NBA All-Star Prediction

Sid Lamsal

Tuesday, May 9, 2023

I. Background

Every year, 24 NBA players are named NBA All-Stars. This recognition is given to players who have shown exceptional play up to that point in the season. The exclusivity of the All-Star title causes a lot of discussion among fans over who will be named an All-Star. Conversations are fueled by the lack of a universal criteria for what qualifies a player as an All-Star. This project aims to determine the player statistics that are historically significant in determining NBA All-Stars. Additionally, this project aims to develop a classification machine learning model to predict NBA All-Stars.

II. The Dataset

The dataset for this project contains NBA player stats from the 2009-2010 season to the most recent 2022-2023 season. The data was scraped off Basketball Reference, a free, commonly used, NBA stats website [1]. Before preprocessing, the dataset contained 8980 sample units and 32 variables. Within the data, there were multiple sample units for the same player in the same year. This was due to players that were traded to a different team mid-season (separate rows for the player's stats for a specific team). In order to convert to row names, only the observation that contained the aggregate stats of the player was kept which reduced the number of observations to 7057. Removing all rows with NA value resulted in 6172 sample units. The general justification for removing the NA rows is that most NA rows were players who typically don't play much. In other words, they are part of the majority of players who are clearly not All-Stars, so their removal will not impact the model much since the model still has plenty of clearly non-All-Star players. In terms of columns, the columns that seemed irrelevant to being an All-Star, such as the player's team, were removed. A description of all 26 columns that were kept can be found in Table 1.

Table 1: Column description

Column Name	Data Type	Description (all numerical variables are per game stats)
G	Integer	Number of games played
GS	Integer	Number of games started (on the court when the game starts)
MP	Numerical	Minutes played
FG	Numerical	Shots made
FGA	Numerical	Shots attempted
FG.	Numerical	Shooting percentage
X3P	Numerical	3-point shots made
X3PA	Numerical	3-point shots attempted
X3P.	Numerical	3-point shooting percentage
X2P	Numerical	2-point shots made
X2PA	Numerical	2-point shots attempted
X2P.	Numerical	2-point shooting percentage
eFG.	Numerical	Effective field goal percentage (weighted efficiency)
FT	Numerical	Free throws made
FTA	Numerical	Free throws attempted
FT.	Numerical	Free throw shooting percentage
ORB	Numerical	Offensive rebounds
DRB	Numerical	Defensive rebounds
TRB	Numerical	Total rebounds (ball caught after missed shot)
AST	Numerical	Assists (pass leads to score)
STL	Numerical	Steals (steal ball from offense)
BLK	Numerical	Blocks (block shot attempts)
TOV	Numerical	Turnovers (offensive player loses the ball)
PF	Numerical	Personal fouls
PTS	Numerical	Number of points scored
AllStar	Categorical	1: All-Star, 0: not an All-Star (response variable)

The variables described in Table 1 are common box score statistics that are commonly used to assess a player's performance. There has been a rise in advanced stat usage such as PER and VORP, however, the stats in Table 1 are easy to understand and most frequently used by casual fans when discussing NBA All-Stars. As such, the variables should have ample predictive power.

To account for the lack of a test set, player data for the 2022-2023 season is reserved (535 observations). One cause for concern is the distribution of the response variable: AllStar. As shown in Figure 1, a large majority of sample units fall in class 0: not an All-Star.

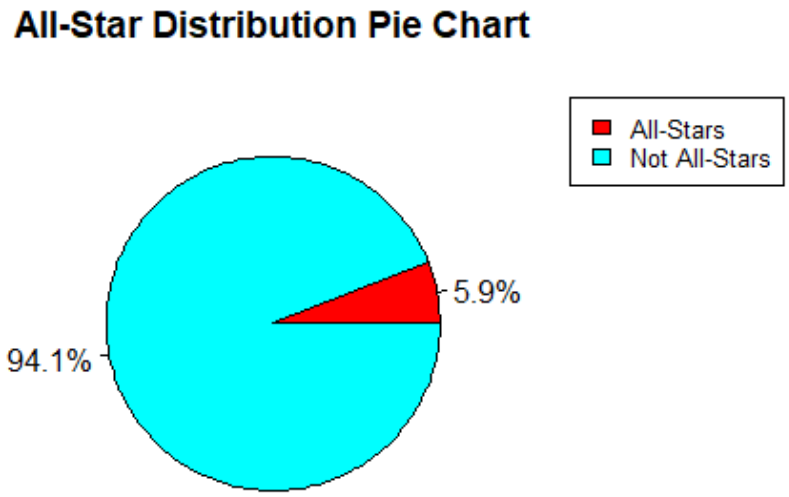

Figure 1: Pie chart of response variable distribution

Intuitively, the distribution makes sense since only 24 players are picked as All-Stars, barring injury. However, this may pose an issue since the model will be disproportionately exposed to non-All-Stars. To account for this, an oversampled training set was created which will be used alongside the original training set to determine if the original distribution skews the model.

The measure used to assess the effectiveness of a model will be accuracy. In any given year, players who produce All-Star level stats are not chosen, as such, it wouldn't be unusual for the model to have a few false positives (players predicted to be All-Stars that actually were not). This suggests that the best measure is true positive rate since when it is maximized, the false

negative rate is decreased. However, if the model labels all players as All-Stars, then true positive rate would be 100%, but the model would not be good. To account for the nuances of any particular measure, this project will consider both accuracy and true positive rate. Considering that 94.6% of players in the test set are not All-Stars, a model would have to be at least over 94.6% accuracy to have value. The test set was not oversampled to better simulate what a real test set would be like.

III. Models

i. Logistic model, LDA, QDA

The first model that was created was a regular logistic regression model with every variable. The optimal threshold of the regular model was .41. This threshold probability resulted in an optimal accuracy of .97 and a true positive rate of .925. The ACI was 793.12 and the AUC was .9875. Figure 2 is the corresponding ROC curve. Evidently, the curve is close to the top left which is ideal. The low ACI and high AUC suggest that the model performed very well.

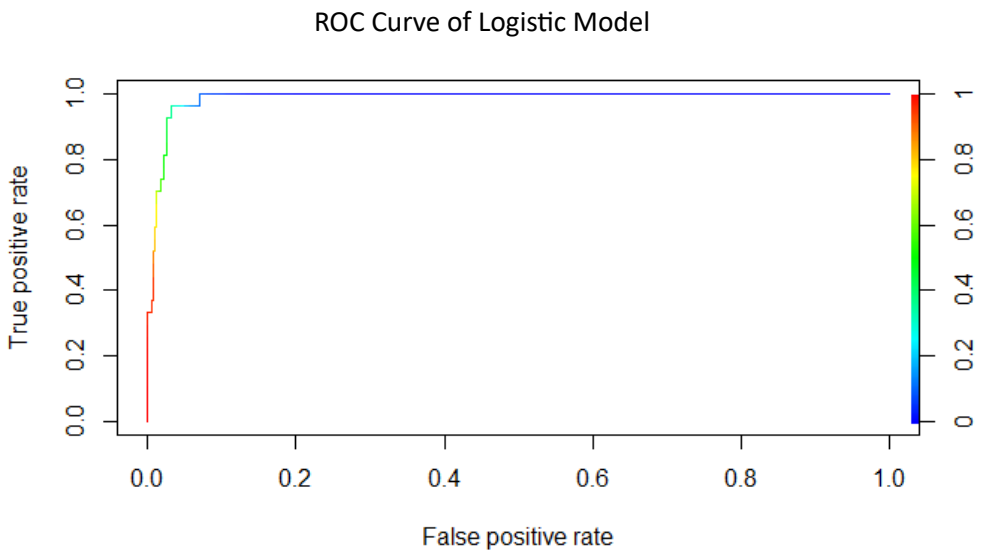

Figure 2: Regular full Log model ROC curve

Additionally, the confusion matrix was very promising. The model inaccurately predicted 2 All-Stars as non-All-Stars (false negative) and 12 non-All-Stars as All-Stars (false positive) as

shown in Table 2. Using 5-fold cross validation, the accuracy of this model was .97 meaning it also fit the training data well.

Table 2: Incorrect classification from the full model

Player	Model	Acutal
Devin Booker	1	0
Jaylen Brown	0	1
Jalen Brunson	1	0
Jimmy Butler	1	0
Anthony Davis	1	0
Darius Garland	1	0
James Harden	1	0
Jaren Jackson Jr	0	1
Zach Lavine	1	0
Kawhi Leonard	1	0
Dejounte Murray	1	0
Kristaps Porziņģis	1	0
Nikola Vučević	1	0
Trae Young	1	0

For an NBA fan, the players that the model incorrectly categorized are not outrageously wrong. All the players in Table 2 are some of the better players in the league. Notably, some of the missed players were All-Stars in the near past meaning they had sample units in the training set. This could have led the model to classify them as All-Stars since the player's stats would be similar. Secondly, All-Star selection takes place approximately ¾ of the way through the season. The dataset for this model is using the full season average stats. This means a player who played significantly better after the All-Star selection may have better season stats than pre-All-Star stats. One example from this model highlighted in Table 3 is Zach LaVine.

Table 3: Zach LaVine Post and Pre-All-Star selection [2]

	Points	FG%	3pt%	Rebounds	Assists
Before All-Star selection	24	46%	37%	4.9	4.1
After All-star	27	53%	39%	3.5	4.7
Season Average	25	49%	37.5%	4.5	4.2

As shown in Table 3, Zach had over all better stats after All-Star selection. This suggests that a dataset that uses season averages is not optimal. However, this conclusion is also not black and white. Although Zach played better, his season average is close to his pre-All-Star stats. His season stats still give a good idea of how he played before All-Star selection, so the model still has an accurate representation of how the player preformed pre-All-Star selection. A player would have to play significantly better for the remaining ¼ of the season to have a drastic impact on their season average. Additionally, the model likely has historical occurrences of a similar situation to learn from. That is to say, the dataset is not necessarily at fault for the model's mistakes, although it is a possibility.

The second model is the same as the full model, but with an oversampled training set. The best threshold for the oversampled model was .96 which resulted in an accuracy of 0.97 and a TP rate of 0.78. Before, the training data had significantly more non-All-Stars, so the model would tend to give lower probability for a player to be an All-Star. This is evident in the data as 31 players had a greater than .5 chance of being an all-star in the full model. On the other hand, the oversampled model has more All-Star observations and can be more "confident" that a player is an All-Star. For comparison, the oversampled model had 69 players with greater than .5 chance of being an All-Star. Figure 3 displays the difference in the probability distribution of the 2 models. Again, it is apparent that the oversample model tended to give higher probabilities for a player to be an All-Star due to its balanced training data. Only 27 players are actually All-Stars in the test set, so the oversampled model needed a higher threshold to weed out players so that its accuracy would be maximized.

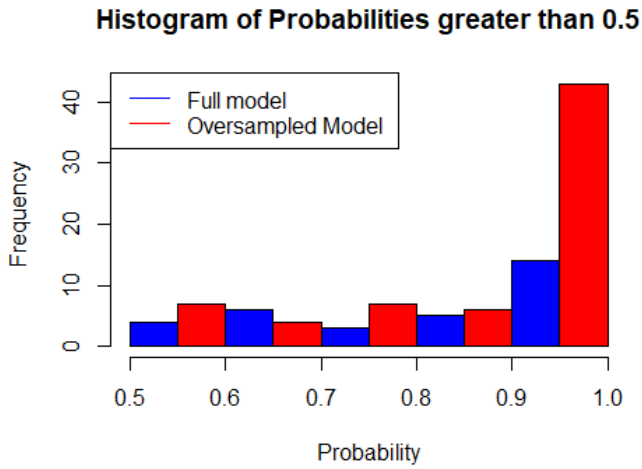

Figure 3: Histogram of Probabilities

The AIC of the model is 3247.9 and the AUC was .987. Figure 4 is the corresponding ROC curve. Again, we see that the curve is towards the upper left suggesting that the model performed well. The cross validation for the oversampled dataset was .95. The validation set is also balanced, so 95% percent accuracy is impressive as random guessing would be 50% accuracy. That is to say, the oversampled model performed well, but AIC would suggest that it is worse than the full model.

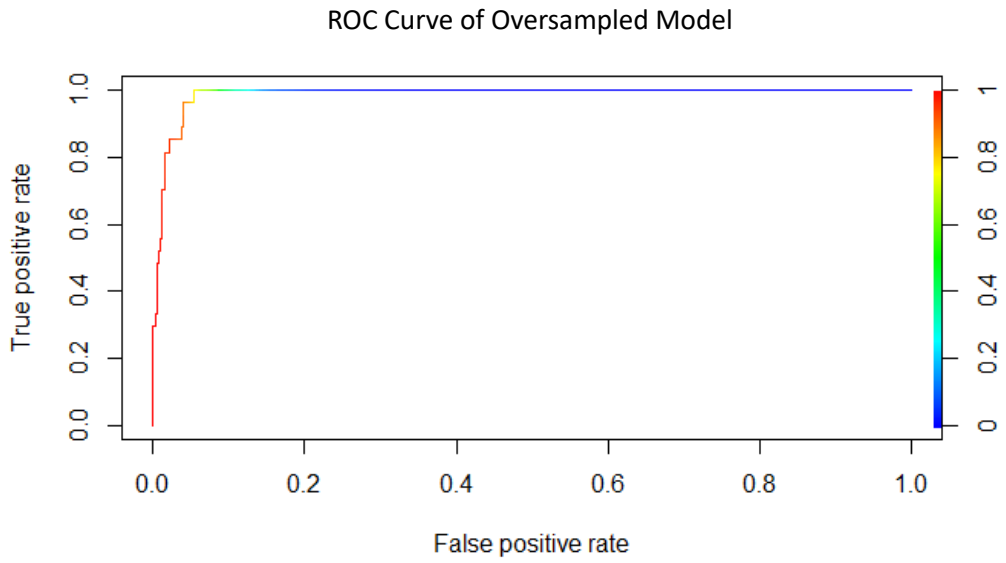

Figure 4: Oversampled full Log model ROC Curve

The confusion matrix of the oversampled model reflected the balancing of classes. The model predicted 14 players incorrectly: 8 non-All-Stars as All-Stars, and 6 All-Stars as non-All-Stars. For comparison, the full model predicted 12 and 2 respectively. Similarly, the players that the model incorrectly categorized were all some of the better players in the league. That is to say, the mistakes were not unusual as all the mistaken players are All-Star caliber or former All-Stars in previous years.

The good performance of the oversampled model suggests that the regular dataset and its imbalance is not overly problematic. Had the oversampled data set performed poorly, this would suggest that the previous model performed well due to being biased towards the majority class. However, since a balanced model had similar performance, it suggests that the distribution of classes is not inflating the performance.

Since all the variables are numerical, LDA and QDA models can be used to attempt to find the best model. Table 4 summarizes the results.

Table 4: Pre-variable selection model summary

Model	Accuracy	TP rate	AIC	AUC	C.V. error	Optimal threshold	Matrix
Full model	0.972	0.926	793.12	0.988	0.975	0.41	prediction 0 1 0 463 12 1 2 25
Oversampled model	0.972	0.778	3247.9	0.987	0.955	0.96	prediction 0 1 0 467 8 1 6 21
LDA	0.976	0.926	n/a	0.989	0.972	0.86	prediction 0 1 0 465 10 1 2 25

<table>
<tr><td>LDA Oversample</td><td>0.966</td><td>0.778</td><td>n/a</td><td>0.985</td><td>0.950</td><td>0.99</td><td>prediction<table><tr><td></td><td>0</td><td>1</td></tr><tr><td>0</td><td>464</td><td>11</td></tr><tr><td>1</td><td>6</td><td>21</td></tr></table></td></tr>
<tr><td>QDA</td><td>0.946</td><td>n/a</td><td>n/a</td><td>0.988</td><td>0.947</td><td>1.0</td><td>prediction<table><tr><td></td><td>0</td><td>1</td></tr><tr><td>0</td><td>475</td><td>0</td></tr><tr><td>1</td><td>27</td><td>0</td></tr></table></td></tr>
<tr><td>QDA Oversampled</td><td>0.95</td><td>n/a</td><td>n/a</td><td>0.988</td><td>0.942</td><td>1.0</td><td>prediction<table><tr><td></td><td>0</td><td>1</td></tr><tr><td>0</td><td>475</td><td>0</td></tr><tr><td>1</td><td>27</td><td>0</td></tr></table></td></tr>
</table>

Looking at the performance of the different models, we can see that the logistic (full) model and the LDA model performed relatively well whereas the QDA model struggled to predict for the test set. LDA has the same TP rate as the full model, but the accuracy is higher since it had fewer false positives. This displays the benefit of using multiple measures for evaluating the model. Ultimately, the confusion matrix is the most detailed and nuanced way of analyzing the model, but measures like TP rate and Accuracy give a nice overall summary of model effectiveness.

ii. Variable Selection

Up to this point, the models have utilized all available predictors. This might cause high variance in the models leading to overfitting. The results in a model that fits the training data well but is not generalized enough to predict for non-training sample units. As such, different variable selection methods can be used to possibly find an optimal set of variables.

One factor to consider before starting variable selection is variable interaction and non-linearity. Considering that the response variable is categorical, determining if the relationship between a variable and the response is non-linear is not possible with a simple graph. Intuitively, some level of nonlinearly is possible. For example, the difference between a 5-point scorer and a

15-point scorer is not as significant as a 15-point scorer and a 25-point scorer. A player going from 15 to 25 points could be the difference between an All-Star or not, while 5 points to 15 points is likely still not an All-Star player. However, this same argument can be applied to a 25–35-point scorer since the jump from 25 to 35 points does not increase the chance of being an All-Star as much as 15 to 25 points. This behavior does not fit a specific and interpretable nonlinear term, and the true relationship cannot be effectively graphed. As such, nonlinear terms will not be used for the interest of simplicity and interpretability, but they might be important for more in-depth modeling. On the other hand, interactions between variables can be determined by plotting. Figure 5 contains the plots of the variables with the highest correlation.

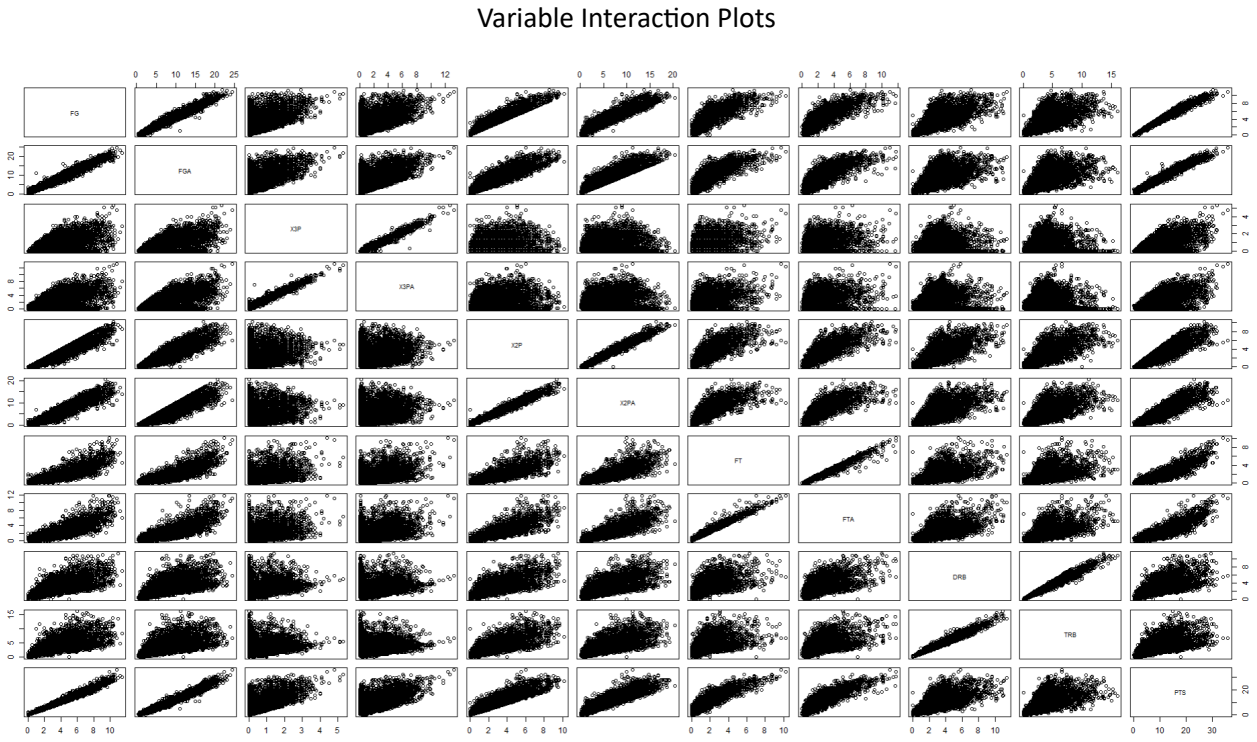

Figure 5: Variable Correlation Plot

At face value, Figure 5 suggests that there are multiple variables with high correlation like FG and FGA, X3P and X3PA, X2P and X2PA, FT and FTA, and DRB and TRB. However, these correlations likely do not require interaction terms because their interaction is already accounted for in the data. For example, FG and FGA measure field goals made and field goals attempted, but there is another variable in the data (FG.) which is FG/FGA, or field goal percentage. It's a similar story for X3P and X3PA, X2P and X2PA, and FT and FTA. For DRB

(defensive rebounds) and TRB (total rebounds), there is another variable (ORB: offensive rebounds) which is the difference between total and defensive rebounds. Figure 5 shows that PTS is correlated with many of the shooting statistics (FG, FGA etc.). To account for this, the model can use an interaction term for PTS and eFG. (Effective field goal percentage) which essentially measures how efficiently a player makes shots adjusted for the value of the shot. As such, it considers 2-pointers, 3-pointers, and free throws. The interaction term will make the biggest difference for players who score a lot of points, but don't do it efficiently. These players are less likely to be All-Stars since they score a lot but are not efficient.

With the interaction term, the regular logistic model increased its accuracy by one unit and the spread of false positives and false negatives was balanced to 7 and 6 respectively instead of 12 and 2. This suggests that the model was better at predicting when it considered how efficiently the players were scoring their points.

With the interactions accounted for, variable selection can be utilized to choose the optimal combination of variables to reduce over-fitting (variance). Variable selection will be performed using the regular logistic model. The log model is more compatible with different variable selection methods, so more observations can be collected on the best/most important variables for determining All-Stars. Additionally, since the log model performed similar with an oversample dataset versus a regular dataset, using the regular training set should not greatly affect the dataset. The different variable selection methods that will be used include algorithmically trying different combination of variables, backwards and forwards selection using AIC values, and Ridge and Lasso regression. Table 5 summarizes the results of the different variable selection methods.

Table 5: Variable selection summary

Method	Accuracy	TP	AIC	AUC	Variables	Optimal threshold	Matrix
Forward selection	0.974	0.74	792.58	0.988	All	0.64	prediction 0: 469, 6 1: 7, 20
Backwards selection	0.974	0.78	771.51	0.985	GS, FG, X3P, X2PA, eFG., FT, DRB, AST, STL, BLK, TOV, PF, PTS, eFG.:PTS	0.60	prediction 0: 468, 7 1: 6, 21
Subsets (using BIC)	.976	0.81	789.31	0.99	GS, MP, FG., X2PA, eFG., FT, FT., DRB, AST, STL, BLK, TOV, PF, PTS, eFG.:PTS	0.7	prediction 0: 468, 7 1: 5, 22
Ridge	.982	0.85	n/a	0.99	All	0.60	prediction 0: 470, 5 1: 4, 23
Lasso	.976	0.78	n/a	0.99	GS, X2P, FT, FTA, DRB, AST, STL, BLK, eFG.:PTS	0.57	prediction 0: 469, 6 1: 6, 21

As shown in Table 5, the Ridge regression model is significantly better than any other model and variable selection method. For comparison, the LDA model misclassified 12 sample units. The ridge regression cuts that down to 9. When the ridge regression was performed using the oversampled dataset, the confusion matrix was the same, which further supports the conclusion that the distribution of classes is not biasing the results. Overall, the variable selection results suggest that the variance of the model has been reduced. Although Ridge regression used all of the variables due to it utilizing L2 penalty, 13 variables were the most significant (had noticeably higher coefficients): FG., X3P, X3P., X2P., eFG., FT, FTA, FT., DRB, AST, STL, BLK, PF.

Many of the most important variables in the Ridge Regression are stats dealing with efficiency. This suggests that the Ridge model especially values the efficiency of the player. In addition, it utilized non-scoring sats like assists, steals, and blocks. Prioritizing these variables would help the model better predict the players who do not score much but contribute All-Star level production. One thing to note is that FT (free throws made) and related stats may have been a proxy for points scored. The best scorers in the league typically take and make a lot of free throws.

iii. Tree based methods

In addition to considering LDA and Logistic models, classification trees can be created to attempt to predict All-Star selections. The main advantage of the decision trees is that they can be plotted which can make for a more interpretable model. The classification trees will utilize Gini index as it is a better measure for tree growth than error rate.

The first classification tree uses the regular training set. This tree returned a classification error rate of 0.017 on the training set and 0.96 accuracy on the test set. The tree utilized 24 variables with 79 terminal nodes. The pruned version was similar. A second tree using the oversampled dataset returned a classification error rate of 0.07 on the training set and an accuracy of 0.87 on the test set. The oversampled tree used 8 variables and had 10 terminal nodes which is also similar to its pruned version. The oversampling had a much larger effect on the tree's accuracy on the test set. This makes sense as trees typically have more variance, so a change in the training set would make a bigger difference compared to a logistic model. Figure 6 illustrates the oversampled decision tree.

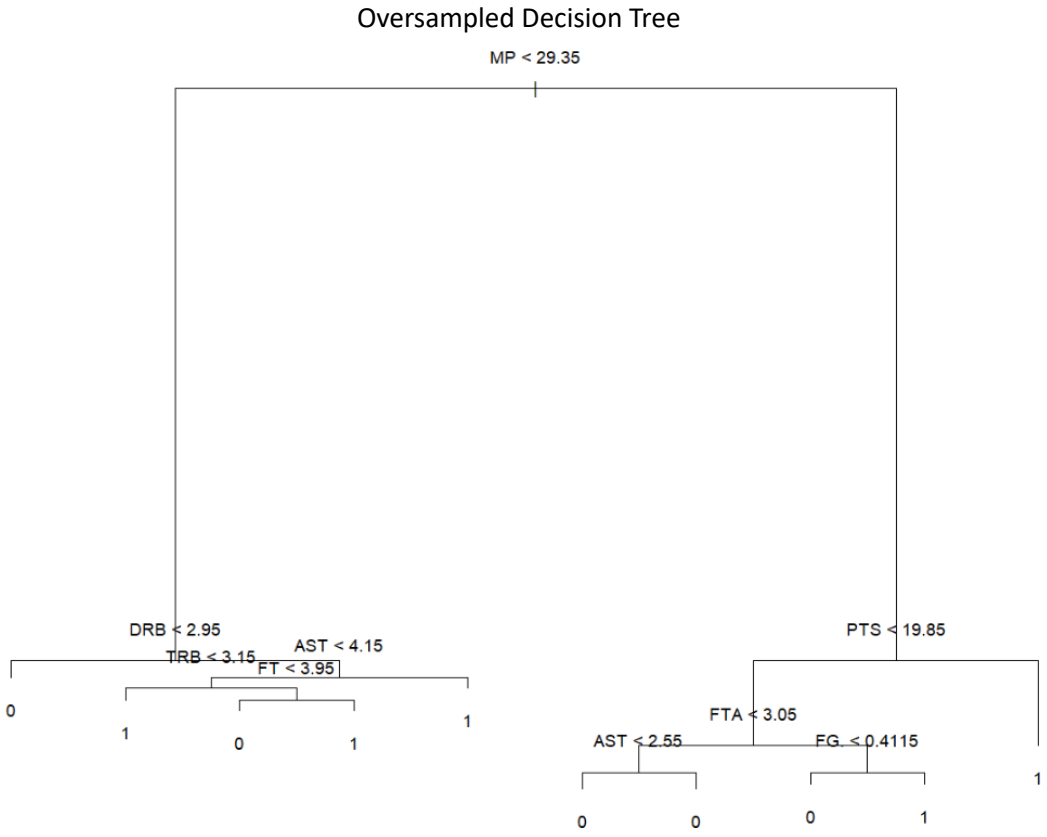

Figure 6: oversampled tree

The splits and variables used in the tree make intuitive sense when contextualized. For example, a basketball game is 48 minutes, so if a player plays 29 minutes and scores over 19.85 points, then that player is a relatively good player and could be an All-Star. The visualization of the tree provides an interpretable framework to understand how the model is arriving at its prediction. If the player plays for over 29 minutes, the tree then considers scoring stats like FTA and FG which suggests that a player that plays 29 minutes or more and scores over 20 points or scores efficiently is an All-Star. On the other hand, if a player plays less than 29 minutes, but contributes a certain number of rebounds (DRB and TRB), free throws (FT), and assists, then that player is also an All-Star. This illustrates 2 distinct groups of players: players who contribute through scoring and players who contribute through complementary stats.

As previously mentioned, trees typically have high variance. Bagging can help to reduce variance by creating multiple trees with bootstrapped datasets. The bagged tree using the regular dataset has an error rate of .028 on the training set and .96 accuracy on the test set along with .89 true positive rate. An equivalent tree using the oversampled dataset has an error rate of 0.0082 on

the training set and an accuracy of .96 on the test set along with a true positive rate of .85. Figure 7 and Figure 8 show the most significant variables in the regular and oversampled bagged trees.

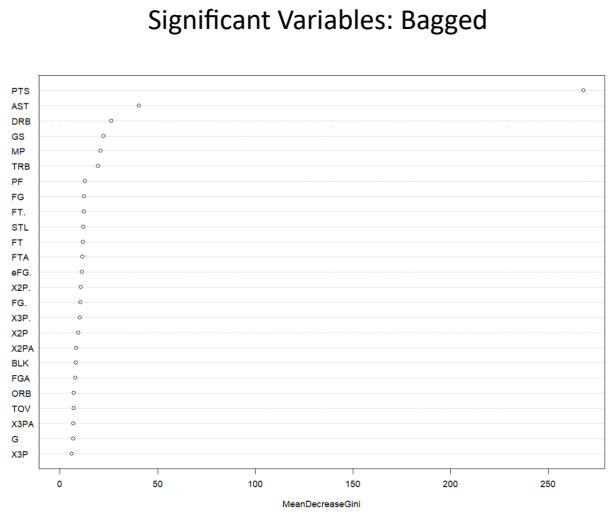

Figure 7: Bagged tree significant variables

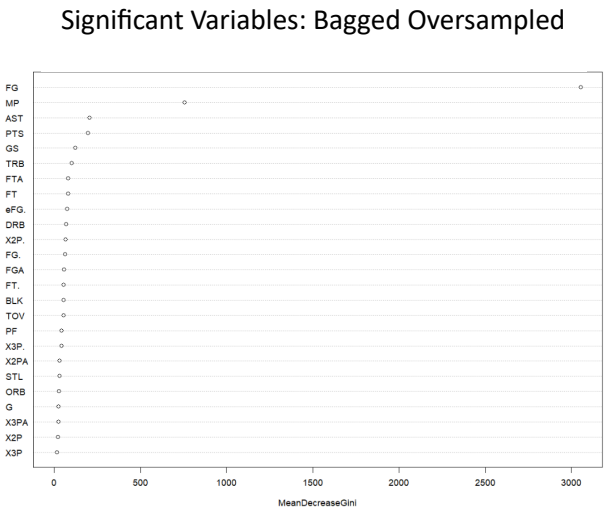

Figure 8: Oversampled significant variables

The plots show that PTS and FG were very strong predictors. This is good for prediction but undermines the effort to reduce variance since every bootstrap tree likely looked similar with PTS or FG as top predictors.

Random forest can be used to counter the effects of overly strong predictors. With random forest, each bootstrapped training set has a fraction of the variables. In this dataset, there are 25 predictors, so random forest will use the square root of the number of predictors (i.e., 5). The random forest model had an error rate of 0.027 on the training set and an accuracy of .97 on the test set along with a true positive rate of .926. Evidently, random forest removed some level of variance in the model as the model became noticeably better at predicting All-Stars (true positive rate). Figure 9 illustrates the significant variables of the random forest model.

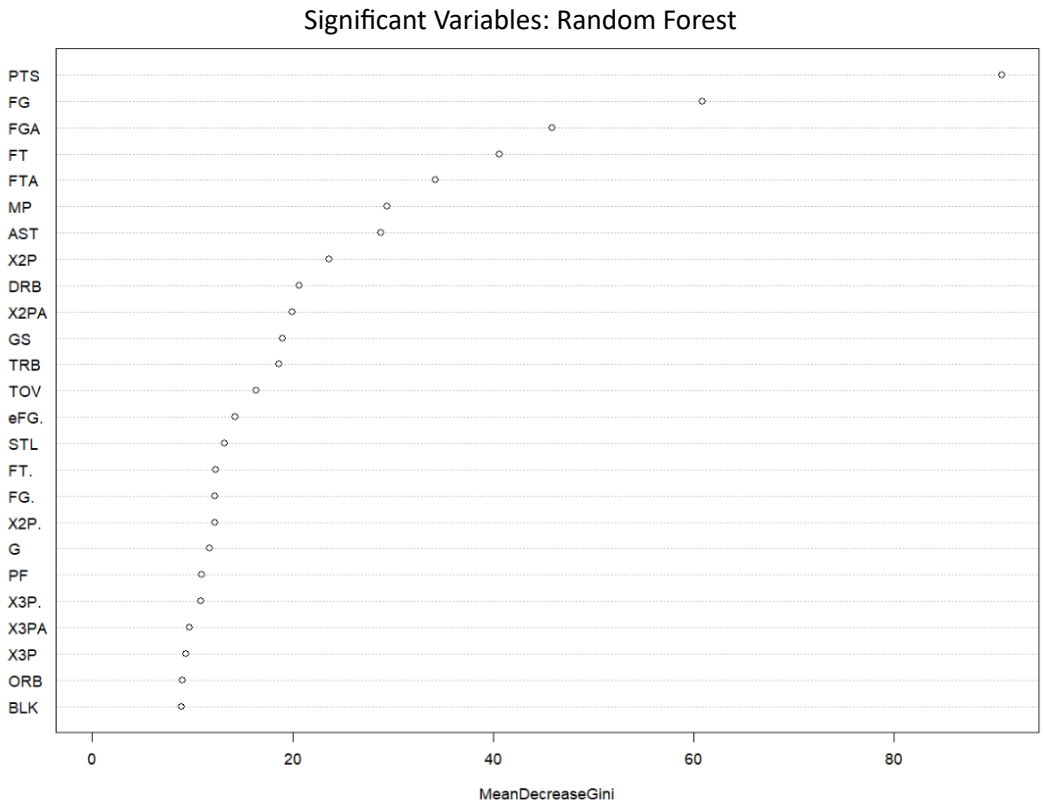

Figure 9: Random Forest significant variables

In the random forest model, variable significance was more balanced since different variables were used in different trees. The random forest model has similar performance to the LDA model.

Another method for improving classification trees, boosting, can be used to possibly improve performance. The boosted tree uses a shrinkage of .1 and depth of 3. Its accuracy on the test set was .972 and the true positive rate was .889. Evidently, the boosted model performed well-on par with the random forest model and the LDA model. However, between the 3, LDA had the least errors. Figure 10 shows the influence of different variables on the model. Again, we see that PTS is the most important predictor for determining All-Stars.

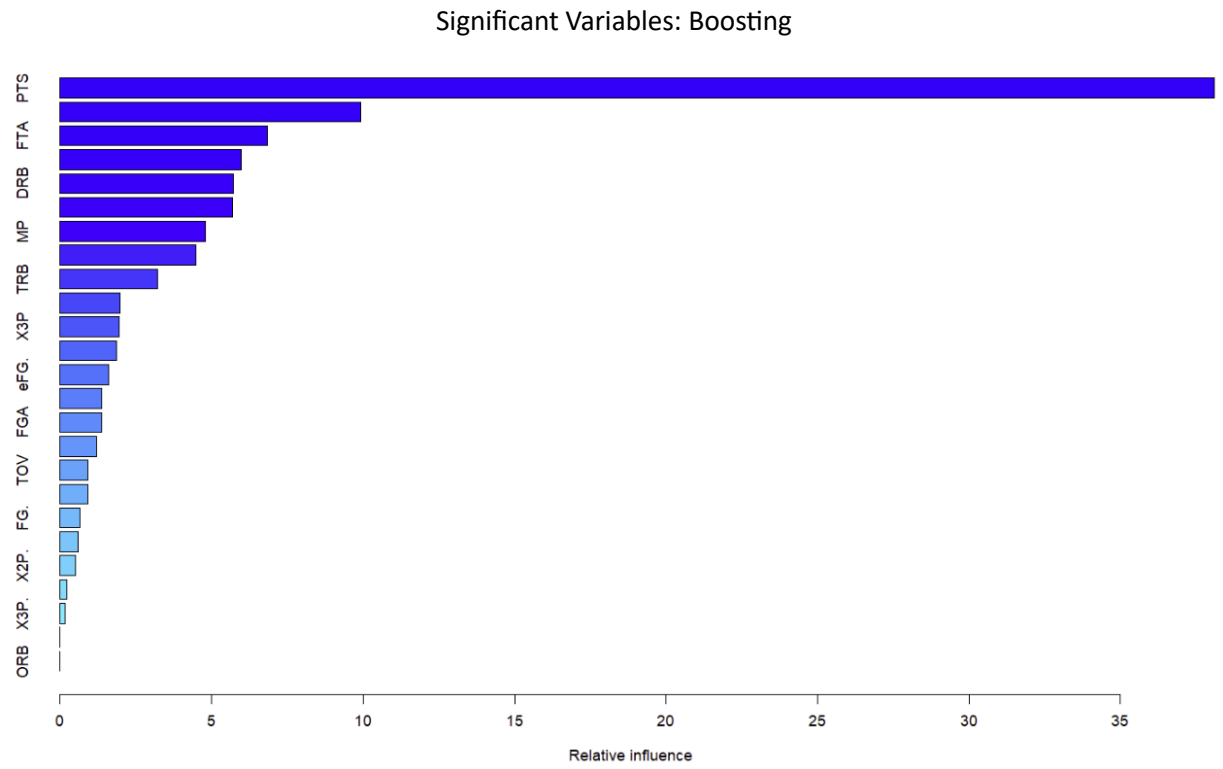

Figure 10: Boosted best variables.

To summarize, as shown in table 6, each of the tree-based methods had better accuracy on the test set while having lower accuracy on the training ser. The better test set performance suggests a reduction in variance, and the lower training accuracy suggests an increase in bias. This is in line with what was expected since tree-based methods predominantly help to lower the variance of decision trees.

Tabel 6: Tree and Tree Based Methods Summary

<table>
<tr><th>Method</th><th>Test Acc.</th><th>Test TP</th><th>Train Acc.</th><th>Used Variables</th><th>Matrix</th></tr>
<tr><td>Regular</td><td>0.958</td><td>0.889</td><td>0.983</td><td>All except FTA</td><td>prediction<table><tr><td></td><td>0</td><td>1</td></tr><tr><td>0</td><td>457</td><td>18</td></tr><tr><td>1</td><td>3</td><td>24</td></tr></table></td></tr>
<tr><td>Bagging</td><td>0.964</td><td>0.889</td><td>0.973</td><td>Top 5: PTS, AST, DRB, GS, MP</td><td>prediction<table><tr><td></td><td>0</td><td>1</td></tr><tr><td>0</td><td>460</td><td>15</td></tr></table></td></tr>
</table>

					1 | 3 | 24
Random Forest	0.970	0.926	0.973	Top 5: PTS, FG, FGA, FT, FTA	prediction | | 0 | 1 | | 0 | 462 | 13 | | 1 | 2 | 25 |
Boosting	0.972	0.889	0.978	Top 5: PTS, AST, FTA, GS, DRB	prediction | | 0 | 1 | | 0 | 464 | 11 | | 1 | 3 | 24 |

iv. SVM

The last form of classification that will be attempted for this analysis will be support vector machine. Unlike trees, SVM uses a hyperplane which separates sample units into different classes. Support vectors are the data points which are on or within the hyperplane. Support vectors effectively softens the requirement that all data points must be perfectly classified.

The two things that need to be specified for SVM are the kernel and cost. Kernel determines the type of hyperplane. For example, the hyperplane could be polynomial or radial to handle non-linear separating space. Cost defines the cost of violating the margins. Larger cost narrows the margins of the hyperplane meaning fewer support vectors on the margin or violating the margin. Essentially, lower cost allows for more errors. Since cost serves as a tuning parameter, its optimal value can be determined with cross validation. With a radial kernel, the gamma value also needs to be optimized. Table 7 summarizes the results of different kernels of SVM at the optimal tuning parameters.

Table 7: SVM summary

Kernal	Cost	Accuracy	TP rate	Training Error	Matrix
Linear	1	0.968	0.852	0.030	prediction | | 0 | 1 | | 0 | 465 | 10 |

<table>
<tr><td></td><td></td><td></td><td></td><td></td><td>1 | 4 | 23</td></tr>
<tr><td>Polynomial</td><td>0.1</td><td>0.974</td><td>0.852</td><td>0.028</td><td>Prediction
| | 0 | 1 |
| 0 | 466 | 9 |
| 1 | 4 | 23 |</td></tr>
<tr><td>Radial</td><td>5, gamma=.5</td><td>0.946</td><td>0.074</td><td>0.047</td><td>prediction
| | 0 | 1 |
| 0 | 473 | 2 |
| 1 | 25 | 2 |</td></tr>
</table>

Table 7 shows that none of the kernels were particularly good at predicting. The best SVM (polynomial kernel) was relative to the log models without variable selection. When compared to the bagging and random forest, SVM had similar performance on training and test sets. The polynomial kernel is only slightly better than the linear, but the fact the best separating hyperplane is polynomial suggests that there might be some level of non-linear relationship between the response and the predictors. This was hypothesized earlier in the variable selection section.

IV. Interpreting Model Results.

Table 8 contains the best model from each of the previous sections. Out of the following, it is evident that the ridge regression model was the best model. A noticeable trend in the models is that the variable selection models typically performed better on the test set. The largest number of misclassifications was 13 for the variable selection models which is less than or equal to the best tree and SVM. SVM and trees typically have high variance, and variable selection is used to reduce variance, so it's evident that the tree and SVM suffered from high variance.

Table 8: Summary of models

Model	Accuracy	TP rate	Train Acc.	Matrix
LDA	0.976	0.926	0.970	prediction 0: 465, 10 1: 2, 25
Ridge	.982	.852	0.974	prediction 0: 470, 5 1: 4, 23
Boosted	0.972	0.889	0.978	prediction 0: 464, 11 1: 3, 24
Polynomial	0.968	0.852	0.972	prediction 0: 463, 9 1: 4, 23

Table 8 presents a fairly clear winner considering Ridge had 3 less misclassifications than the next best model. However, there are a few more factors worth considering: player position and conference locks, and practical severity of errors.

A factor that the models did not account for is the position and conference locks for the All-Star team. There are 24 players voted as All-Stars, but that is split between 12 for the Eastern Conference and 12 for the Western Conference. Out of the 12, there can be 4 guard players, 6 forward players, and 2 wild cards [3]. The terms “guard” and “forward” refer to the position that the player plays. Lastly, there may be only 24 spots on the team, but there can be more players who were named All-Stars but could not play on the official team since they were injured. These restrictions can be applied to the model to more accurately depict its predictive power. Since a

real-time prediction would not have data on injury replacement players, the model will be evaluated on its ability to correctly predict the initial 24 All-Stars with the restrictions in place. Table 9 summarizes the results.

Table 9: Position and Conference Locked Predictions

Predicted		Actual [3]	
East	West	East	West
Forwards: Joel Embiid, Giannis Ante., Jayson Tatum, Kevin Durant, Jimmy Butler, Julius Randle	Forwards: Nikola Jokić, LeBron James, Anthony Davis, Domantas Sabonis, Zion Williamson, Lauri Markkanen	Forwards: Giannis Ante., Kevin Durant, Joel Embiid, Jayson Tatum, Bam Adebayo, Julius Randle	Forwards: LeBron James, Nikola Jokić, Lauri Markkanen, Zion Williamson, Paul George, Domantas Sabonis
Guards: Trae Young, James Harden, Donovan Mitchell, Kyrie Irving	Guards: Luka Dončić, Shai Gilgeous, Damian Lillard, Stephen Curry	Guards: Jaylen Brown, DeMar DeRozan, Kyrie Irving, Donovan Mitchell	Guards: Ja Morant, Stephen Curry, Luka Dončić, Shai Gilgeous
Wild Cards: Pascal Siakam, DeMar DeRozan	Wild Cards: Ja Morant, Anthony Edwards	Wild Cards: Jrue Holiday Tyrese Haliburton	Wild Cards: Damian Lillard, Jaren Jackson Jr.

The method by which the predictions were determined was using probabilities. Essentially, the top 6 forwards with the highest probability of being All-Stars were chosen. This essentially penalizes the model for not only misclassifying, but also the specific probabilities it assigned to each player. As shown in Table 9, in a more realistic application, the Ridge regression incorrectly predicted 6 players out of the 24 for a total true positive rate of 0.75. The yellow highlighted players are those who the model predicted as All-Stars but were not. The red highlight represents players who were All-Stars but not predicted as such by the model.

The players who the model failed to predict or the players that it incorrectly predicted are relatively the same caliber of player. That is to say, the severity of the errors is not an issue. The highlighted players are all All-Star caliber players. In fact, a USA Today Sports article titled "The biggest 2023 All-Star snubs" [4] includes all 6 players that the model incorrectly predicted to be All-Stars. Additionally, as illustrated in figure 11, the players that the Ridge model incorrectly predicted were among the most missed among the 4 top models.

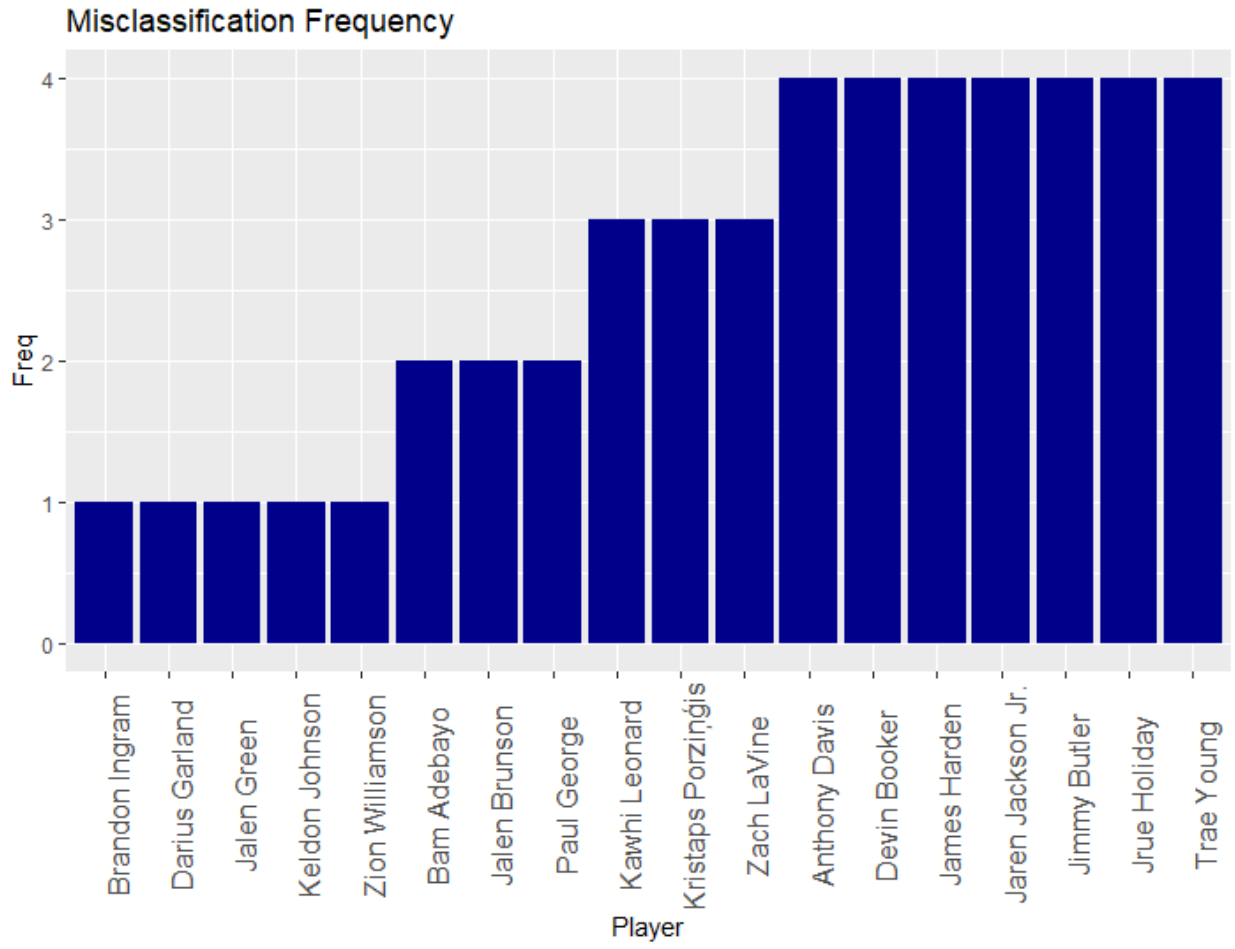

Figure 11: Misclassification Frequency

V. Final thoughts

The original goal of this project was to explain the most significant variables for determining All-Stars, and to predict All-Stars. The results have proven to be a mixed bag. For instance, PTS has consistently been a top predictor in the trees and variable selection model, however, it is difficult to say that PTS was the best predictor when the best predictive model didn't have PTS as a top 13 predictor. Realistically, there is no doubt that the number of points scored is a very important factor for determining All-Stars. Beyond PTS, different models all prioritized different stats. As for predictive power, the best model in this project would have gotten 6 out of 24 players incorrect.

The All-Star voting is not known for its clear consistent criteria. The top 10 players are voted in by fan votes. Additionally, the rest of the players are chosen by head coaches. Fan voting has been criticized in the past for turning All-Star selection into a popularity content. For example, for the past 2-3 years, LeBron James has not been the best player in the league in many people's eyes. However, he has received the most votes by far of any player during All-Star voting. In terms of coach votes, the top priority of most coaches is winning games, a factor which is not considered in this analysis. Additionally, as mentioned earlier, the data for this analysis contains season stats, not pre-All-Star selection stats. As such, 3 things could have been done to possibly better the dataset: include fan vote totals, include wins, and use pre-All-Star Stats.

Although many awards in the NBA seem to lack consistency, that is arguably one of the biggest factors in the NBA's success. Fan discourse on social media and the unexpected nature of the sport is what keeps fans interested and in large part makes the NBA successful. Part of the beauty of the sport is talking basketball, making predictions, being wrong, and complaining on Twitter. It's all a part of enjoying basketball, and many fans wouldn't want it any other way.

References

[1] Sports Reference LLC. (2023). Basketball-Reference.com - Basketball Statistics and History. https://www.basketball-reference.com/

[2] NBA Media Ventures, LLC. (2023). NBA Advanced Stats. https://www.nba.com/stats/players/traditional

[3] NBA Media Ventures, LLC. (2023). *2023 All-Star Game: Jayson Tatum makes history, leads Team Giannis to victory.* https://www.nba.com/news/2023-all-star-draft-all-star-game-explainer

[4] Urbina, Frank. (2023). *The biggest 2023 All-Star snubs.* HoopsHype, USA TODAY Sports. https://hoopshype.com/lists/2022-23-nba-biggest-all-star-snubs-season/