data sets according to three different positions of a player: Forward, Centre and Guard.**

We aim to find relations between various player stats and use them to predict which of the three positions a player is best suitable to play in. Furthermore, we also want to predict the salary of each player according to their performance in the games over time.

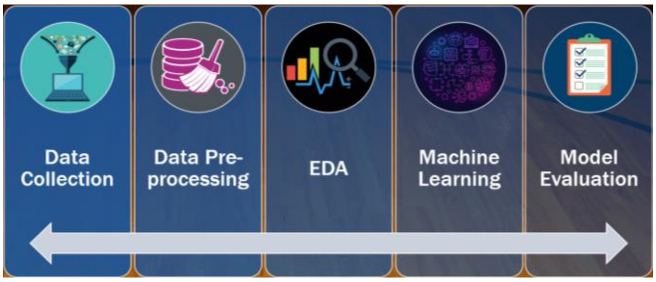
I. Introduction

The overall data available for sports analytics have been pretty basic up until the 1990's, with the only information available being a team's points, assists and number of rebounds. Then with the emergence of technological advancements like video tracking players on the field, an abundance of data has been available for sports predictive analytics. This data has proven to play an important role in bettering the overall performance of players and teams. In 2013-14 when the NBA started investing heavily in video tracking all its teams during practices as well as matches, players and teams' achievements were noticeably noted.

Statistics have shown that one of the top improvements achieved was a team's three-pointer goals made (3PM) per game. It has seen a 46% increase. "The average three-point shot taken in 2012 was about 18.4 three-point shots per game, and in 2017 the average team took about 27 shots per game"[1]. Hence, players are always encouraged to aim for a three-pointer attempt (3PA).

After tidying and transforming our datasets, we performed Exploratory Data Analysis to visualize the relationships between the various features we are interested in. We explored the change in various features over time, how those features played a role in the change of salaries and how each position affects the change in those features. Our results complied with what research had suggested of an overall advancement in performance due to the abundance of data available and the implementation of predictive analytics. Our dataset consisted of both, players and teams' statistics, but since a team's statistics is built on individual players statistics, we only included the latter in fitting our machine learning models (to avoid linear dependencies in our models).

II. The Flow



III. Methods

In this section all the technical methods used to complete the flow mentioned above will be explained with respect to the project uploaded on GitHub.

A. *Data Collection and Integration:*

The following files are responsible for data collection and integration:

- *Codes/Scarping/**
- *Codes/data_integration_and_cleaning.rmd*

<u>Codes/Scraping/*:</u> This file contains all the scraping code. A total of six scripts were written to get the complete dataset. Two main websites were scraped [2][3]. A total of 10 datasets are generated.

<u>Codes/data_integration_and_cleaning.rmd:</u> In this file, we have joined all the scraped datasets, and creates a final csv file. Over time team names have changed, so that's why we manually updated the team names so as to follow the latest 30 teams. Finally, our data was ready for preprocessing.

B. *Data Preprocesing:*

The following files are responsible for data preprocessing:

- *Codes/final_preprocessing.Rmd*

We start off by Dropping near zero variance variables. Then we impute the missing data. We used two methods for imputing missing data, first one was k-Nearest Neighbors and the second one was miss-forest. Results generated by miss-forest were much closed and accurate to the actual values. Hence, we decided to go with that.

After this, we identified variables with a co-relation higher than 60% with respect to the target variables. Finally, we identify variables with linear combinations and removed overlapping data. (overlapping data: players that have played at more than one position in a single year for the same team)

C. *Exploratory Data Analysis:*

The following files are responsible for data collection and integration:

- *Codes/final_EDA.rmd*

1. EDA w.r.t Position:
   We created a function that would generate animating (year-wise) graphs with x axis as Positions and y axis as rest of the stats. You can access these animated plots in our PPT and on GitHub. In order to understand the different positions in basketball and their responsibilities, please refer to the diagram below:

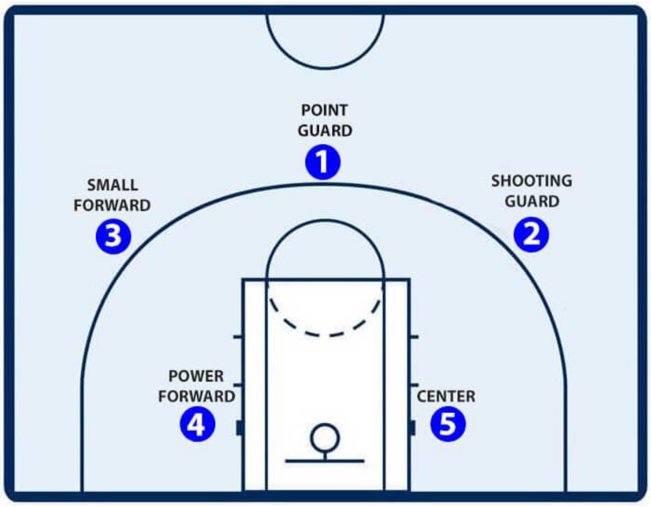


Fig 1.1: Positions in Basketball

Center:
The tallest player in the team, responsible for collecting rebounds and making missed shots count.

Forward:

i) Small Forward: Responsible for shooting three pointers and at times driving in the zone.

ii) Power Forward: Responsible for covering small forward and collecting rebounds.

Guard:

i) Point Guard: Responsible for getting the ball in the opponent's court and rotating the ball among players.

ii) Shooting Guard: Responsible for shooting three pointers.

All the position related animations can be better understood after referring to the above chart.

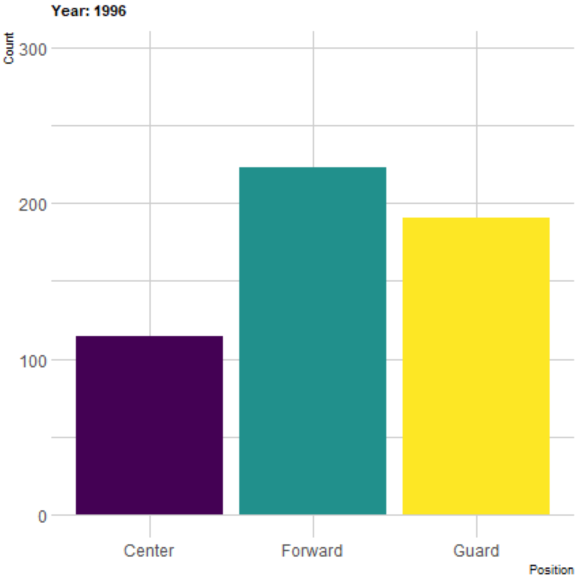


Fig 1.2: Distribution of Position

Inference: We realized there is a class imbalance, hence, we decided to apply some sampling techniques.

2. EDA w.r.t Player Stats:

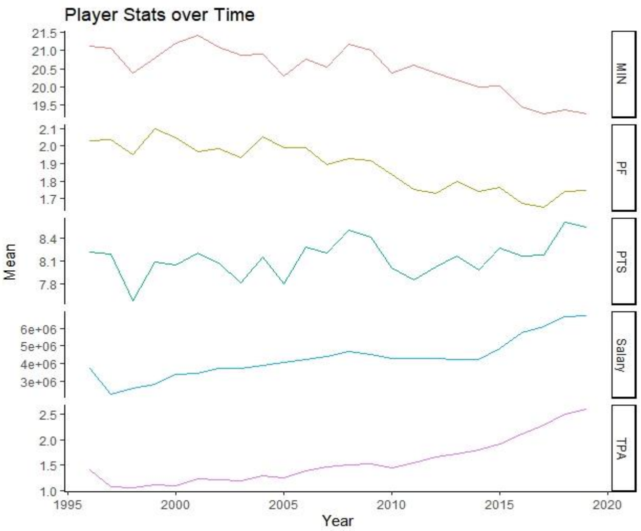


Fig 2.1: Player Stats over Time

Inference: From the graph we can conclude that the players have become more efficient over the years.
This is depicted in the increase in Points Scored and Three Pointers and decrease in Minutes Played over time.
The rate of Personal Fouls has also decreased.
The Salary of the players also sees an increasing trend. Two factors play an important role here: Inflation and better performance of players. (Better players are paid more)

Next, we wanted to explore the relationship between types of points and their contribution to total points. Furthermore, we wanted to see the relationship between points and salary.

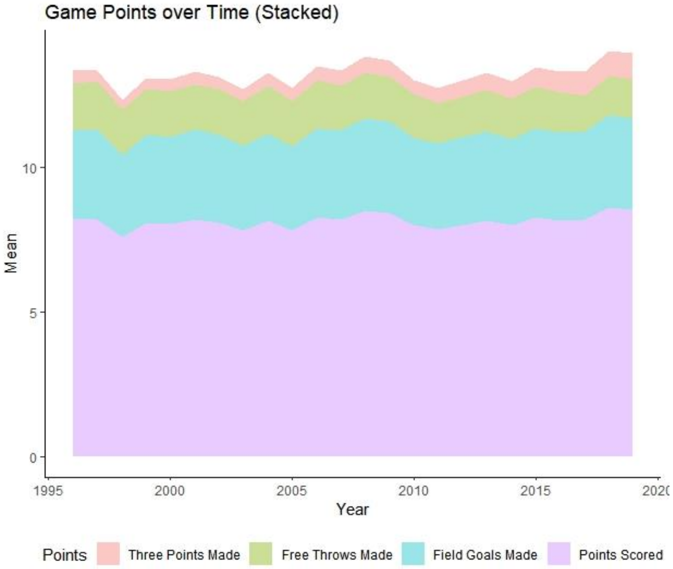


Fig 2.2: Game Points over Time

Inference: As depicted by the graph, we can see the contribution of different types of points in overall Points. Field Goals Made contribute the most towards Total Points Scored; Next are Free Throws; Three Pointers contribute the least.

D. Machine Learning: Salary

The following files are responsible for Machine Learning w.r.t Salary:

- *Codes/ML_Salary.rmd*

After the pre-processing of our data, we decided to drop all team related stats and those variables which did not satisfy the correlation threshold of 60%.

This resulted in our final dataset with 16 variables and 13,416 rows.

We implemented stepwise model selection with 5-fold cross validation and 5-fold resampling of our data to avoid overfitting. However, before we did that, we had to transform our variables such that the predictor variables showed a linear relationship with Salary. The transformation was fairly consistent with Salary being transformed to log(Salary) and the rest of the variables were transformed to their appropriate square root.

Once we got our results, we found out the best RMSE was with seven out of the 16 variables as shown by the graph below.

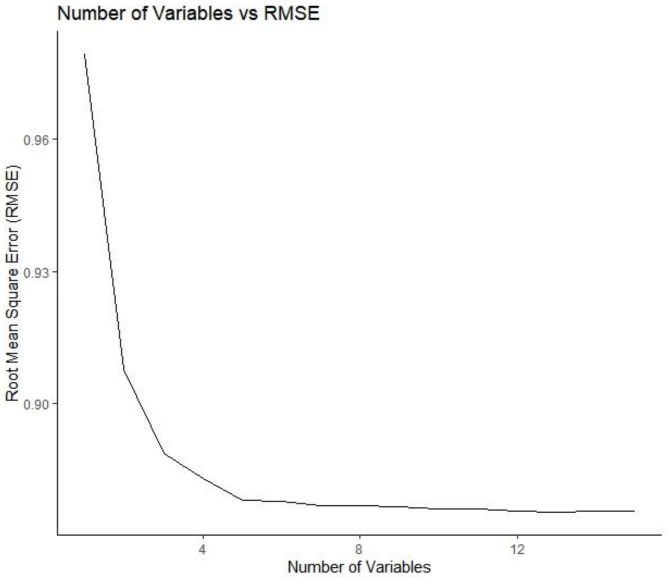


Fig 3.1: Stepwise Model RMSE

Plots that depict a linear relationship between Salary and the predictor variables [appendix 1]

E. Unsupervised Machine Learning: Position

The following files are responsible for Unsupervised Machine Learning w.r.t Position:

- *Codes/ML_Position.rmd*

Initially we thought an unsupervised approach would give us better results than that of a supervised approach, thus we implemented K-Means Clustering. As discussed above, we have a case of class imbalance, so we ran k-means for a total of six iteration, original data, up sampled data, down sampled data for three and five clusters.

We could identify the clusters clearly, however with some overlapping data, which was not surprising since a player could play multiple positions in a game, or in different games of the same season and so forth.

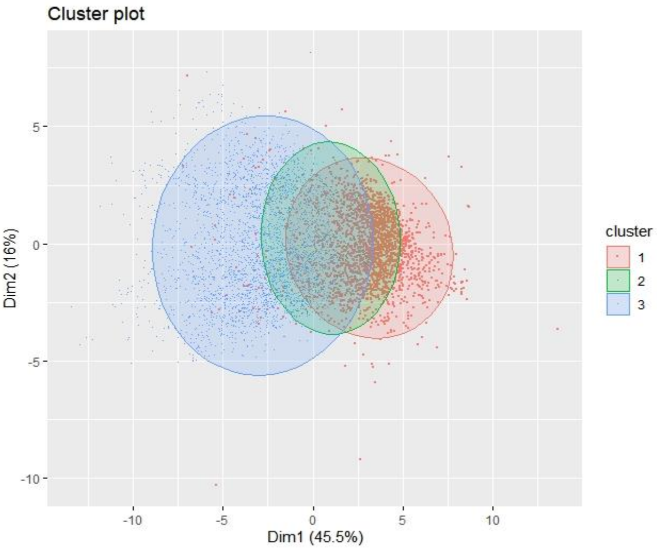


Fig 4.1: K-Means Clustering with 3 clusters

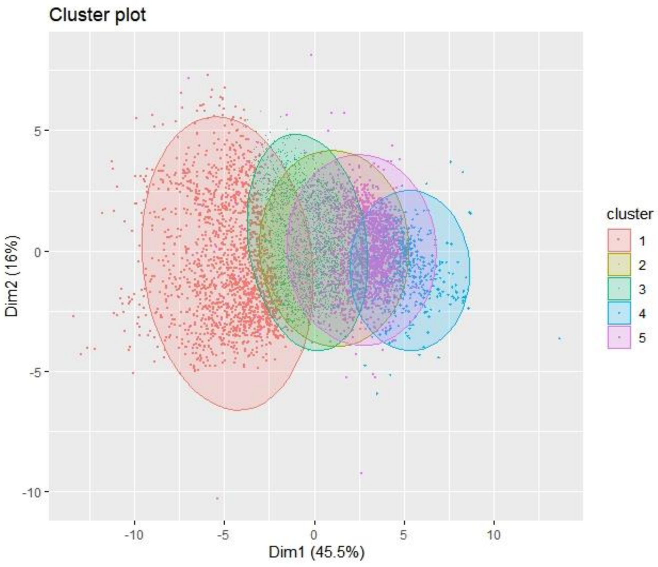


Fig 4.2: K-Means Clustering with 5 clusters.

We were not satisfied with the results we got, hence we decided to give Supervised Machine Learning a shot.

F. *Supervised Machine Learning: Position*

The following files are responsible for Supervised Machine Learning w.r.t Position:

- *Codes/ML_Position.rmd*

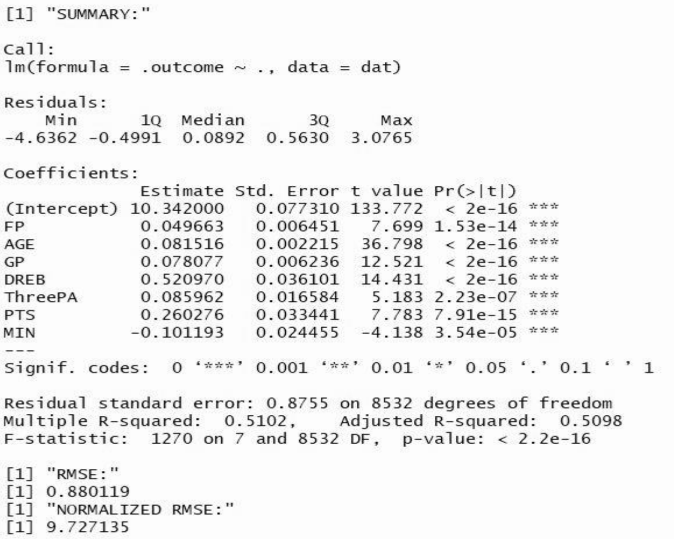
Keeping in mind the data imbalance, we ran three iterations of the following Machine Learning Models to predict Position:

i) Multinomial Logistic Regression
ii) Decision Trees
iii) Support Vector Machines
iv) Neural Nets
v) Extreme Gradient Boosting

We were able to generate ML models that justified their behavior. The results, comparison and behavior of the models will be discussed in the "Results" Section.

IV. RESULTS

A. *Machine Learning Results: Salary*



```
[1] "SUMMARY:"

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.6362 -0.4991  0.0892  0.5630  3.0765 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10.342000   0.077310 133.772  < 2e-16 ***
FP           0.049663   0.006451   7.699 1.53e-14 ***
AGE          0.081516   0.002215  36.798  < 2e-16 ***
GP           0.078077   0.006236  12.521  < 2e-16 ***
DREB         0.520970   0.036101  14.431  < 2e-16 ***
ThreePA      0.085962   0.016584   5.183 2.23e-07 ***
PTS          0.260276   0.033441   7.783 7.91e-15 ***
MIN         -0.101193   0.024455  -4.138 3.54e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8755 on 8532 degrees of freedom
Multiple R-squared:  0.5102,	Adjusted R-squared:  0.5098 
F-statistic:  1270 on 7 and 8532 DF,  p-value: < 2.2e-16

[1] "RMSE:"
[1] 0.880119
[1] "NORMALIZED RMSE:"
[1] 9.727135
```

Fig 5.1: Linear Regression to predict Salary

The scale of log(Salary) is 8.52 to 17.64. The Normalized RMSE lies between this range and is used to compare different linear models of the same scale.

Residual Plots of the Linear Model [appendix 2]

B. *Unsupervised Machine Learning Results: Position*

TABLE I. K-MEANS CLUSTERING RESULTS

No. of Clusters	***Percentage***
3	67.5
3 (down sampled)	67.9
3 (up sampled)	68.7
5	76.4
5 (down sampled)	77.0
5 (up sampled)	77.7

Initially the data that we had consisted of many overlapping values, hence, when we fed this data to the k – means clustering algorithm, it wasn't able to distinguish (categorize) variables with the same predictor variables but a different category of the response variable i.e. Position. For example, a player can have the same stats but a different Position in a different row.

To tackle this problem, we decided to create a function which gets rid of such duplicates. Now, we have around 11500 rows which is about 2500 duplicate values (which are actually not, since position is different).

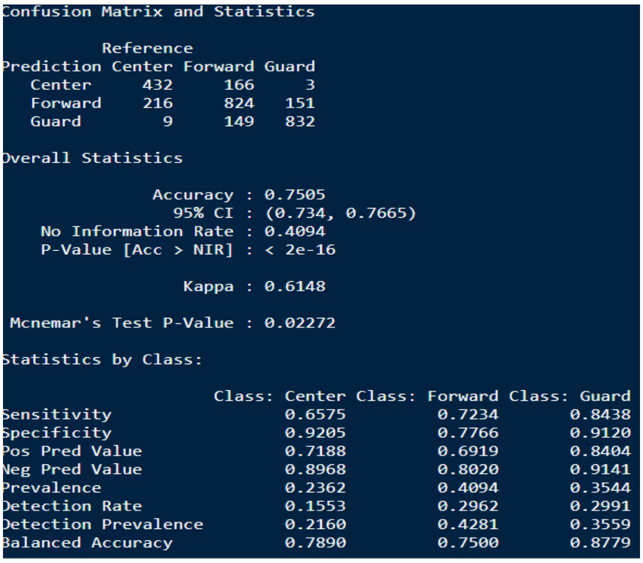
The aforementioned results above are generated when we use this newly created data which is specifically for this algorithm.

C. *Supervised Machine Learning Results: Position*

Comparison of Accuracies of Various Models [6]

The outputs below are of the five best models that we were able to train:

1) *Multinomial Logistic Regression:*



```
Confusion Matrix and Statistics

          Reference
Prediction Center Forward Guard
   Center     432     166     3
   Forward    216     824   151
   Guard        9     149   832

Overall Statistics
                                          
               Accuracy : 0.7505          
                 95% CI : (0.734, 0.7665) 
    No Information Rate : 0.4094          
    P-Value [Acc > NIR] : < 2e-16         
                                          
                  Kappa : 0.6148          
                                          
 Mcnemar's Test P-Value : 0.02272         

Statistics by Class:

                     Class: Center Class: Forward Class: Guard
Sensitivity                 0.6575         0.7234       0.8438
Specificity                 0.9205         0.7766       0.9120
Pos Pred Value              0.7188         0.6919       0.8404
Neg Pred Value              0.8968         0.8020       0.9141
Prevalence                  0.2362         0.4094       0.3544
Detection Rate              0.1553         0.2962       0.2991
Detection Prevalence        0.2160         0.4281       0.3559
Balanced Accuracy           0.7890         0.7500       0.8779
```

Fig 5.2: Multinomial Logistic Regression

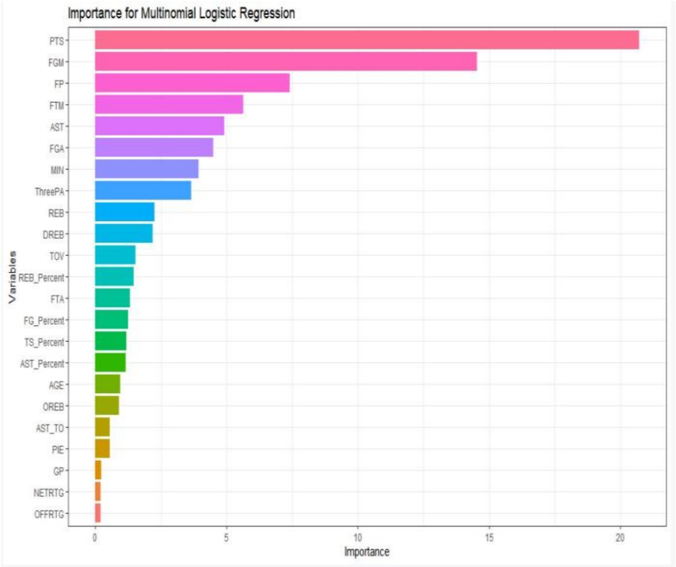
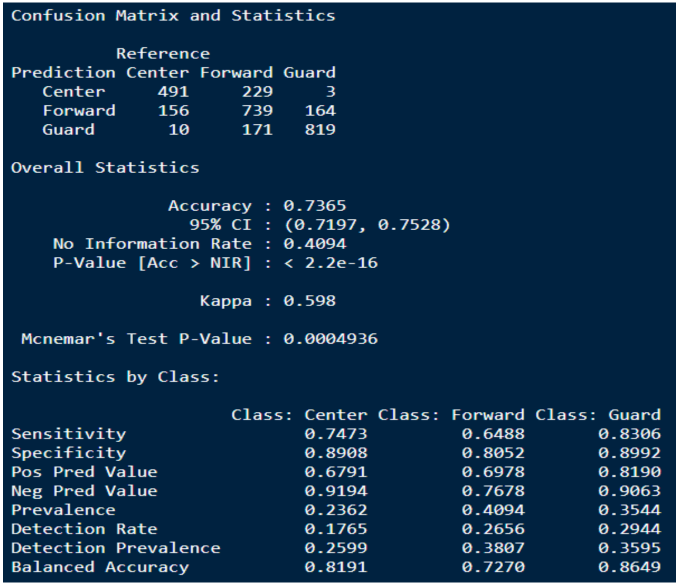


Fig 5.3: Important Features for Multinomial Logistic Regression

Inference: This model gives the most importance to point based features (PTS, FGM, FP etc). According to our EDA, guard and forward have scored the highest points and hence our model has a good sensitivity to those positions.

2) *Decision Trees*



```
Confusion Matrix and Statistics

             Reference
Prediction Center Forward Guard
   Center     491     229     3
   Forward    156     739   164
   Guard       10     171   819

Overall Statistics

               Accuracy : 0.7365
                 95% CI : (0.7197, 0.7528)
    No Information Rate : 0.4094
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.598

 Mcnemar's Test P-Value : 0.0004936

Statistics by Class:

                     Class: Center Class: Forward Class: Guard
Sensitivity                 0.7473         0.6488       0.8306
Specificity                 0.8908         0.8052       0.8992
Pos Pred Value              0.6791         0.6978       0.8190
Neg Pred Value              0.9194         0.7678       0.9063
Prevalence                  0.2362         0.4094       0.3544
Detection Rate              0.1765         0.2656       0.2944
Detection Prevalence        0.2599         0.3807       0.3595
Balanced Accuracy           0.8191         0.7270       0.8649
```

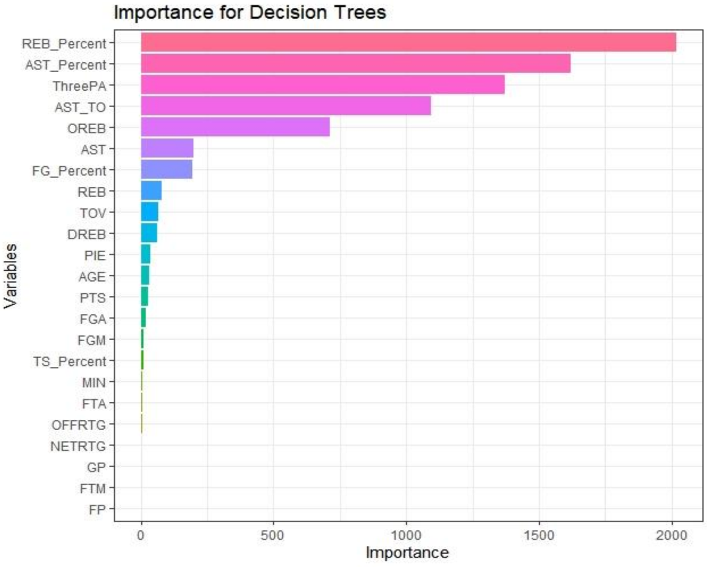
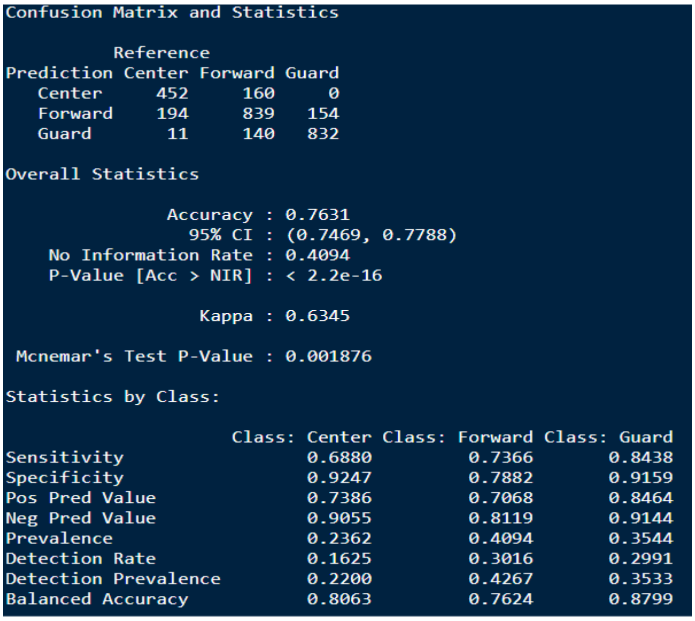


Fig 5.4:Model Output and Important Features for Decision Trees

Inference: This model gives the most importance to features that can be classified as “Support” which are traits of a player playing as a Center, hence we see an increase in sensitivity towards Center, a drastic decrease for Forward and minimal change for Guard.

3) *Support Vector Machines*



```
Confusion Matrix and Statistics

             Reference
Prediction Center Forward Guard
   Center     452     160     0
   Forward    194     839   154
   Guard       11     140   832

Overall Statistics

               Accuracy : 0.7631
                 95% CI : (0.7469, 0.7788)
    No Information Rate : 0.4094
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6345

 Mcnemar's Test P-Value : 0.001876

Statistics by Class:

                     Class: Center Class: Forward Class: Guard
Sensitivity                 0.6880         0.7366       0.8438
Specificity                 0.9247         0.7882       0.9159
Pos Pred Value              0.7386         0.7068       0.8464
Neg Pred Value              0.9055         0.8119       0.9144
Prevalence                  0.2362         0.4094       0.3544
Detection Rate              0.1625         0.3016       0.2991
Detection Prevalence        0.2200         0.4267       0.3533
Balanced Accuracy           0.8063         0.7624       0.8799
```

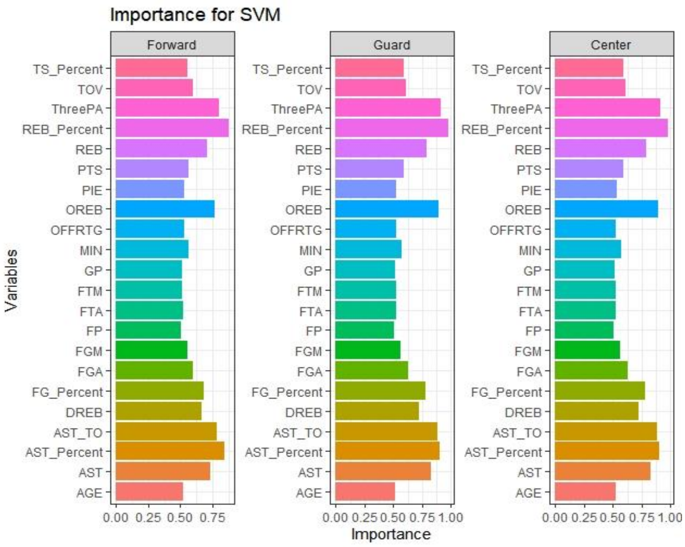
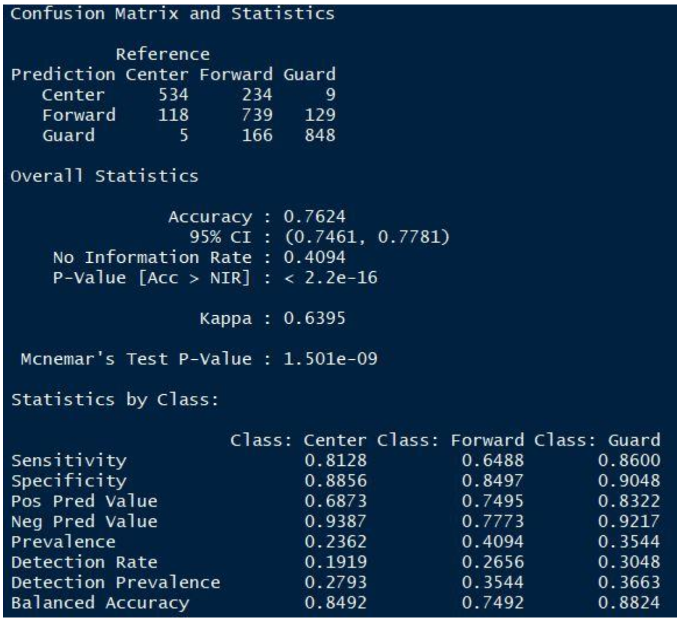


Fig 5.5: Model Output and Important Features for Support Vector Machines

Inference: It behaves similar to the Logistic Regression Model; however, it seems like it gives equal importance to support type features, hence giving a better overall accuracy and sensitivity of each category.

4) *Neural Nets*



```
Confusion Matrix and Statistics

          Reference
Prediction Center Forward Guard
   Center     534     234     9
   Forward    118     739   129
   Guard        5     166   848

Overall Statistics

               Accuracy : 0.7624
                 95% CI : (0.7461, 0.7781)
    No Information Rate : 0.4094
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6395

 Mcnemar's Test P-Value : 1.501e-09

Statistics by Class:

                     Class: Center Class: Forward Class: Guard
Sensitivity                 0.8128         0.6488       0.8600
Specificity                 0.8856         0.8497       0.9048
Pos Pred Value              0.6873         0.7495       0.8322
Neg Pred Value              0.9387         0.7773       0.9217
Prevalence                  0.2362         0.4094       0.3544
Detection Rate              0.1919         0.2656       0.3048
Detection Prevalence        0.2793         0.3544       0.3663
Balanced Accuracy           0.8492         0.7492       0.8824
```

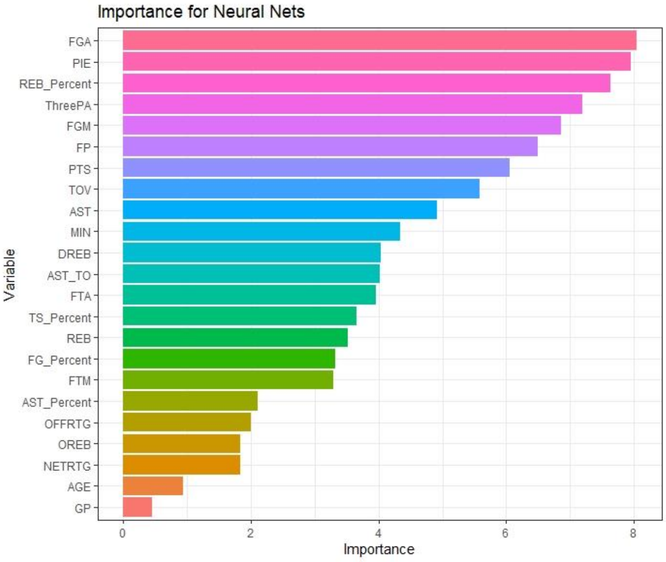
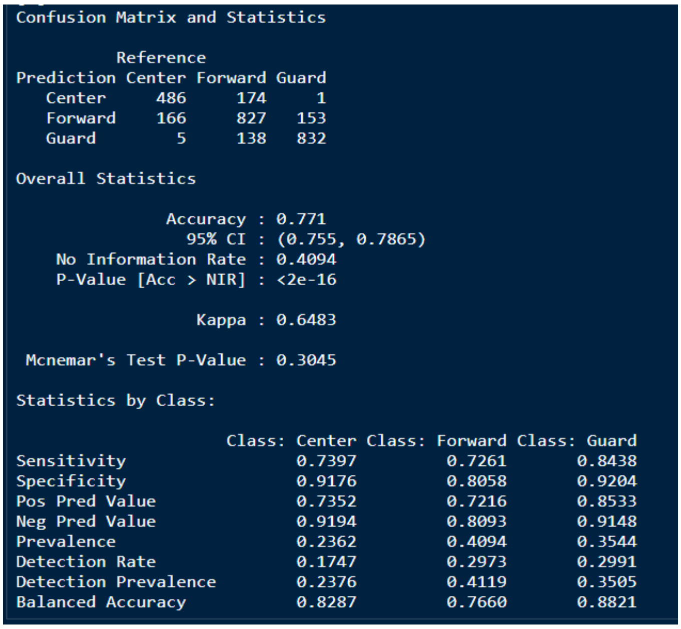


Fig 5.6: Model Output and Important Features for Neural Network.

Inference: From the aforementioned plot, we can see that this model gives more importance to a higher number of features as compared to the previous models. Hence, we can see that its able to classify Center and Guard really well, but Forward, not so much! This may be because Forward is a versatile player and possesses skills which are common to both, center and guard, hence is more difficult to classify and may be misclassified as the other two positions.

5) *Extreme Gradient Boosting*



```
Confusion Matrix and Statistics

          Reference
Prediction Center Forward Guard
   Center     486     174     1
   Forward    166     827   153
   Guard        5     138   832

Overall Statistics

               Accuracy : 0.771
                 95% CI : (0.755, 0.7865)
    No Information Rate : 0.4094
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6483

 Mcnemar's Test P-Value : 0.3045

Statistics by Class:

                     Class: Center Class: Forward Class: Guard
Sensitivity                 0.7397         0.7261       0.8438
Specificity                 0.9176         0.8058       0.9204
Pos Pred Value              0.7352         0.7216       0.8533
Neg Pred Value              0.9194         0.8093       0.9148
Prevalence                  0.2362         0.4094       0.3544
Detection Rate              0.1747         0.2973       0.2991
Detection Prevalence        0.2376         0.4119       0.3505
Balanced Accuracy           0.8287         0.7660       0.8821
```

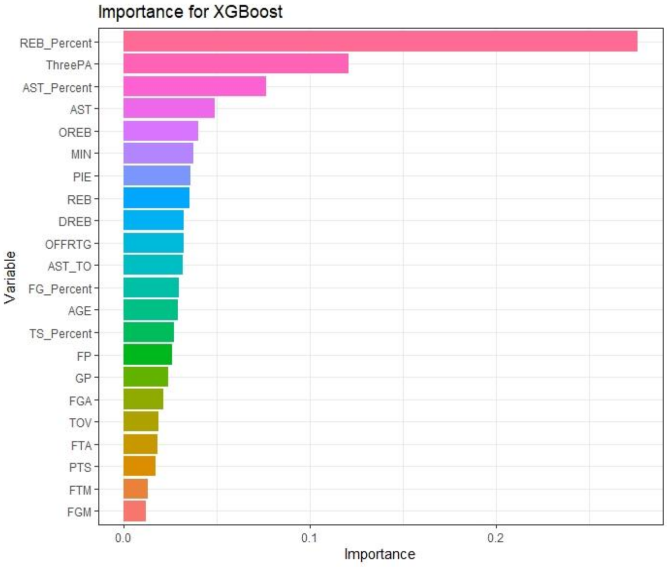


Fig 5.7: Model Output and Important Features for Neural Network.

Inference: This model gives the most balanced sensitivity of all three classes, furthermore it also makes sense that this model gives the same importance as that of Decision Trees. This model also gives the best accuracy.

V. DISCUSSION

We can see clear relationships between the player stats and position. Our models work as expected. We identified key stats that differentiate the three categories of position.

In further works of this project, we would like to consider various other advance stats and pin point such features that would improve the unsupervised approach.

We also plan on using various algorithms to extract the most correlated variables (such as the Boruta Package and PCA).
In the next iteration of this project, we plan on using Supervised Machine Learning Methods to classify five different positions rather than three.

VI. STATEMENT OF CONTRIBUTION

It was a joint contribution by all the team members. Everyone was involved in idea development, EDA, education around models, and general discussion of approach.

1. *Sahar and Jay:* Helped in performing the Exploratory Data Analysis and for designing the regression model for the prediction of Salary of a player.
2. *Arsh and Omkar:* helped in performing the EDA specifically for the positions and running the different supervised and unsupervised machine learning models for the prediction of the Position of a player.

Equal contribution and efforts by all team members in preparing the presentation and final report.

VII. REFERENCES

[1] https://medium.com/@randerson112358/how-the-nba-uses-data-analytics-6eac3c43a096

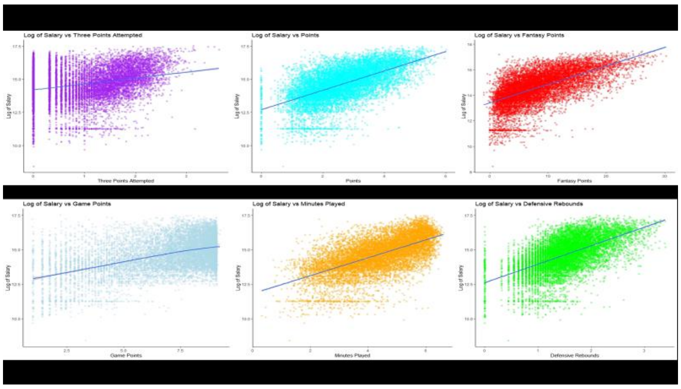
[2] https://stats.nba.com/players/

[3] https://hoopshype.com/salaries/players/

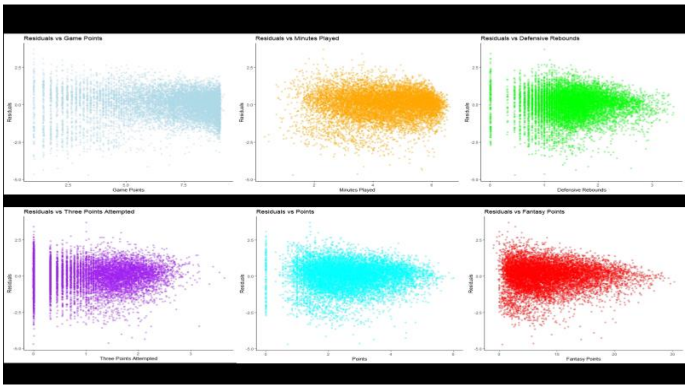
https://www.basketball-reference.com/

VIII. APPENDIX

[1]



[2]



GitHub Link: https://github.com/arshmodak/Predicting-NBA-Player-Salaries-and-Position-using-Machine-Learning