Classifying NBA Players by Position

STAT471: Modern Data Mining
Final Project
Jackson Joffe, Jake Iyoob, Jane Huang
May 2, 2021

Contents

I. Executive Summary

Until recently, NBA positions were extremely static. Historically, positions have been exclusively decided by size. However, in the early 2010s, positionless teams emerged as viable lineups that teams could use. Today, player position is more ambiguous than ever, and this study attempts to investigate how player position relates to certain NBA statistics to better understand the role of each position. In our project for STAT471: Modern Data Mining, we examine the differences in positions for NBA players. The goal of this study is to create, tune, and compare positional classifiers for NBA players. Ideally, this algorithm would be useful for coaches, general managers, analysts, and financially savvy players.

In this study, we first introduce our topic and describe how we obtained our data. Using the nbastatR package[1], we scraped NBA statistics from Basketball Reference[2] from 1959 to 2019. We identified 14 features we wanted to examine in our study. They were: ptsTotals, pfTotals, astTotals, blkTotals, trbTotals, stlTotals, fg3aTotals, pctFT, ftaTotals, ratioWS, ratioOWS, ratioDWS, orbTotals, minTotals. After EDA revealed some worrisome concerns about multicollinearity, we removed some variables and settled on six final features: astTotals, blkTotals, trbTotals, stlTotals, fg3aTotals, and pctFT. Next, we split our observations into four different time periods, and we train and test different multiclass classifiers on each of the four time periods (Era 1: 1959-1974, Era 2: 1975-1989, Era 3: 1990-2004, Era 4: 2005-2019) as well as the full 60-year period (1959-2019).

After spending time tuning appropriate parameters for each model, we made predictions on the test set. We compared the tuned models we created, and we found that in general, the random forest multiclass classification model performed the best. Tree-based models performed worse in general, while the gradient boosting classification model had high variance in testing accuracy. The classifier valued assists and rebounds highly during all periods, and it began to give importance to steals, blocks, and three-pointers as they were gradually recorded throughout time. We finally settled on a final model for each era and the entire period after hours of runtime.

The final section of this study delves into our study's conclusions, limitations, and potential future directions. We suggest other multiclass classifiers that might be worth trying, and we also note that many potentially useful advanced NBA statistics were not recorded during most of NBA history.

[1] http://asbcllc.com/nbastatR/reference/index.html
[2] https://www.basketball-reference.com/

II. Introduction

A. Background and Study Goals

Why would one need to use machine learning to classify NBA players by position? A casual basketball fan would likely have watched enough basketball to identify which player plays which position. Even a person with no knowledge of the game could find some YouTube clips of the all time greats to better understand what each position does. So what extra information might a machine learning algorithm offer us that would justify its use?

When it comes to the task of classification, machine learning algorithms are using a predicted probability to decide whether or not something belongs to a group. In the case of an NBA positions, a machine learning algorithm is going to predict the probability that the player is either a guard, forward, or center. The algorithm will give us a probability for each position for any given player. It's these probabilities that could prove to be useful to a front office or discussion among NBA analysts.

Let's also consider another use for these positional probabilities. Each year a panel of sportswriters and broadcasters vote for various awards like Most Valuable Player, Defensive Player of the Year, and the All-NBA teams. Players are eligible for a supermax contract (30% of a team's salary cap) if they meet the following criteria:

- Make an All-NBA team in the previous season or in 2 of the 3 previous seasons
- Be named Defensive Player of the year in the previous season or in 2 of the 3 previous seasons
- Be named Most Valuable Player in at least one of the three previous seasons

Because these media members' votes can affect the future earnings of players, it is paramount that we have equitable metrics for evaluating and comparing players. The inputs to an NBA positional classifier algorithm would be some group of recorded statistics that describe a player's performance and impact on a game. The algorithm would then take those inputs, perform computations, and provide us with the positional probabilities that we would use for classification.

The goal of this study is to create, tune, and compare positional classifiers for NBA players. We split the observations we have into four different time periods, allowing us to analyze the algorithm's effectiveness historically. Ideally, this algorithm would be useful for coaches, general managers, analysts, and financially savvy players.

B. Data Summary

Using the nbastatR package[3], we scraped NBA statistics from Basketball Reference[4] from 1959 to 2019, excluding the COVID-19-shortened 2020 season from our analysis. This full dataset includes 20,281 total observations each with 67 features. To analyze the algorithm over time, we divided the 60-year period into four 15-year periods: 1959-1974, 1975-1989, 1990-2004, and 2005-2019. We also used the entire 60-year period for parts of our analysis. For our analysis, we limited predictors to 14 features, choosing `groupPosition`, which classifies player position as guard (G), forward (F), or center (C), as our response variable. Though we could have chosen to classify player position by slugPosition — point guard, shooting guard, small forward, power forward, and center — we chose to simplify our

[3] http://asbcllc.com/nbastatR/reference/index.html
[4] https://www.basketball-reference.com/

analysis by considering both point and shooting guards as guards and both small and power forwards as forwards.

C. Overview of Analysis

Exploratory data analysis (EDA) revealed a number of remarks, including that centers block significantly more shots on average than forwards and guards while corralling a larger share of rebounds, but guards score more points, dish more assists, and net more steals than centers and forwards by a significant amount.

We began the main analysis by splitting the cleaned dataset (filtered for the 14 features, the response, and player name for interpretability) into four smaller datasets. Each dataset was then assigned to the 15-year periods mentioned above. We then split each dataset into training data to generate models to identify NBA statistics associated with and predictive of NBA player position (80%) and testing data to evaluate these models (20%).

The three methods that we employed to generate these models are as follows: 1) Tree-based Methods, 2) Random Forest Multiclass Classification, 3) Gradient Boosting Multiclass Classification. For each of these methods, we tune the models where appropriate. Following model generation, we evaluated the three models using the testing mean-squared error (MSE), and we selected the model with the lowest testing mean-squared error in conclusion.

III. Data Description and Explanation

A. Data Cleaning and Preparation

The data we used contains information for four different eras each spanning 15 years, discluding 2020, given the shortened season: from NBA games between 1959 and 1974 (2,333 observations), from NBA games between 1975 and 1989 (4,531 observations), from NBA games between 1990 and 2004 (6,294 observations), and from NBA games between 2005 and 2019 (7,123 observations). We obtained these by calling a function from the imported package and specifying the seasons we want to select. Below, we tidy the data, removing unnecessary columns and renaming features to help interpretability. These features include idPlayerNBA, urlPlayerThumbnail, urlPlayerHeadshot, for example. The full list can be found in our R code. The total observation count for all of our data is 20,281.

The features we selected for analysis initially were as follows:

1. ptsTotals - Total points scored by a player in a season
2. pfTotals - Total personal fouls by a player in a season
3. astTotals - Total assists by a player in a season
4. blkTotals - Total blocks by a player in a season *blocks introduced as an NBA stat in 1973-74
5. trbTotals - Total rebounds (offensive and defensive) by a player in a season
6. stlTotals - Total steals by a player in a season *steals introduced as an NBA stat in 1973-74
7. fg3aTotals - Total three point field-goal attempts by a player in a season *the NBA introduced the three-pointer in 1979-1980
8. pctFT - Free Throw Percentage; the formula is FT / FTA.
9. ftaTotals - Total free throws attempted by a player in a season

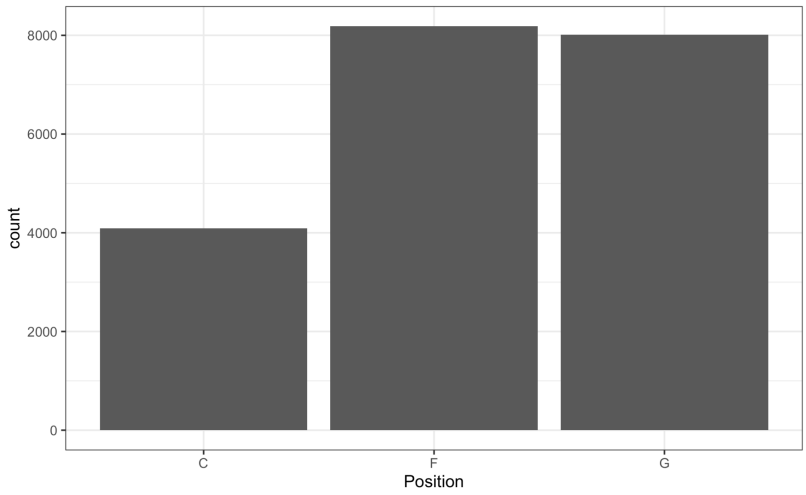
10. ratioWS - Win Shares is a player statistic which attempts to divide credit for team success to the individuals on the team. It is calculated using player, team and league-wide statistics and the sum of player win shares on a given team will be roughly equal to that team's win total for the season
11. ratioOWS - Similar to Win Shares but for offensive possessions
12. ratioDWS - Similar to Win Shares but for defensive possessions
13. orbTotals - Total number of offensive rebounds by a player in a season *offensive rebounds tracked beginning in 1973-74
14. minTotals - Total number of minutes played by a player in a season

The above fourteen features, in addition to player and year which were kept as indices, were the final variables we selected for our analysis at this stage; later, though, we will remove certain variables to avoid multicollinearity concerns. Before moving forward, we note that our selected NBA statistics dataset is complete, with no missing values. We use slugPosition as our response; this variable classifies position as guard (G), forward (F), and center (C). Lastly, we finalized our four datasets for each of the four eras to prepare for exploratory analysis.

B. Exploratory Data Analysis

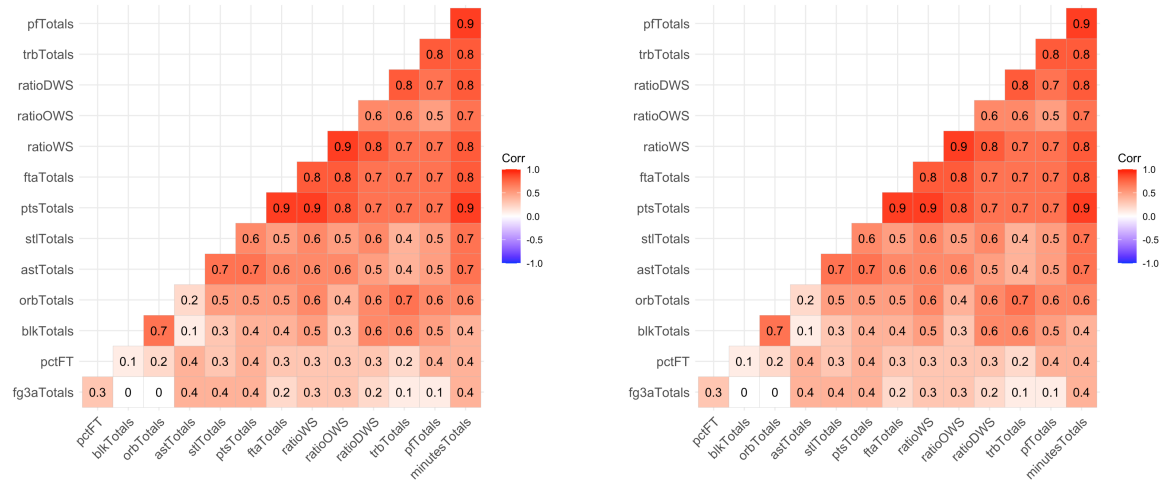
In order to explore any high-level feature/response relationships, we first analyzed the entire data set, which includes NBA statistics for all 14 features from 1959 to 2019. This analysis will hopefully allow us to examine important cross-temporal patterns in certain statistics and, later, in our models.

The first task of our EDA was to explore the distribution of our multiclass categorical response. This was partly to ensure each position was adequately represented in the data and also to see if any of the three (G, F, or C) are more common than the others. The barplot featured below shows that each position has at least 4000 observations with F and G having almost double that of C at around 8000 each. This difference simply reflects the fact that a team typically has two guards, two forwards, and one center on the floor. Given this fact, the number of guards and forwards in the data should reflect this 2 to 1 ratio. The counts for each category are sufficient to continue with EDA and model building.

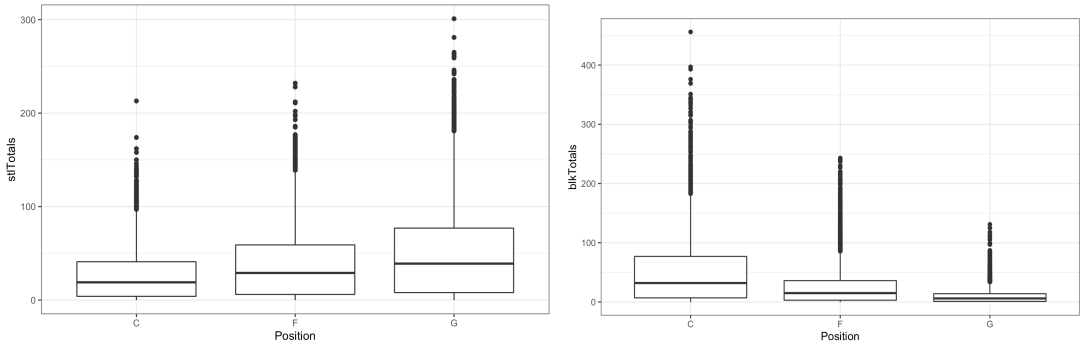


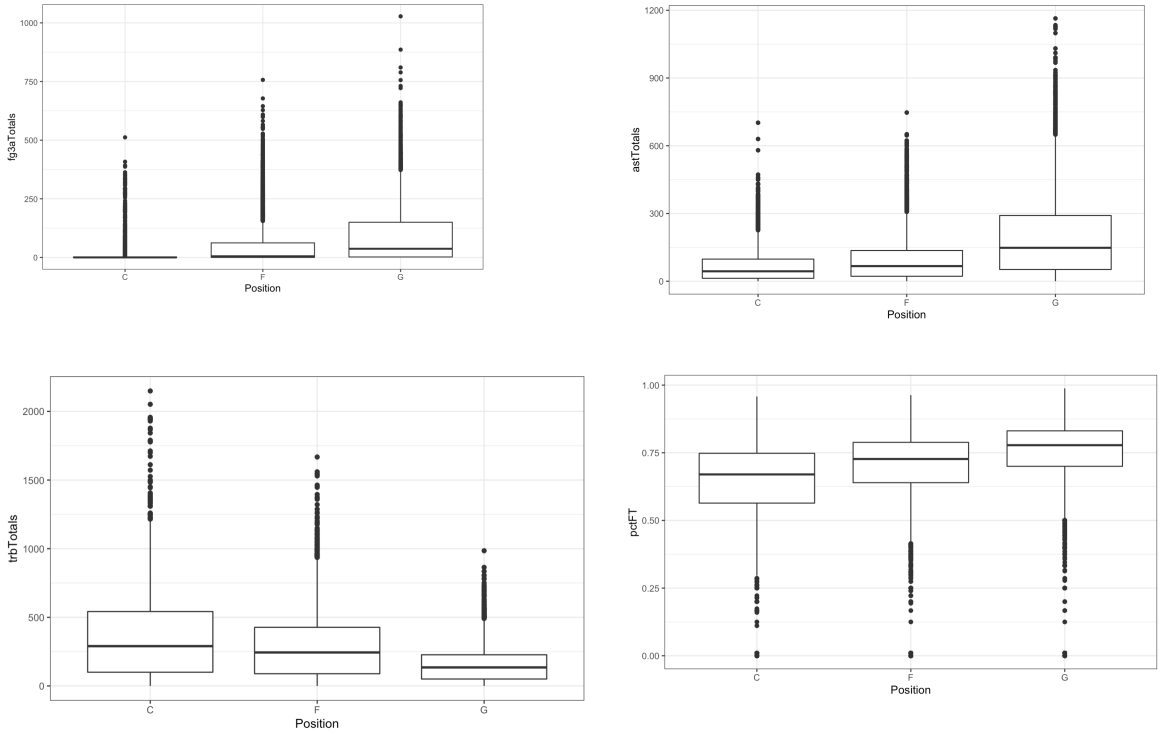
We then decided to analyze correlations between features to check for any possible multicollinearity issues. In order to do so, we used an imported package called "ggcorrplot". This allowed us to call a function which computes the correlation matrix for the features. The correlation plot revealed

high correlations between personal fouls, rebounding, and minutes. We chose to remove pfTotals, ratioOWS, ratioDWS, ratioWS, orbTotals, minTotals, and ptsTotals because they are significantly correlated with other variables. This left us with just 6 features: astTotals, stlTotals, trbTotals, blkTotals, pctFT, and fg3aTotals. Below, we display the correlation matrix with 14 original features and the updated correlation matrix of 6 features. Though blocks and rebounds and assists and steals seem to be somewhat highly correlated, this could be due to underlying similarities in the way the stats are accumulated. We will further address this in our conclusion section.



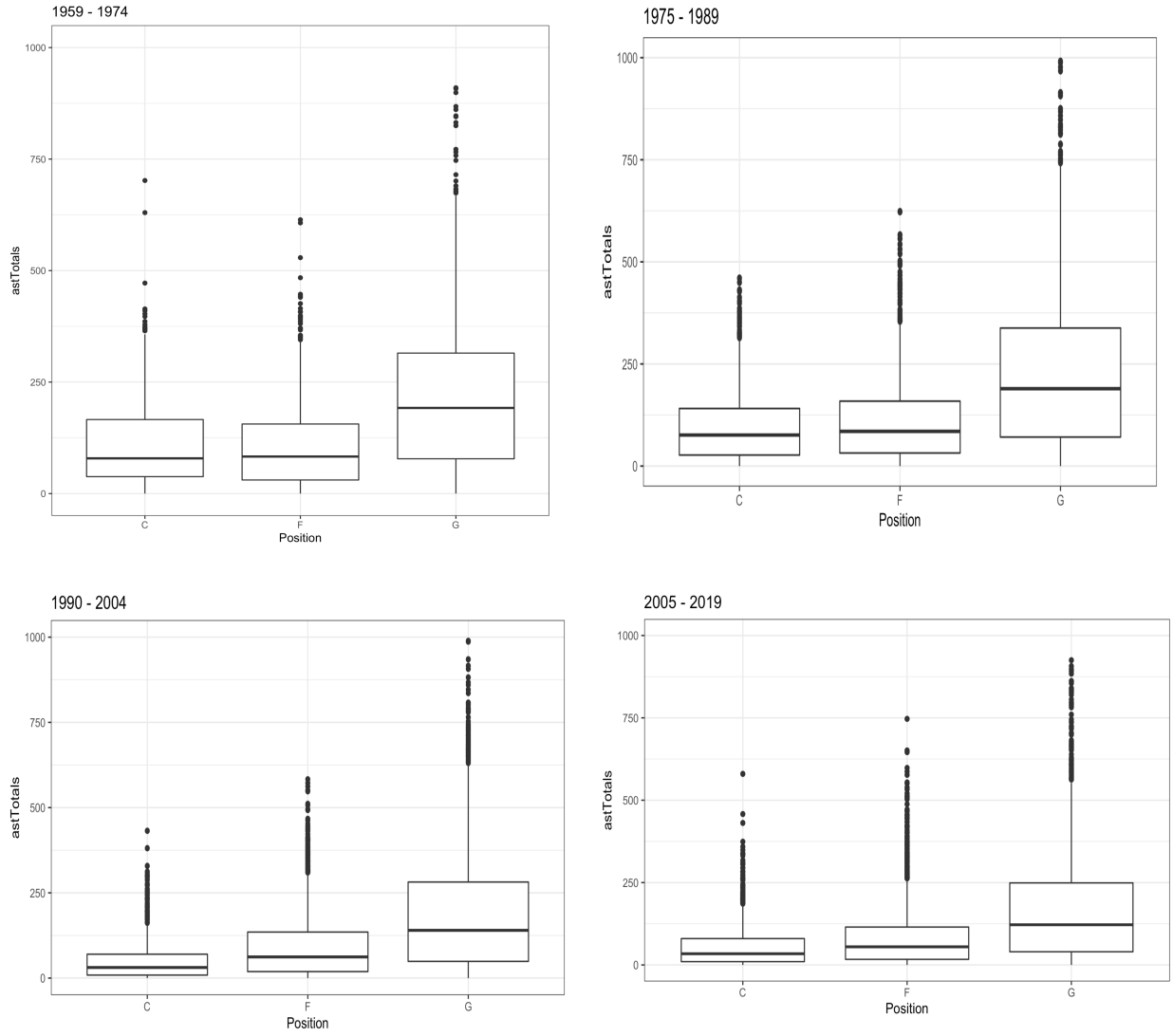
Using all of the data we graphed boxplots of select features, setting the x-axis to the position and the y-axis to the feature. This allowed us to compare the distribution of these features among the three different positions. The boxplots for each feature are featured below.





A few remarks can be made on these summary statistics and their distributions. Centers block significantly more shots on average than forwards and guards, and they also appear to grab a larger share of rebounds. Nonetheless, we still see forwards accumulating lots of rebounds too. Guards seem to dish more assists than centers and forwards by a significant amount; they also net more steals. Notably, there is very little difference in free-throw attempts between each position. Finally, it is very clear that few centers attempt three-pointers; fg3aTotals appears to be one of the most variable features by position in our dataset. Finally, forwards seem to fall somewhere between guards and centers for each feature. This may reflect their all-around playing style.

The process of analyzing the distributions of these variables while also separating the data by era would require too much space and will be covered in more depth in the modeling section. However, we used one feature, astTotals, to demonstrate how a features distribution between positions might change through the 4 eras.



The above boxplots, compared with each of the other eras, is an example of how features and their distributions might change over time. As one can see, the general trend of guards out-assisting both forwards and centers by a large margin does not change. However, the graphs show that centers and guards average significantly less assists in the most recent two eras. This represents an evolving trend of passing less and scoring in a solo fashion (1 on 1), which negates many possible assists. The older two eras featured players who focused more on passing and preset plays. This is just one example of a feature's distribution changing with time. This type of analysis will be further explored in evaluation and interpretation sections.

C. Training and Testing Split

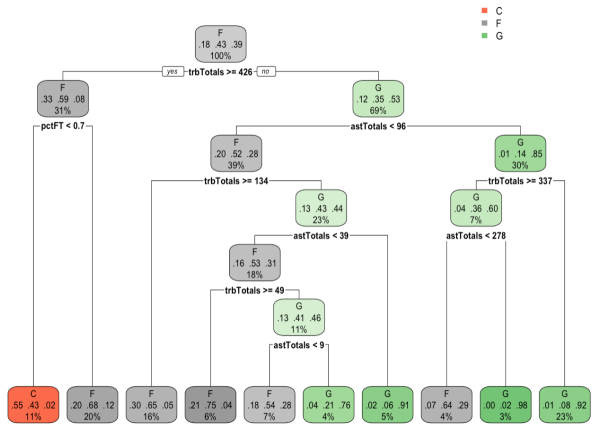
Before beginning to build our predictive models, we split each era's data into training — used to build and generate models — and testing datasets — used to evaluate models. Each training set consists

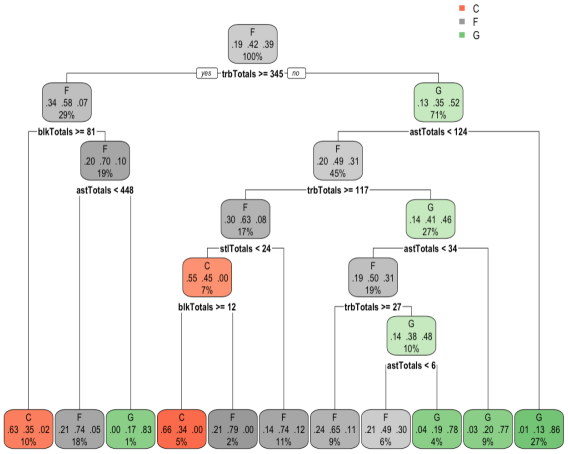
of 80% of the observations in that era, and the testing set has the remaining 20%. We train our models on each era and the entire dataset in the following section.

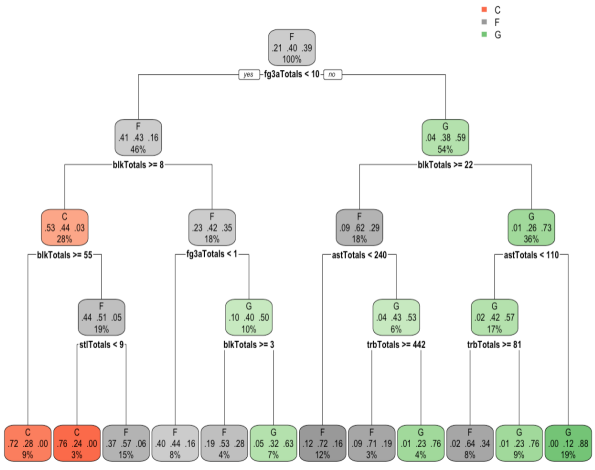
IV. Model Building, Evaluation, and Interpretation

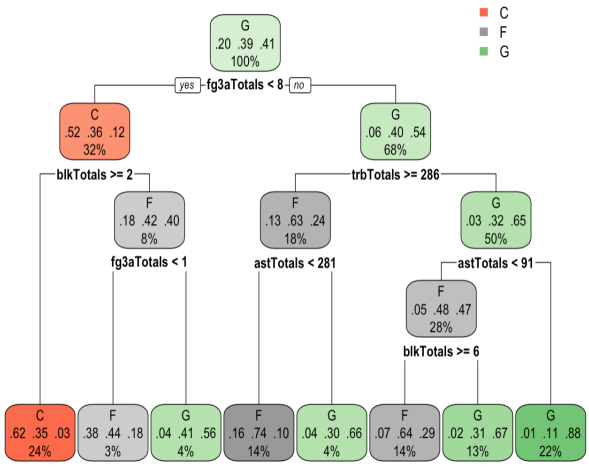
A. Tree-based Classifier

We begin our model building by constructing and analyzing simple tree multiclass classifiers. Trees essentially split the data into multiple nodes by finding ideal features and ideal values of that feature to split on which will reduce the models gini index (the purity of all the terminal nodes in the model). As you can see in the trees below, you can step through these splits, eventually ending at the prediction. We constructed simple trees that are displayed below for each era. These trees are very interpretable, but for our final tree-based classification prediction, we will use a pruned tree model that is not shown in this section.









For the first era (1959-1974), it appears that rebounds, assists, and free-throw percentage are the most important variables in classifying position. Notably, this era was played almost entirely without blocks and steals recorded; these stats were only recorded beginning in the 1973-74 season, the last

season in this set of observations. Thus, the tree does not consider these stats to be as important. For the second era (1975-1989), rebounds and assists are still important in classifying position. However, with the introduction of steals and blocks in the 1973-74 season and the three-point shot in the 1979-1980 season, we see these variables usurp free-throw percentage in this era's tree model.

For the third era (1990-2006), we see the tree gets slightly smaller but maintains some of its general structure. Three-pointers, blocks, steals, and rebounds are the important features in the model. The model can adeptly differentiate between guards and centers in most nodes, analyzing three-point totals, block totals, and assist totals. For the fourth and final era (2005-2019), the tree gets even smaller. The model now only includes three-point totals, offensive rebounding totals, assist totals, and block totals. For centers, blocks are extremely important — having at least one block with less than 8 total threes taken on the year is a good indicator of being a center for this era. To differentiate between guards and forwards, the model looks at offensive rebounds and assist thresholds; we would expect forwards (who are taller on average) to grab more offensive rebounds while we would expect guards to dish more assists.

We then grew out the deepest tree for each era and the entire period by setting `minbucket` to 1, `minsplit` to 2, and `cp` to 0. Then, we pruned back each tree and selected the optimal number of terminal nodes using the one-standard error rule. Unfortunately, the trees constructed this way were large, making them difficult to interpret meaningfully.

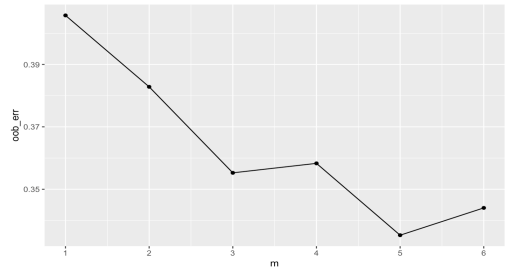
Making general observations, we note that the model on the full data appears to mostly rely on values for three-point field goals attempted, assists, steals, and blocks. While three-pointers, steals, and blocks weren't recorded throughout the entire dataset, assists are recorded throughout NBA history and are clearly an important variable in classifying player position. From these initial trees, it seems the algorithm is understanding some of the key differences between guards, forwards, and centers. Centers in general accumulate more blocks and rebounds due to their relative height and strength, and most big men shoot less threes. Guards also do tend to rack up more assists than forwards because they often are the team's primary ballhandler.

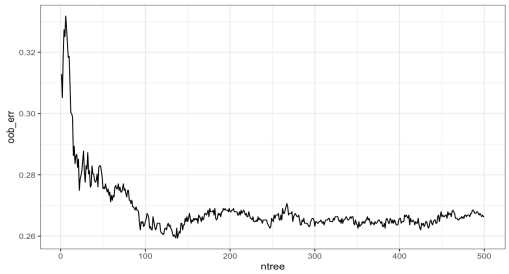
B. Random Forest Multiclass Classifier

We next tried a random forest classification method with 500 trees to start. Random forests are somewhat like trees that use random samples of the training data with replacement and aggregate m trees. These trees are averaged together to get the final predictions, reducing the overall model variability.

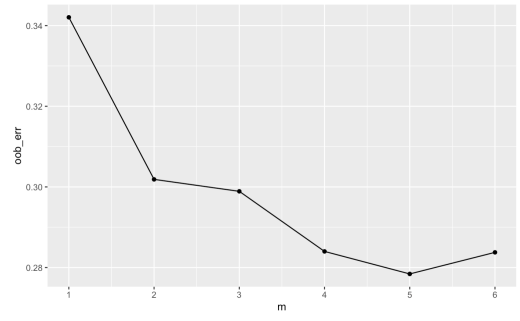
We observed that the OOB error usually bottomed out quickly for each era and the full period, which was a good sign. Instead of building on these random forest models, we instead chose to optimize our models by tuning `mtry` to see which value produced the lowest OOB error. `mtry` is the number of variables available for splitting at each tree node. Below, we display the OOB error for each era and the full period. We also display the `mtry` tuning graph.

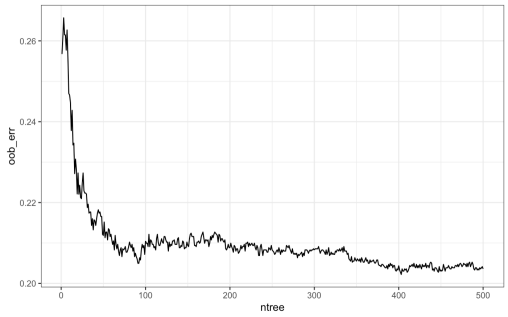
Era 1 (1959-1974)



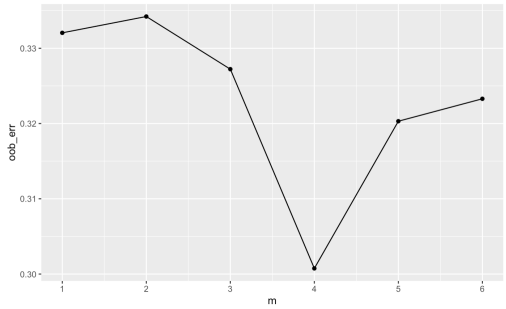


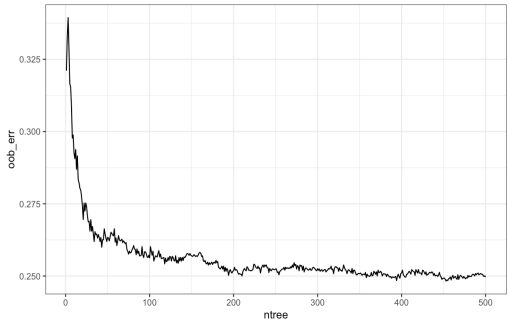
Era 2 (1975-1989)



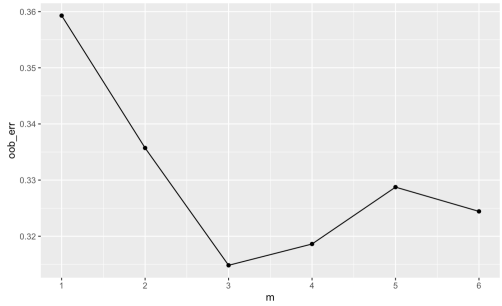


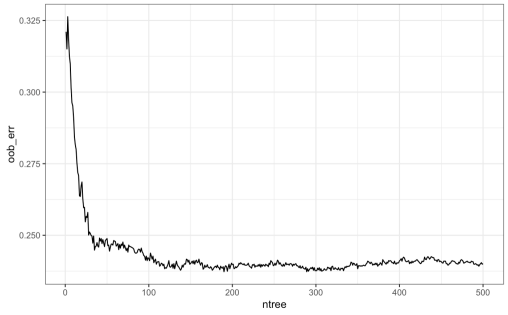
Era 3 (1990-2004)



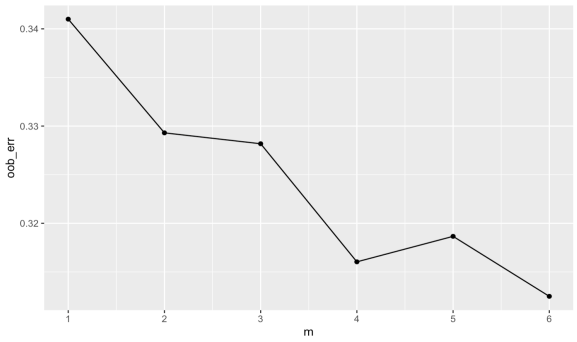


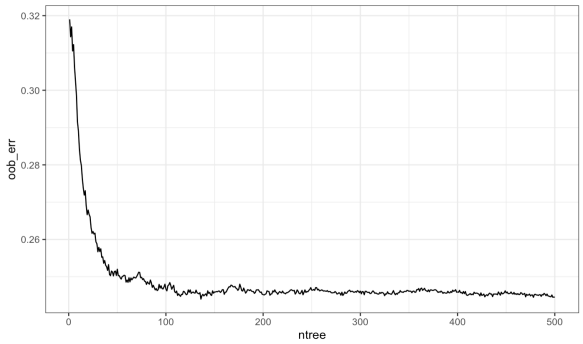
Era 4 (2005-2019)





Full Period (1959-2019)





Now we will discuss the variable importance plots for each era, which are omitted due to lack of space but can be found in our code. In the first era, the model is able to discern that assists are important. Since guards and forwards are smaller, quicker, and more agile players in general, they are entrusted with primary ball handling duties. As a result, they often distribute the ball and accrue lots of assists because they have the ball the most often. Interestingly, free-throw attempt totals were not a significant factor in this era, despite the tendency in basketball's early years to foul bad free-throw shooters like Wilt Chamberlain. During the second era, steals and blocks were recorded during all years, so we see the model place more importance on these features in this era. Three-point shots were introduced in 1979-80, but the three ball was not a primary focus of teams' offensive strategies at this time, so the model deems it least important out of all our features.

In the third era, we see three-point field goals become more important in classifying player position. The three-point shot was used more frequently in this period than both prior periods as teams recruited players who could shoot a high percentage from beyond the arc. Finally, the features in the fourth era have relatively similar importance to the third era. Blocks, assists, three-point shots, and rebounds appear to be the most important features. In today's modern NBA, teams have placed a greater emphasis on the three-point shot, and we see this reflected in this era. Still, the number of rebounds a player accrues (usually more if the player is taller/stronger) and the number of assists a player makes (usually more if the player is smaller/agile) are the most important aspects of this model.

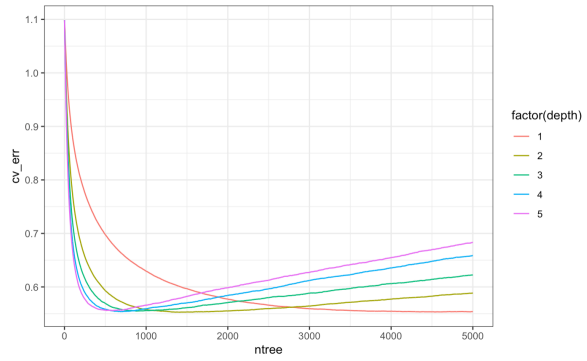
When analyzing the entire dataset, we note that trbTotals and astTotals are biased statistics because they were recorded during all 60 years in the dataset; the same cannot be said for other features. Thus, these variables' importance is inflated relative to fg3aTotals, for example, which wasn't recorded until the 1979-80 season.

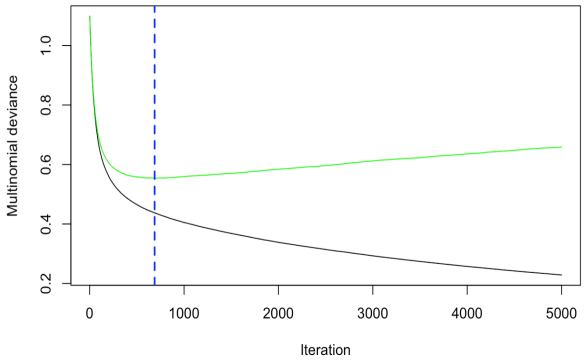
C. Gradient Boosted Multiclass Classifier

Lastly, we tried gradient boosted multiclass classifiers on each era of data and the entire dataset. Gradient boosted models use trees of a reduced depth in order to iteratively fit the data and the residuals. A tree is fit to the training data, and the predictions are multiplied by a fraction. Then a new tree is fit to the residuals of the train data, and the previous steps are repeated many times, fitting to the residuals and multiplying by a fraction to predict.

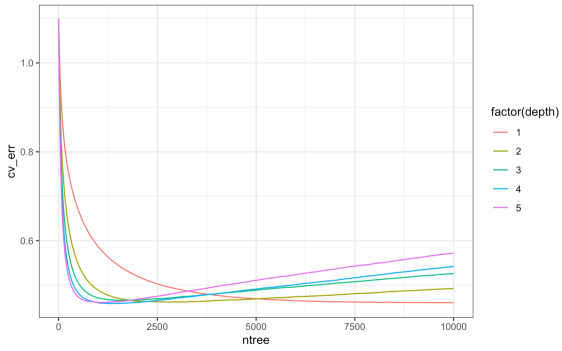
We set a multinomial distribution in the gbm() function, with 10 cross-validation folds, enough trees to bottom out the multinomial deviance curve, and a shrinkage factor of 0.01. At first, we employed simple models, but we will display the results from our tuning process. Below, we display the cross-validation error plot for each era for each level of interaction depth (1-5) and a visualization of `ntrees` in the optimal model.

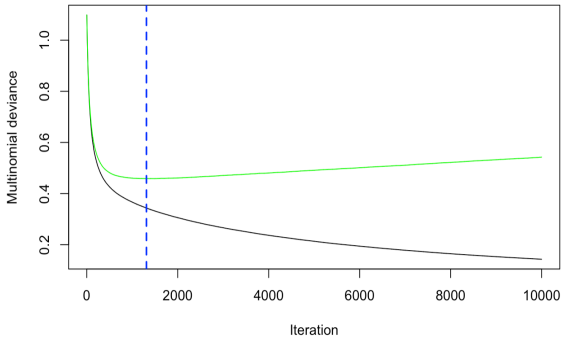
Era 1 (1959-1974) with 5,000 trees (interaction depth 4 selected)



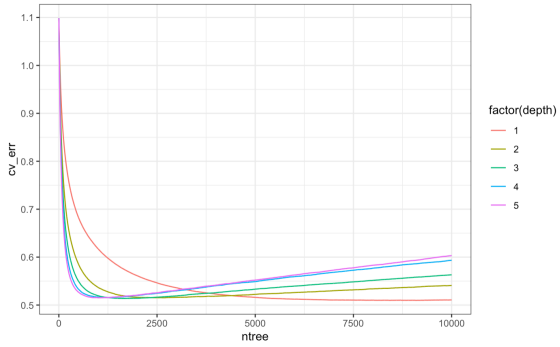


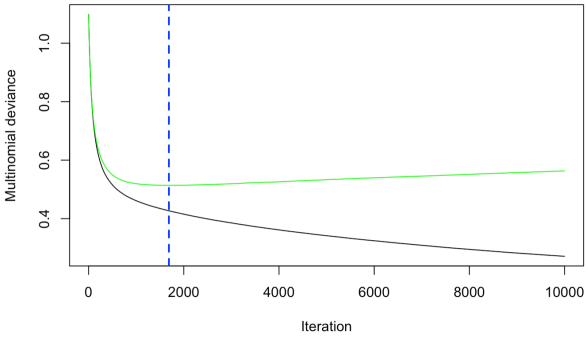
Era 2 (1975-1989) with 10,000 trees (interaction depth 4 selected)



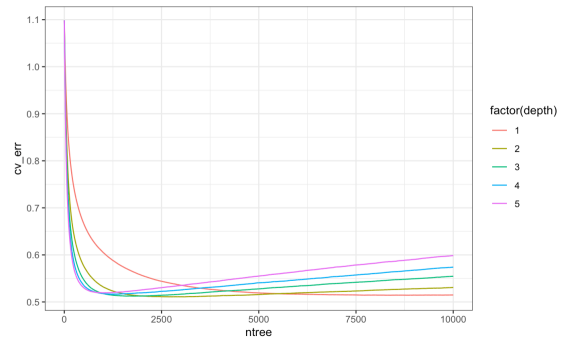


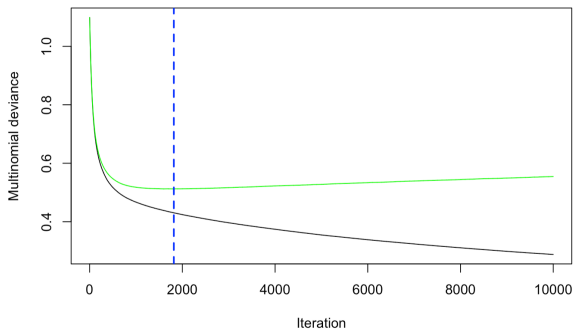
Era 3 (1990-2004) with 10,000 trees (interaction depth 3 selected)



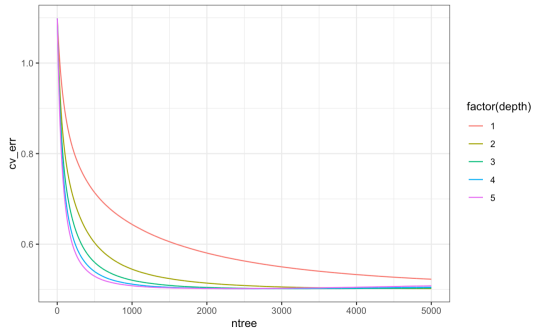


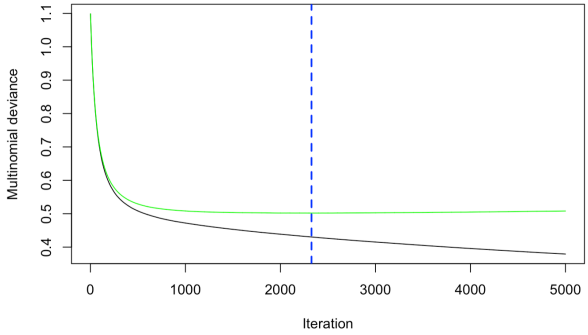
Era 4 (2005-2019) with 10,000 trees (interaction depth 4 selected)





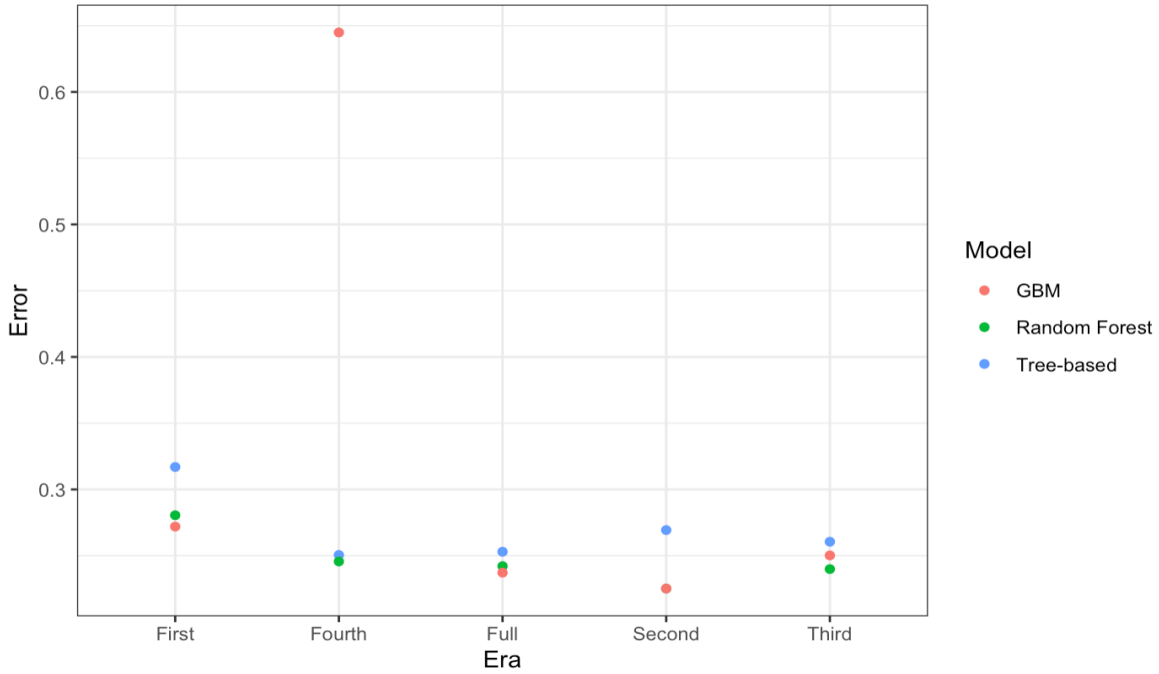
Full Period (1959-2019) with 10,000 trees (interaction depth 5 selected)





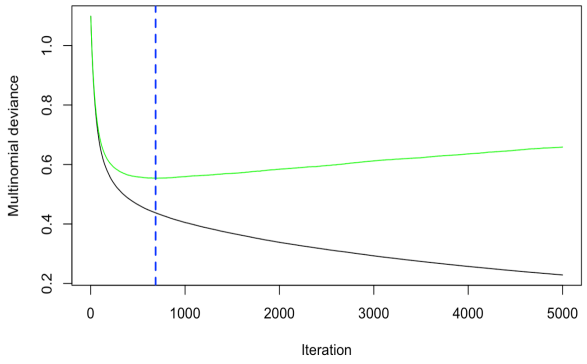
V. Final Model and Conclusions

A. Model Performance, Evaluation, and Comparison

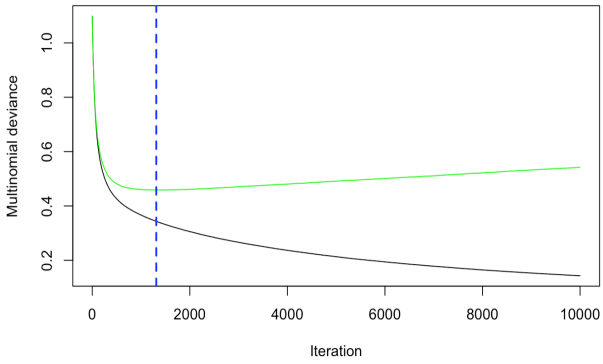


The graph above shows the test errors for the three different models in each era and for the full training data for the most part. In comparing the models, the random forest classifiers seem to result in the lowest test errors. As one can see, the tree-based models have test errors which are higher for the most part. Additionally, the random forests are much more stable than the GBM models. Given these facts, it appears as though random forest models are a suitable choice for our data. Below, we present the error curves for the best-performing model in each era and the full time period.

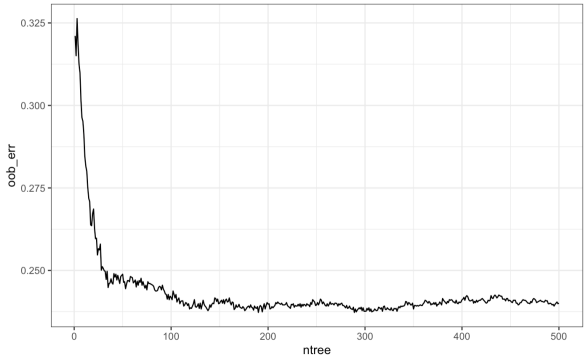
Era 1 (1959-1974)



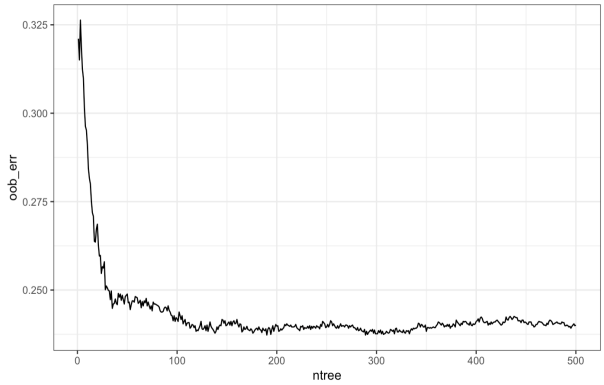
Era 2 (1975-1989)



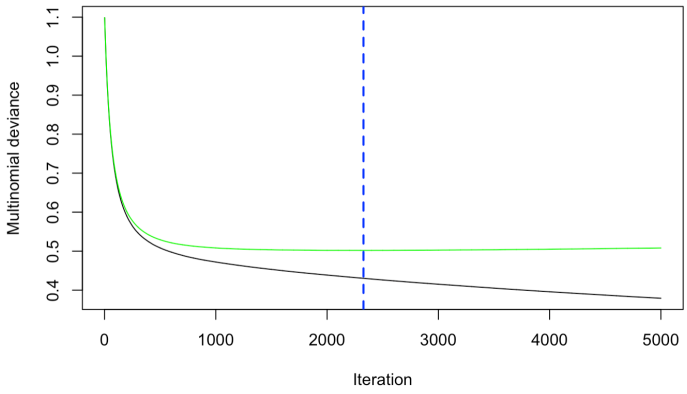
Era 3 (1990-2004)



Era 4 (2005-2019)



Full Period



The current study is limited in a few ways. The first issue is that data for years near the NBA's inception is scarce; comparatively few statistics were actually recorded in these early years, and more advanced NBA statistics like PER and VORP weren't introduced for a long period of time. As a result, it was impossible to compare the first era to its successors, though it could have made for interesting discussion to see how the significant stats evolved quickly over the first few years of play. Lastly, although our study classified positions as guard, center, and forward, it fails to differentiate between point and shooting guards and small and power forwards.

B. Conclusions and Future Directions

Our robust analysis has helped us build an optimal model for each era and the entire dataset, and it has also given us valuable insights into what variables are important in each model. In the first era, assists and rebounds were of the most importance to the model. However, as other stats began to be recorded, these metrics became more and more important in the models. For example, three pointers were non-existent from 1959-1974. In the next era, you see three pointers are still not a significant feature, even though they had been introduced as a stat; this represents teams reluctance to attempt threes and stray from the strategy they had historically practiced. The next two eras listed three point attempts as a significant feature, becoming more significant in the most recent era; this represents teams increased adoption of the three point shot and namely the influx of threes in recent years which can be attributed to players like Steph Curry, Damian Lillard, Klay Thompson, etc.

We anticipate a number of possible follow-up analyses that naturally arise from our study. For example, instead of classifying players as guard, center, and forward, we could classify players according to the traditional five-man lineup: PG, SG, SF, PF, C. Further, we could also use different features in subsequent analyses in order to gain better insights into how NBA statistics relate to each position. Another possible follow-up analysis could involve the same features trained and tested on international basketball data. International basketball is played much differently from the NBA in terms of strategy, playstyles, personnel, etc. It would be interesting to see what statistics might matter in predicting positions in basketball leagues outside the U.S.. Given that the NBA is the ultimate goal for many international professional basketball players, insights from this analysis could better shed light on the hump that international players must overcome to break into the league. Finally, we could also employ other multiclass classifier methods, including K-Nearest Neighbors, Multiclass Logistic Regression, and Neural Networks.

As the study is concluded, it is important to mention how the analysis might be important to several stakeholders involved in the NBA and basketball in general. Analysts might find this information useful in assessing whether certain players can be categorized as positions other than what they are currently listed as. For example, if a player is listed as a guard, but they grab a large share of reabounds, could they be considered a center based on the fact that rebounds are significantly attributed to centers? This type of question is worth considering in voting for all NBA teams which feature two guards, two forwards, and a center. If a talented center is unable to make all NBA at that position, but they play much more like a forward, could they be categorized as a forward for the purposes of the award? These types of questions could have a significant impact on the fluidity of positions in basketball.