**NBA Game Prediction**

**Abstract**

This study aims to compare the efficacy of different modeling approaches in both predicting and explaining the outcome of National Basketball Association (NBA) games from 2015-2023. Modeling of game outcomes is undertaken and leveraged by a number of industry stakeholders, from the teams involved to media analysts to sportsbooks offering odds for sports gambling, which collectively earn hundreds of millions of dollars in revenue annually. The scope of this study was to use previous home and away team statistics to predict each game’s winning team, with the goal of minimizing misclassification. We found that while the most accurate modeling approach was the linear discriminant analysis (LDA) model, there were some reservations about potential bias. Further research into this area could extend the data used from basic counting statistics to other measures of team success, drill deeper into player performance and availability, and improve the statistical rigor of our approach.

**Introduction**

In this report we’ll be exploring the prediction of NBA games using typical box score statistics. While there is significant investment and research in this field, much of the development in this space is done within organizations with motivation to preserve competitive advantage either against other competition - other teams in the case of NBA teams, and other gambling companies and the public in the case of sportsbooks. Despite this, machine learning approaches have been applied to the problem to varying results.

What we hope to add to this space is both a comparison of the effectiveness and communication ability of a selection of machine learning models. Additionally, we hope to provide a straightforward mechanism for adding additional data to the models, providing a framework for future enhancements to our approach with more data and additional modeling approaches.

**Data Sources**

The data that is the basis for our modeling (located at [https://www.kaggle.com/datasets/nathanlauga /nba-games](https://www.kaggle.com/datasets/nathanlauga%20/nba-games)) was sourced from Kaggle. We took the game-level statistics for home and away teams and prepared a season long average for key statistics for both teams leading up to each game: points for and against, assists for and against, and rebounds for and against.

This data wrangling approach gave us data points that would be available before each game for both the home and away teams. We excluded games where a team had not yet played that season. This led to twelve total variables that underpinned measuring team performance offensively and defensively.

In addition to these predictor variables, our outcome variable is whether the home team wins (HOME\_TEAM\_WINS), which was assigned as binary where 1 represents the home team winning. In our dataset, around 60% of games are won by the home team (often referred to as a home court advantage), making the designation of home and away teams important to maintain through our data analysis.

**Proposed Methodology**

With our two goals of both accuracy in predicting wins and diving deeper into the drivers of those wins, we will take a look at three populations of data: NBA games across all teams, and then games for the most and least successful teams from the sample - the Golden State Warriors and Orlando Magic, respectively. Our primary focus for modeling is to use the entire population for tuning and comparing the accuracy of models that predicts games for all NBA teams, while producing case studies in the drivers for Orlando and Golden State that show the differences in individual team success.

For those case studies, we’ll use a basic logistic regression model that details the most important variables for the overall population compared to each of those teams specifically and compare the magnitude and direction of the coefficients for the three models.

We will then use the whole dataset to compare four different modeling approaches to determine the most accurate approach in predicting whether the home team wins.

Taking the wrangled data and splitting it to a 75/25 train-test distribution, we’ll feed it into four different models, each of which will be tuned individually. We’ll use a logistic regression model, a K-Nearest Neighbors model, a linear discriminant analysis model, and a random forest model, taking the most accurate. We’ll then compare each of these models in a Monte Carlo simulation and determine the statistically superior model.

**Literature Review**

With the increased popularity of sports betting, attempting to create models to predict sports outcomes has been greatly explored, and not just for the NBA. Bunker (2017) provides a extensive framework to be applied across many vastly different sports. As sports analytics has grown and datasets become more available, sports bettors are constantly looking for that next edge to help them bet with confidence. Basketball and the NBA have not been immune to this, and discussions around the NBA have become more and more betting centric as it’s become more mainstream. Below are a few intriguing topics around predicting NBA games that stood out to us as we were conducting research.

An important piece that we are missing from our dataset is capturing individual player performance. With our dataset being team focused, we are unable to account for things like matchups, or the presence of a star player (perhaps multiple) on a given team. In Osken and Onay’s (2022) paper, they approach predicting game outcomes by clustering players into groups where they are evaluated within, the player groups within each team are then used to predict outcomes. By capturing a team's composition, player profiles, and their roles within teams, this modeling approach captures a different element of the sport.

Another intriguing aspect of predicting NBA games is the grueling schedule that the teams go through. Often teams will play games on consecutive nights, potentially taking long flights in between as well. The schedule features 82 regular season games over a 6.5-month span, and the most successful teams will go onto play at least another 12+ games in the playoffs. First introduced by Etine and Small (2007), their paper dives into the impact of rest days in these grueling schedules, especially when looking at it from the home team perspective. A typical scenario could be a team is coming into your home arena, having just played the night before in another city. This is a tall task for an away team, and being able to capture that disadvantage in a predictive model could increase model performance.

When looking into research around other team-based models, we came across a paper by Chao (2012) that interestingly, used the teams' most recent 10 game averages rather than the season long averages. Given the length of the NBA, with constant lineup changes and player injuries, teams can be quite inconsistent. By taking the most recent 10 game sample, the team averages are capturing the recent trends and performance of a team and would not be skewed in any way by the performance beyond the last 10 games. We think this approach has both advantages and disadvantages. For instance, if it is game 61 and the model is making predictions based off of game 50-60, then it is missing potentially valuable observations from the first 50 games of the season. While NBA teams can certainly be streaky, this shouldn't render early season data useless. An ideal approach would consider both recent and season-long performances, but perhaps weighing more toward the season long averages.

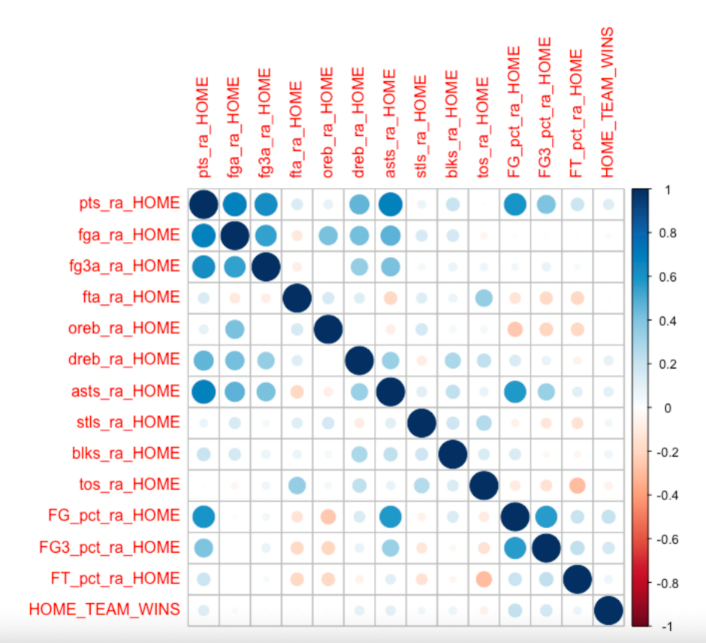
Lastly, a constant theme in our research was the advanced statistics that have been created over the past 25 years. A couple team wide statistics that particularly stood out was True Shooting %, and Pace. Both dating back to 2005, featured in an NBA article “NBA Statistical Primer”, and we’re likely theorized on long before. True shooting takes into account efficiency on all types of shots, including standard 2-point shots, but also free throws and 3-point shots. True shooting will pay a lot of influence on performances with high scoring outputs, but with low overall attempt numbers, meaning a team is making the most of their opportunities – no matter where they come. Pace on the other hand, gets into the amount of possessions a given team has in a game. This is an important stat because it essentially captures opportunity. The more possessions a team has, the more opportunities to score they have, and therefore high pace teams will often rank near the top in points per game. We have points per game in our model, but by pairing that with pace, the element of team efficiency can once again be captured.

There have been many approaches done to predicting NBA outcomes, and we expect more and more novel approaches to spring up as sports betting continues to grow. While we particularly looked at team statistics, from our research we can see the benefit of being able to capture these other elements. As we continue to adapt the model, we will keep in mind potential additions around player performance, scheduling, and more advanced statistics.

**Data Analysis**

Looking at our data as a whole, a few characteristics could eventually be troublesome for some of our modeling approaches. Despite the large number of observations in our dataset (9973), they have a couple of key overlaps that could be troublesome. For one, thirty teams produce this data, with overlap in performance season over season. Given the nature of our variables as averages of game data, we’d expect to see some level of correlation especially among statistics like assists and points, which often go hand in hand.

Let’s take a visual look at the correlation between all the variables in the dataset. The first step we’ll look at is a heatmap of correlation between each of the variables, home and away (Figures 1 & 2):

A screen shot of a graph

Description automatically generated

Figure : Home Team Correlation Plot Figure : Away Team Correlation Plot

We see that though there is some level of correlation between our predictors and response, it is minimal. The strongest correlation value of 0.2132 between HOME\_TEAM\_WINS and FG\_pct\_ra\_HOME is a relatively weak relationship. Below are the top 6 strongest correlation values. Notice how 4 of the top 6 are variables associated with the away team, and therefore possess a negative correlation value with home team wins.

|  |  |
| --- | --- |
| *Variable* | *Correlation value* |
| FG\_pct\_ra\_HOME | 0.2132 |
| FG\_pct\_ra\_AWAY | -0.1773 |
| FG3\_pct\_ra\_HOME | 0.1624 |
| Pts\_ra\_AWAY | -0.1554 |
| Dreb\_ra\_AWAY | -0.1447 |
| FG3\_pct\_ra\_AWAY | -0.1437 |

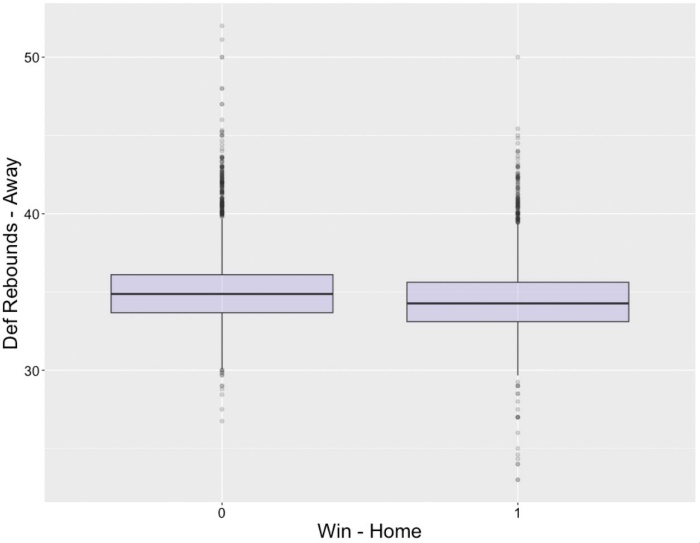
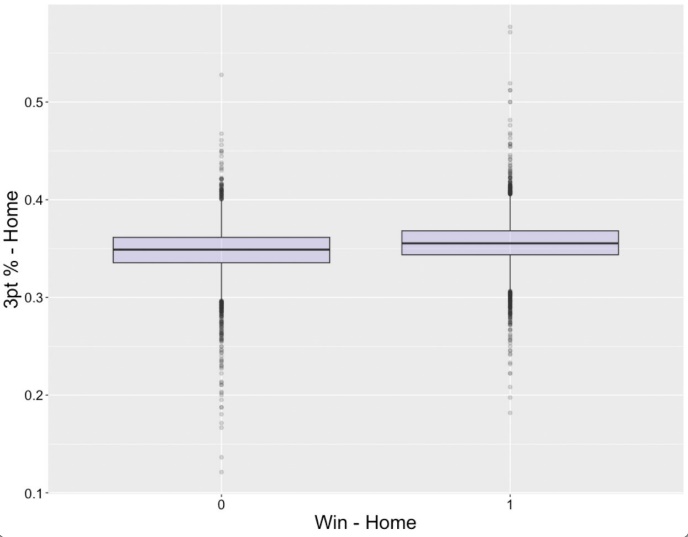


Figure : Home Team Boxplot Figure 4: Away Team Boxplot

In Figures 3 and 4 above, box plots were created for a home team statistic, and an away team statistic, both falling within the top 6 strongest correlation with home wins. Figure 3 features FG3\_pct\_ra\_HOME, where in wins, the percentage is slightly higher. In Figure 4, Dreb\_ra\_AWAY is featured, where the mean value in home losses is greater than in home wins, indicating that in home losses, the away team on average records more defensive rebounds. Overall, there was not much variability observed in these in-game statistics between home wins and losses, likely contributing to the low correlation values.

Conversely, there are strong relationships between predictor variables themselves, especially between home and away assists and their corresponding points, field goals made, and 3pt field goals made. This tracks intuitively given assists always lead to points, although points can be scored without assists. We will have to be careful in the treatment of highly correlated predictors.

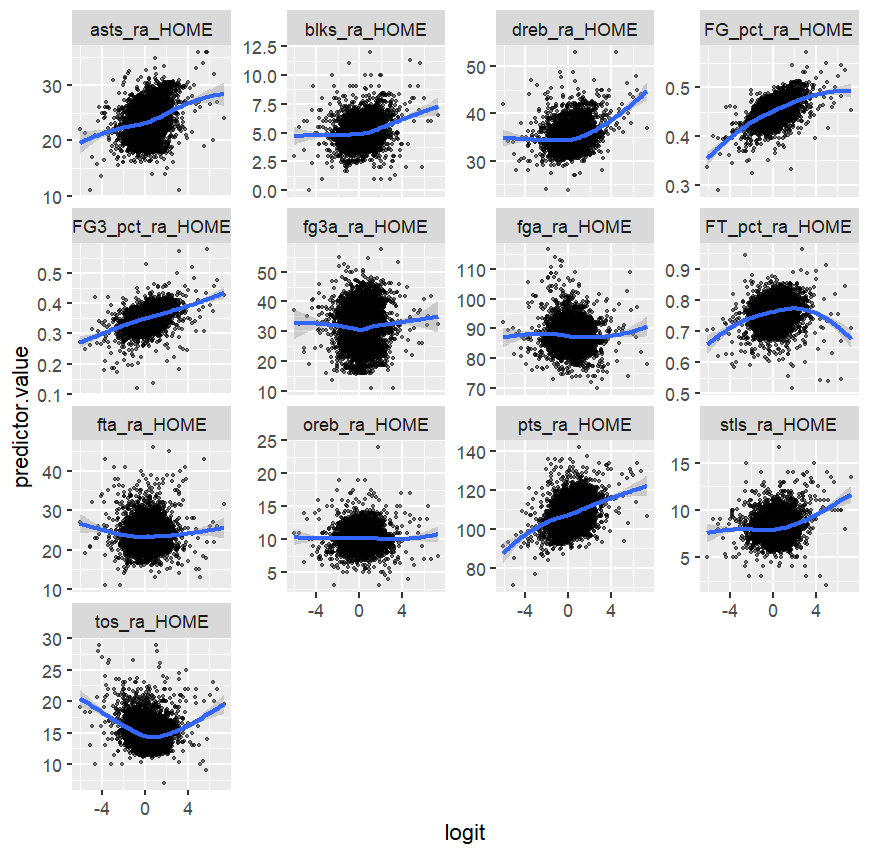
Next, we’ll look at our models and the assumptions that underpin them. Logistic regression will take the most space here, as it is a model that works best when the data fits certain criteria.

**Logistic Regression**

In logistic regression, we have four main assumptions to avoid bias in our results:

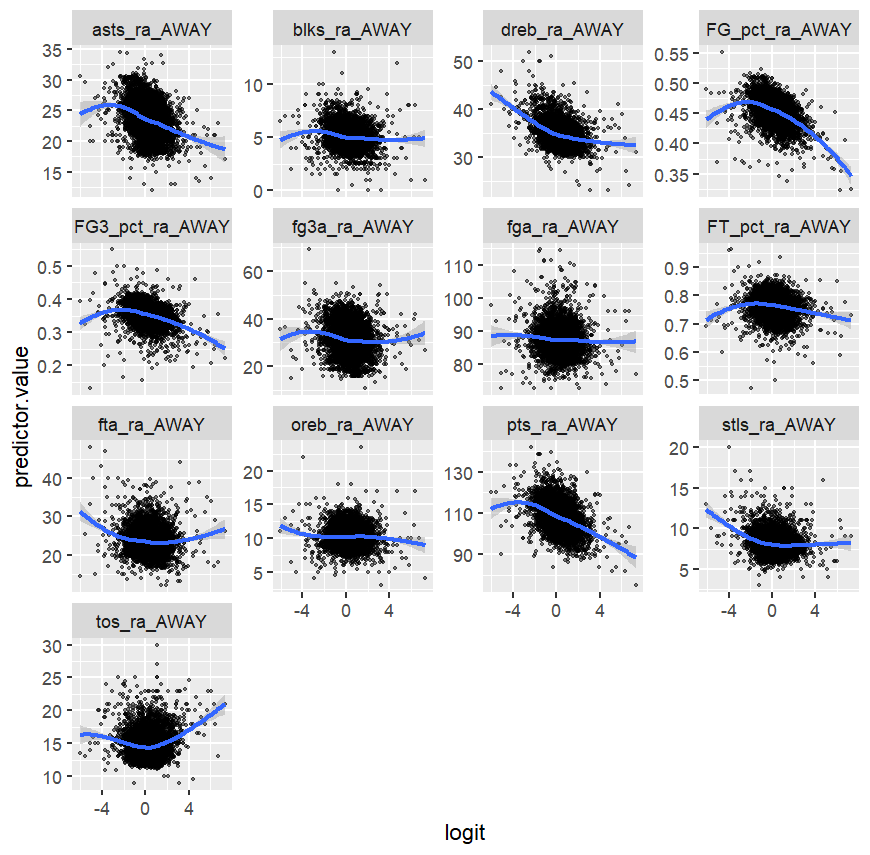
* That the response is binary (0,1)
* Multicollinearity does not exist among the predictors
* No outliers exist in the continuous predictors
* There is a linear relationship between the logit of the response and the predictor variables

To start, we’ve transformed our predictor variable to be binary, so the first assumption holds. Moving to the evaluation of a linear relationship between the logit of the response and the predictor values, we can plot these two elements against one another, again separated by home and away statistics (Figure 5):



*Figure 5: Logit vs Home Predictors*

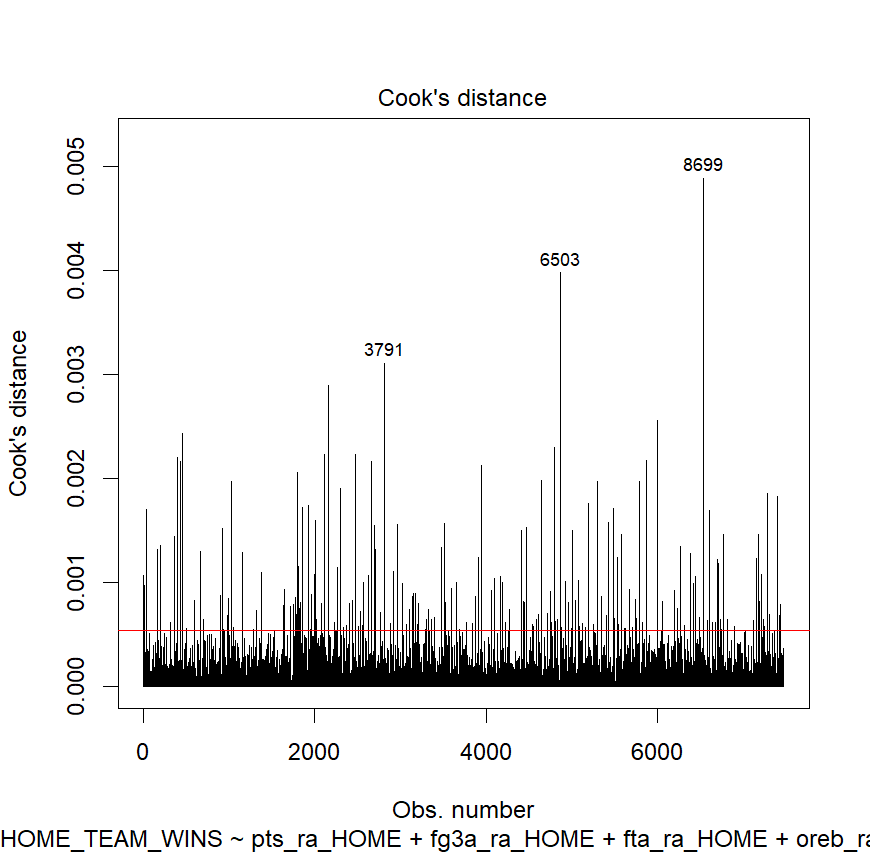
For the home statistics, we see that while on the main, the variables and the logit function have a linear relationships, there are two main exceptions- home turnovers and home free throw percentage, which appear to be decidedly quadratic.



*Figure 6: Logit vs Away Predictors*

In Figure 6, we see a similar story, with away turnovers, away free throws attempted, and away field goal percentage the most extreme examples shown. Clearly, nonlinear relationships between the logit function and the distribution of the continuous predictors is something that we have to take into account with variable selection.

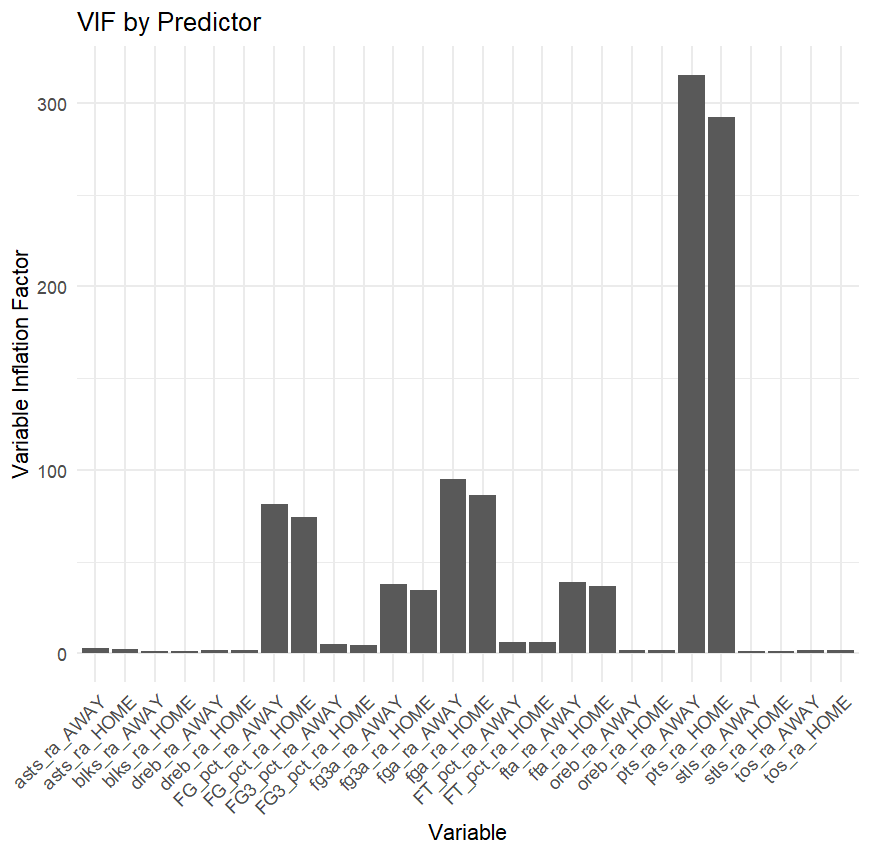
Moving on to an analysis of influential points, we’ll use a Cook’s distance plot to determine which points, if any, clear the rule-of-thumb threshold of 4/n for our training dataset of 973 observations. Our threshold is given as .000401 by that calculation:



*Figure 7: Cook's Distance*

In this case, it looks like many points clear this threshold! Options here would include transforming the data, excluding points, or applying some other model that is more comfortable with more extreme values. In total we have 250 of our 10,453 observations that are influential by the 4/N threshold. We’ll have to address this in some manner as well in our modeling approach.

The final assumption of multicollinearity can be assessed using a simple Variance Inflation Factor (VIF) evaluation for each variable:



*Figure 8: VIF for each predictor (full model, before variable selection)*

In this chart, we have a worrisome level of multicollinearity in the overlap between home and away points and other factors. There is a significant amount of multicollinearity elsewhere, but points are the most concerning. Points for and against are critical factors for our model, so it may be necessary to live with them as collinear points, violating to some degree the assumption of the linear regression model. Variable selection should reduce this appreciably, but to what extent remains to be seen.

In summary, our variable selection approach will be critical to ensuring that the results from the training of the logistic regression model will reflect in the testing data. We also have decisions to make about the data points we keep in and exclude in the

**K-Nearest Neighbors**

Moving finally to K-Nearest Neighbors, we have a somewhat looser assumption underlying this classification model. Generally, the idea is that similar data points exist in close proximity, and while the above tests of correlation and covariance are not strong, the points do have association to one another. Given these associations, we can move on confident that the results of the K-Nearest Neighbors training data will hold out into the test set.

**Random Forest**

Similarly to K-Nearest Neighbors, we make no assumptions about the structure of the data as no probabilistic models underly the Random Forest’s approach. Rather, we rely on the bootstrap sampling of the training data to grow a large number of decision trees which vote on the classification outcome, which can effectively capture the variance of nonlinear relationships like the quadratic relationships demonstrated above in Figure 2.

**Linear Discriminant Analysis**

Linear discriminant analysis’ assumptions are strong:

1. Predictors are independent of one another
2. Predictors are normally distributed

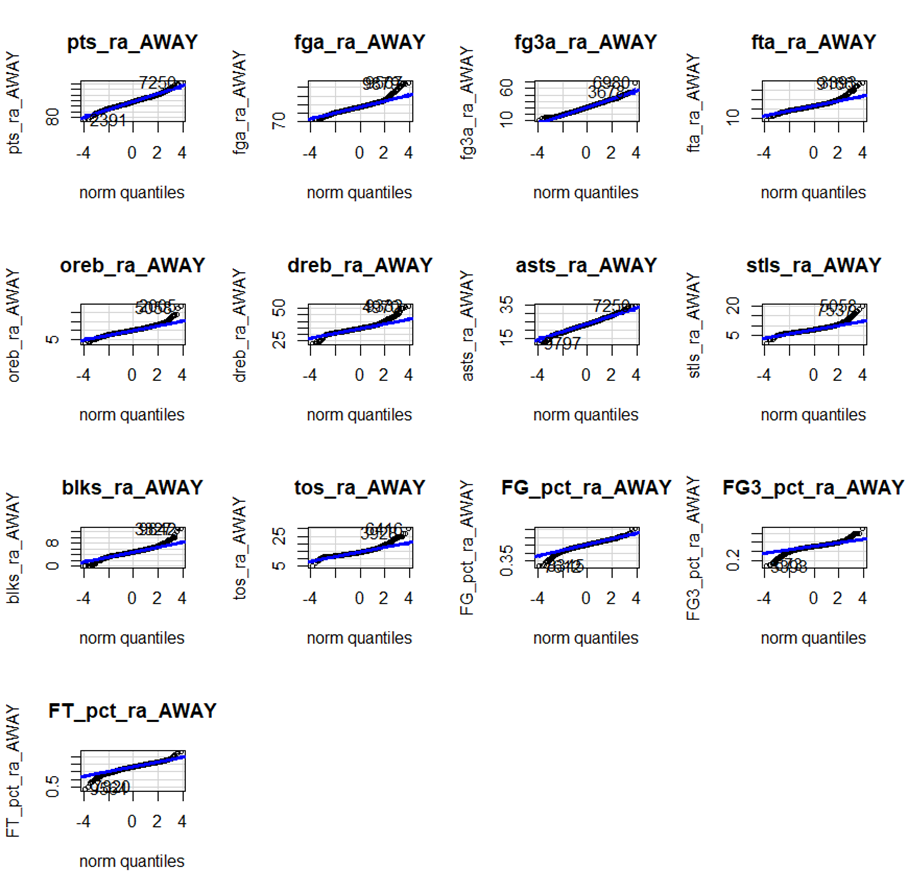
As we’ve already demonstrated with the heatmap analysis above, the covariance for each predictor varies from the response, so that is one assumption for LDA that does not hold. Looking at the graphs of the distribution of each predictor below:

A group of graphs with black lines

Description automatically generated with medium confidence

*Figure 9: Home Predictors vs Normal Distribution*

We see that there are several departures from the normal distribution in the tails of the home predictors, although the heart of each distribution aligns well with the normal distribution. In the away predictors:



*Figure 10: Away Predictors vs Normal Dist*

We see much of the same - some tail departure but no appreciable core differences. Despite these violations of the assumptions of LDA, we’ll proceed to see if we can yield a viable model.

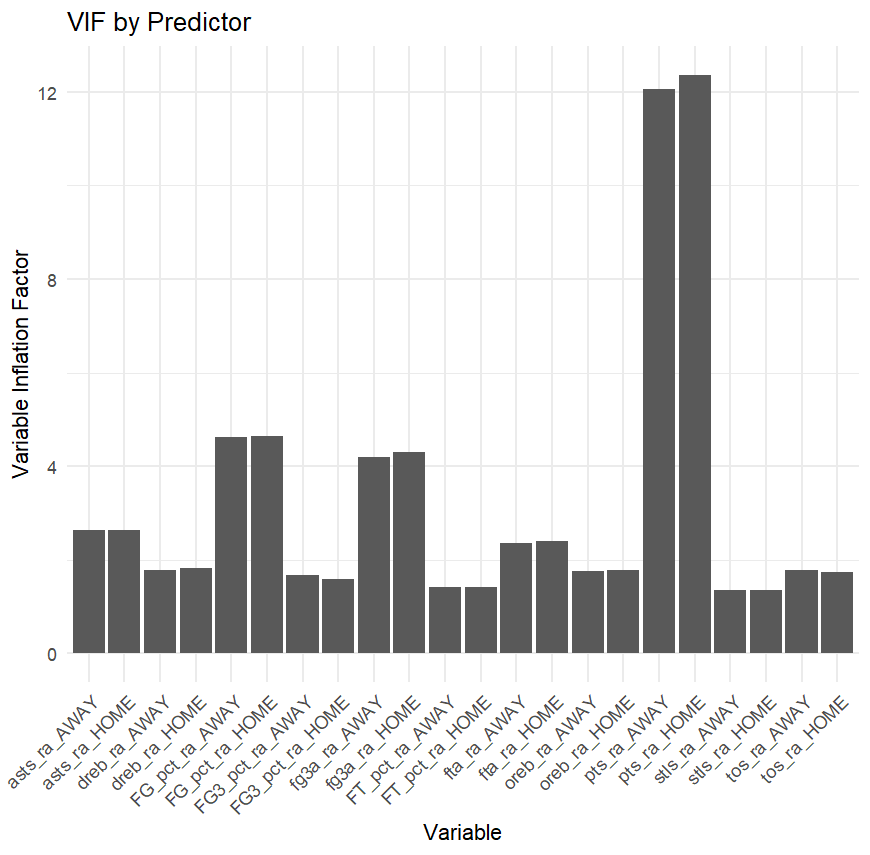
**Methods**

Moving on from the exploration of the data, we are initially tasked with creating a training and testing dataset and evaluating each of the following modeling approaches using this initial distribution of data.

**Logistic Regression:** In our logistic regression approach, we model the log-odds (and by extension, probability) of the classification between wins and losses. Our initial analysis of the model assumptions indicates that we have quite a bit to do to prepare the data for a logistic regression implementation.

We’ll first look to remove some of the most influential points as outliers. We decided to implement an approach of removing any points that were twice the 4/n threshold (so 8/n) in Cook’s distance. This removes 69 of the 250 points that we had noted as outliers by the 4/n threshold, balancing concerns about removing points that may explain variation effectively with true outliers.

Additionally, we need to perform variable selection. After stepwise variable selection, we came out with the factors that make up the below chart. We see a drastic reduction in VIF measures:



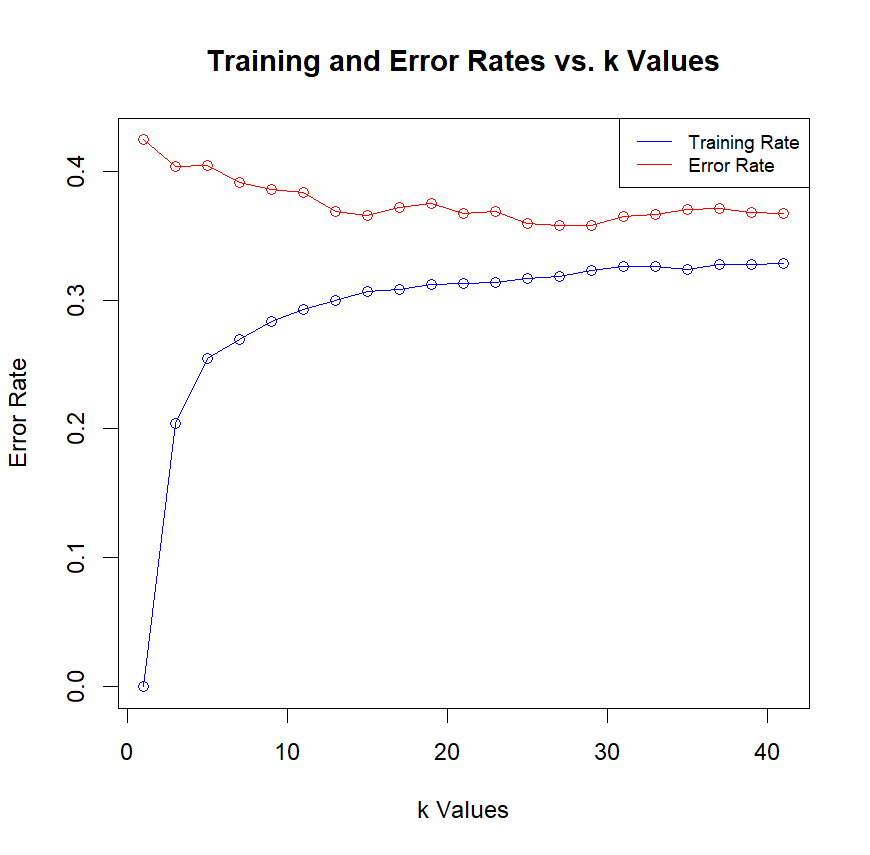
*Figure 11: VIF by Predictor (post variable selection)*

Points for and against are again collinear by our rule-of-thumb measure of 5 but given their importance to the accuracy of the model, we decided to push through with them included. Looking deeper at the stepwise variable selection, we saw that many of the defensive statistics either had a lower magnitude of impact or were left out altogether than offensive statistics like assists. Points were, perhaps expectedly, critical. In the cross validation we performed stepwise variable selection for each iteration.

Next, we determined the best cutoff for the logistic regression model. Through looping over various cutoffs on the training data from 0 to 1 at increments of .05, the most effective cutoff point for the predictors was .5, where probabilities below that point are classified as losses for the home team, and above that as wins. This cutoff was held constant in the cross validation.

**K-Nearest Neighbors:** In the K-Nearest Neighbors approach, we will use each of the variables of a new observation to classify based on the shortest distance a set of k training observations. The votes of those respective classes determine the classification of the new observation. Here, we looped with k values from 1 to 41.

One of the limitations of running this approach on training data to determine error on the training set is that k=1 will always generate an accurate classification, so we ran each k on both the test and training data. In comparing the results between each k value on the accuracy of the classifier, k=27 performed best and was carried forward into the cross validation. The errors are shown in the below plot:

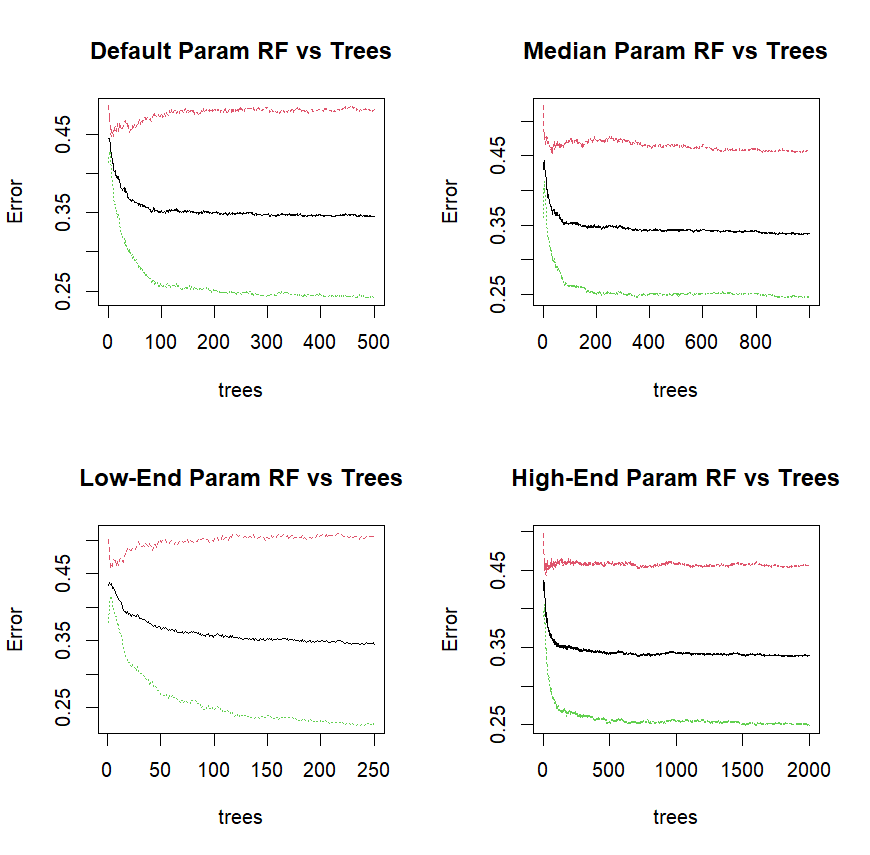


*Figure 12: Training and Testing Error Rate vs K*

**Random Forest:** For the Random Forest model, there were several hyperparameters that needed tuning. The initial variables we were concerned with were the number of trees (num.trees), number of variables to use for each split of the tree (mtry), minimum number of nodes allowed in each terminal tree (min.node.size), and the sample size of the bootstrap used for each split (sample.fraction).

First, we wanted to see if the number of trees used in the modeling approach would dramatically impact the shape of the error curve. In the default model, we initially saw the error curve flatline well in advance of the maximum number of trees that we specify, typically between 500 and 600 trees. This of course can change based on the other parameters given, but it gives us a starting point.

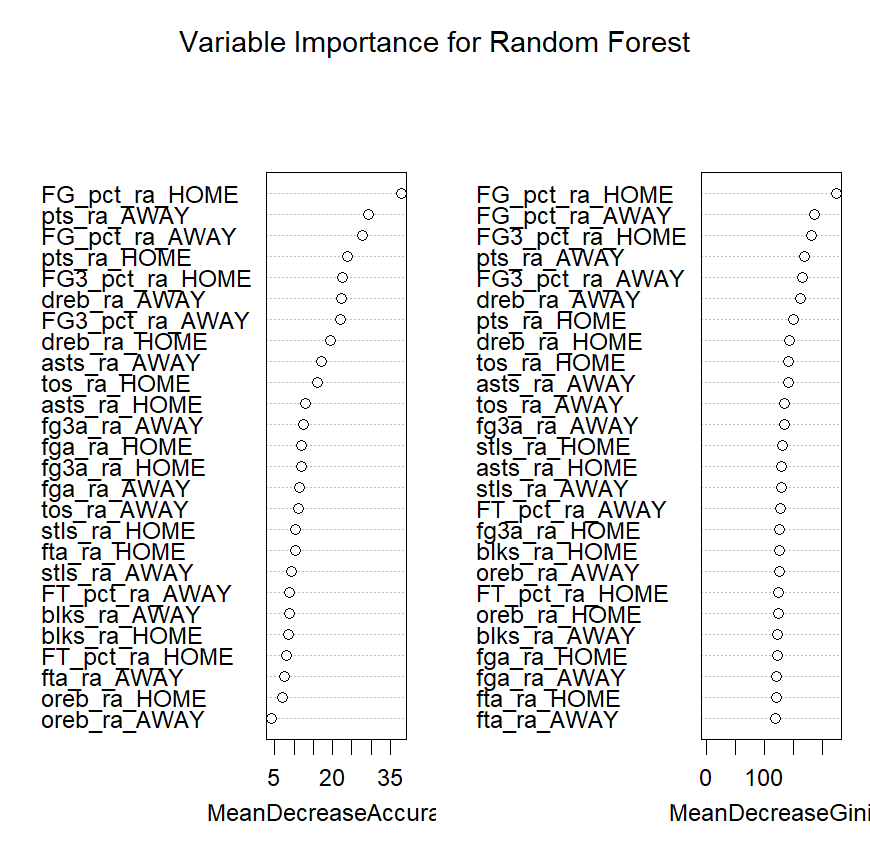
We tested a subset of the models that would eventually be included in our grid search, the default Random Forest (500 trees with other default model parameters), median of our grid search parameters (1000 trees), and the lower and upper end of our grid search parameters (250 and 2000 trees respectively). This is not an exactly an apples-to-apples comparison as the other hyperparameters were varied as well, but it helped to form intuition for the effectiveness of the model:



*Figure 13: Trees vs Error Rate*

From the above graph, we can see that the error rate does indeed flatline between 500 and 1000 for the combinations explored. Therefore, we wouldn’t expect to see much variation in error rates in the full grid search based on the change in the maximum number of trees in each of the forests.

Although the lack of interpretability of Random Forest models is a drawback, we still can see which variables drive changes in accuracy and node impurity:



*Figure 14: Variable Importance*

This shows a similar result to the variables that remained in the logistic regression model after variable selection - points and field goal percentages mattered quite a bit, while ancillary stats like blocks and steals matter less in determining the outcome of a game.

Moving to the full grid search, we varied each of the aforementioned parameters in training each Random Forest model, validating the results using the testing data of our initial 75/25 split:

Tree: 500, 1000, 2000

Mtry: 1, 3, 5, 7, 9, 11

Node Size: 1, 3, 5, 7, 9

Sample Size: .55, 65, .75, .85

This grid resulted in 360 total runs (which despite using the ranger package for optimized random forest runs, still took quite some time). We found that the best performing parameters were tree = 1000, mtry = 1, node\_size = 5, and sample size = .55. Most of the top performing Random Forest models had mtry = 1, making that seem like the most important variable for our dataset.

**Linear Discriminant Analysis:** Given our initial analysis of the assumptions of the linear discriminant analysis model, there are elements of this dataset that are not ideal for this modeling approach. Persisting through those statistical methodology questions, we wanted to present the best performing version of the LDA model we could. The main tuning element for that is determining the cutoff point for the posterior probability of home wins or losses. In this, we took an initial LDA model and, similarly to logistic regression, looped through thresholds in increments of .05 to determine the .55 threshold to be best for win and loss classification.

**Case Study - Golden State Warriors and Orlando Magic**

In order to take this analysis a step further, we also created logistic regression models for individual teams. For this report, we specifically analyzed the Golden State Warriors and the Orlando Magic, as these teams were the best and one of the worst teams respectively during the eight-season timeframe that we observed. During this time, the Warriors won four championships, and won 66.67% of their games, while the Magic only won 35.4% of their games, only making the playoffs twice and losing in the first round each time.

Winning four championships in an eight-year period is very challenging in today's NBA, and this Warriors team will go down as one of the greatest teams ever put together. With this data compiled, we wanted to look at what made them so special during this era. What variables in this dataset were most meaningful for predicting Warriors home wins?

To visualize the differences, we divided the coefficients of shared features between the case study teams and the overall logistic regression model. A large difference would reflect a different magnitude of driver, while a negative value for the divided model value would represent a change in sign.

A graph with red and blue dots

Description automatically generated

*Figure 15: Comparison Between Overall and Warriors Model*

After creating the model, we quickly observed the increased influence of 3pt shooting %, with the coefficient being nearly three times larger in the logistic regression model as compared to the overall model. This is no surprise knowing the 3pt shooting prowess of the Warriors during this time, with players like Stephen Curry and Klay Thompson, two of the greatest shooters ever. Interestingly, however, the coefficient for assists was actually appreciably larger and negative as compared to the overall model.

Additionally, when looking at variable importance, it was important to note that the top nine variables included seven “home team” (Warriors) variables, and only two “away team” variables. In our analysis, that could be because the Warriors often had their opponents outmatched if they were playing as they could. The Warriors-specific model ended up predicting home team wins with an accuracy of 78.5%, yielding an increase in accuracy of ~12.7% as compared to the initial stepwise logistic regression.

The Orlando Magic, however, struggled to maintain sustained success during this time period. They went through six different coaches, and only had two total playoff wins over this span. What stood out the most when observing the Magic model, was that five of the top nine variables in terms of variable importance, were “away team” variables. This would indicate that the Magic specific model paid more attention to the signals from the away team (non-Magic) statistics. This, intuitively, makes sense as they were not a good team and thus were more likely to face stronger teams on a nightly basis.

A graph with blue and red dots

Description automatically generated

*Figure 16: Comparison Between Overall and Magic Model*

Based on the coefficient values, home assists, home free throws, and away 3-point % were the most divergent from the overall model. In terms of accuracy, we once again observed increased accuracy on the individual Magic model, accurately predicting home wins 73.7% of the time, a 7.9% increase from the initial stepwise logistic regression model.

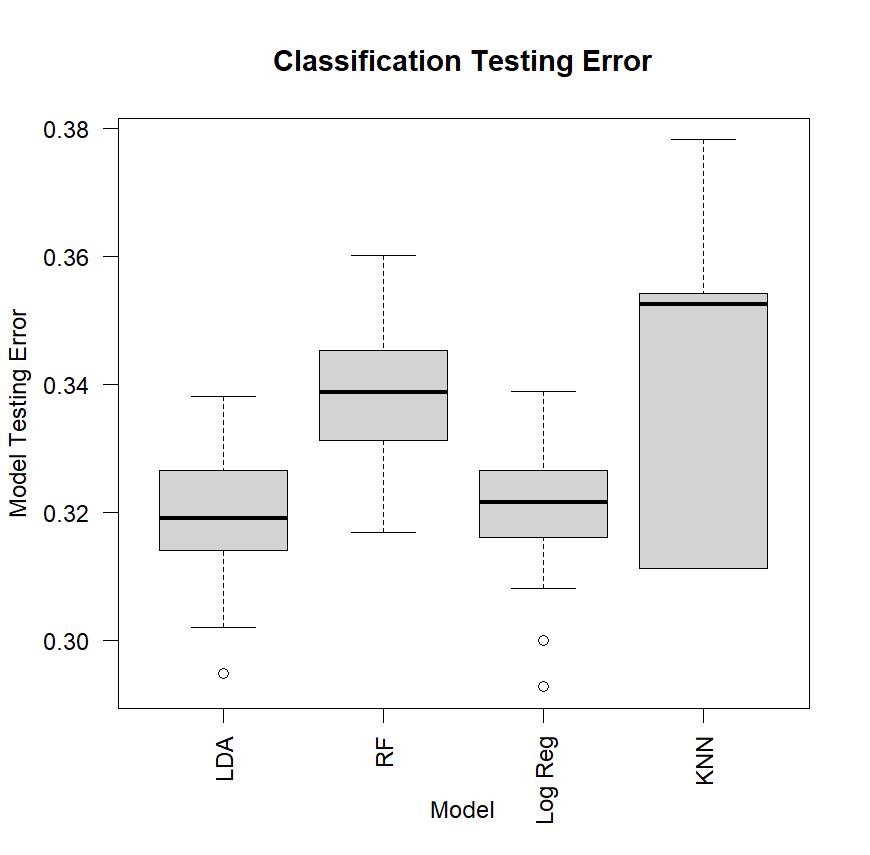
With the uptick in 3-point shooting that the NBA has observed over the last 20 years, it is no surprise to see the variables around 3-point shooting appear in the team specific models. Additionally, we see increased model performance when creating team-specific models, in both comparing to our league wide model, but also comparing to public research around predicting NBA games. This would indicate that, in the case of this data, team specific signals are being better captured in modeling when comparing to models that represent the league as a whole.

**Results**  
Given the concern in potential bias and overfitting of the initial model implementation, we used cross validation to confirm our choice of optimal model for classifying wins. We ran a Monte Carlo simulation that ran through 150 iterations of a random split between training and testing data for each model. We then averaged the testing error for each iteration of each model and compared the error and its variance. These results are available both as a box plot and as a table below.



*Figure 17: Cross Validation Error and Variance*

Visually inspecting the plot and table, the results from the cross-validation performance are very close between each of the models. The best model appears to be the linear discriminant analysis model. But the question of whether there was statistically superior performance of the model needs to be tested.



*Figure 18: Box Plot Error*

We then ran the model through an ANOVA test to see if there was a demonstrably different model in the group. The null hypothesis would be there is no significant difference in mean error, while the alternative hypothesis would be one model does have a significantly different mean error.

Using the aov() function, we saw that the test for the mean error produced an F value of 0.00013, which is significant at our chosen threshold of .05. We took this to mean that there was a significantly superior model out of our set. The alternative hypothesis would now be that the best performing model from an error perspective, the linear discriminant model, was the best out of the set.

We then developed a pairwise t-test of that hypothesis, comparing the linear discriminant model against all others. This resulted in significant differences between the linear discriminant model and every other, with the 95percent confidence interval not including 0 for each comparison. So- we can reasonably assume that the linear discriminant model was superior in terms of mean error.

**Findings and Conclusion**

Overall, we were able to determine that classifying whether NBA wins using basic statistics for home and away teams is best done using linear discriminant, at least when looking at the data we sampled. An interesting takeaway is that despite our reservations about potential bias in the linear discriminant model, it performed the best on the dataset. Expanding our scope to different types of prediction models, or expanding our dataset to include more statistics, could absolutely influence this decision. The dimensionality of the data could have something to do with the success of the linear discriminant analysis compared to logistic regression and KNN, as linear regression is an approach favored in high-dimensional data. Although we only totaled 26 factors, LDA could utilize each of them to capture variation.

Looking at our new optimal model choice, we can use this analysis in a couple of different ways. We can compare our result to sportsbook odds and pick appropriate games to bet on. A team could take and implement a version of the model to determine the statistics it wants a new player to contribute to improve its player contribution.

Extensions of this approach are very possible. Exploring related models like quadratic discriminant analysis could improve on the success of the linear discriminant model in terms of accuracy and statistical rigor, while investing more computational resources in the tuning of the expensive random forest could improve that approach. The success of the individual team models in terms of accuracy over the initial stepwise logistic regression implies another area of further research- could we create an ensemble model approach to tune models for each team to better predict success?

Even beyond that, there are still layers that were not captured in our analysis. Elements like player-specific performance and availability, scheduling challenges like games in a row away from home, or even matchup-specific features could help to improve accuracy. In taking this study a step further, we would look to capture these elements in our dataset, along with the team metrics. Based on our success here and the research conducted, we believe a combination of these elements could lead to continued increased accuracy, especially in the case of making team specific models. Despite the accuracy that we achieved, there is certainly more room to continue exploration in this space.

**Lessons Learned**

Overall, this class effectively showed the process of running analysis and eventually the application of modeling on a dataset. This project felt like less of a big leap and more of a step past what we did in the homework assignments, and we felt comfortable in that process of exploring the data, tuning the models, and presenting the results. We feel that even past all the models that we explored and built, the confidence in the process is what will be the best component of this class for us. The layering of the statistical rigor of the lectures and quizzes on top of the more expressive components of the homework helped build some of that intuition for the models themselves while maintaining the focus on application, which we enjoyed.

**References**

AFIT Data Science Lab. (n.d.). Random Forests · AFIT Data Science Lab R Programming Guide. Retrieved April 14, 2024, from afit-r.github.io website: <https://afit-r.github.io/random_forests#tune>

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15*(1), 27–33. <https://doi.org/10.1016/j.aci.2017.09.005>

Cao, C. (2012). *Sports Data Mining Technology Used in Basketball Outcome Sports Data Mining Technology Used in Basketball Outcome Prediction Prediction*. Retrieved from <https://arrow.tudublin.ie/cgi/viewcontent.cgi?article=1040&context=scschcomdis>

Entine, O. A., & Small, D. S. (2008). The Role of Rest in the NBA Home-Court Advantage. *Journal of Quantitative Analysis in Sports*, *4*(2). <https://doi.org/10.2202/1559-0410.1106>

Kassambara , A. (2018, March 11). Logistic Regression Assumptions and Diagnostics in R - Articles - STHDA. Retrieved from www.sthda.com website: <http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/>

Osken, C., & Onay, C. (2022). Predicting the winning team in basketball: A novel approach. *Heliyon*, *8*(12), e12189. <https://doi.org/10.1016/j.heliyon.2022.e12189>

Pius, E. H. and R. (n.d.). 9.4 Model assumptions | R for Health Data Science. In *argoshare.is.ed.ac.uk*. Retrieved from <https://argoshare.is.ed.ac.uk/healthyr_book/model-assumptions.html>

Statistical Analysis Primer. (n.d.). Retrieved from www.nba.com website: <https://www.nba.com/thunder/news/stats101.html>