NBA All-Star Prediction Project Final Report

Problem Statement

Every year, 30 NBA players are selected as All-Stars, and although the All-Star game itself is an exhibition without much meaning, being selected as an All-Star has major benefits for NBA players. Many players have clauses in their contract that award them more money if they are selected as an All-Star, and players finishing their rookie contracts will be eligible for a higher max contract if they started in two All-Star games. Additionally, brand power and sponsorships are affected, and future awards like Hall of Fame decisions are affected by All-Star appearances.

Being able to accurately predict All-Star players would be important for many constituents. Teams would value this information when trading for new players. Sponsors will want to know if they should expect a player to make the team. Sports betting has become prolific, and companies like FanDuel make money off bets on All-Star selections.

Furthermore, they is controversy as to how players are selected for All-Star teams, especially since 50% of the decision is left to fan voting. Although, in the moment, people recognize popularity contests can influence the voting, years later people will look back and assume the all-stars fully warranted their placement. Given so much is at stake for players, teams, sponsors, and bettors, perhaps the NBA may decide to test out a less biased selection process in the future. Presumabley, part of the decision could be left to a predictive model like the one in this project.

By using a mix of player, team, and advanced stats data from 1979 to present[1], I created a tool to predict whether a player would make the All-Star team in a particular season. The same statistics are updated on a regular basis, so the model could be used to predict All-Star selections before they are made mid-way through the season. I employed exploratory analysis, scaling, hyperparamter tuning, and several supervised models before selecting the one which was most accurate.

After selecting the 33 features that were most impactful, my Random Forest Classifier was able to achieve a macro F1 score of 0.88. The model can be used to predict all-stars with new datasets, as I did with the 2023 season data.

Data Wrangling

The original data set from Sumitro Datta came in three tables: Player Totals, which included stats like total rebounds and assists, Advanced, which included advanced statistics like Value Over Replacement Player, and Team Summaries which I used for wins, losses, and fan

[1] Sumitro Datta formatted this data from Basketball-Reference.com in csv files on Kaggle: https://www.kaggle.com/datasets/sumitrodatta/nba-aba-baa-stats

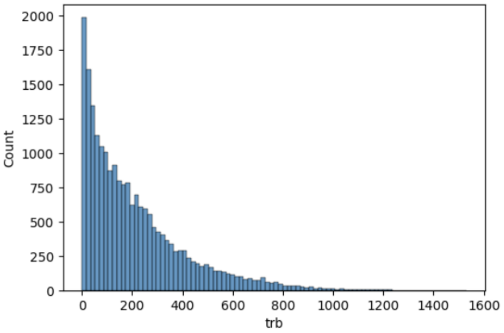
attendance. I created a win percentage column from wins and losses, and also calculated the average attendance per game for each season. I removed all data from before 1979 because that was the first season that included a three-point arc.

There were some null values that needed to be imputed. For example, players who were traded mid-season had their team labeled as ‘TOT’ for ‘two or more teams.” I gave those players a 50% win percentage. Some teams did not record the attendance at home games, so I imputed the average attendance that season into those null values. There were some players with zero minutes of playing time, which led to various advanced stat calculations dividing by zero. I dropped those rows since players who didn’t play would not be consider for all-star awards. The remaining null values were also caused by statistics dividing by zero, such as a center who never shot a three-pointer and therefore had a NaN for their 3-point percentage. I filled those null values with zero. The one exception was turnover percentage, for which I imputed the mean since 0 would have been an excellent score in that category.

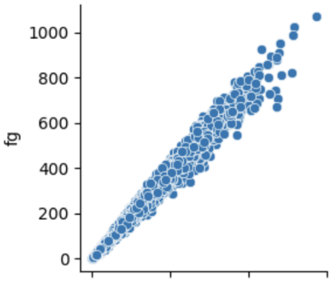
The final shape of this dataset was 22462 rows and 56 columns.

Exploratory Data Analysis

During exploratory data analysis I checked the distributions of my features using histograms and saw that the basic statistics were heavily skewed to the right, since the majority of players had low values for a given statistic, such as total rebounds (shown to the left). When checking for outliers, it also became apparent that some players with very few games or minutes played had extreme values on different advanced statistics. For example, one player had an offensive box plus-minus score of 199.4, (suggested his team would score an additional 199 points per game if he were to play the entire game) but had only played 1 minute that season. I decided to drop all rows with fewer than five games played, since those players would not be considered for All-Star games anyway.



It also became apparent that many statistics were highly correlated, which would later cause issues with modeling. For example, there was much overlap between statistics like total rebounds, total defensive rebounds, and total offensive rebounds. Total field goals and total points scored had a correlation coefficient of .99 (shown in the plot to the right). I trimmed down my features to about 35 after checking for correlations between them.



Exploratory analysis also showed that my classification data was imbalanced; less than 3% of the records led to All-Stars. I would need to resample my data depending on which model I was using.

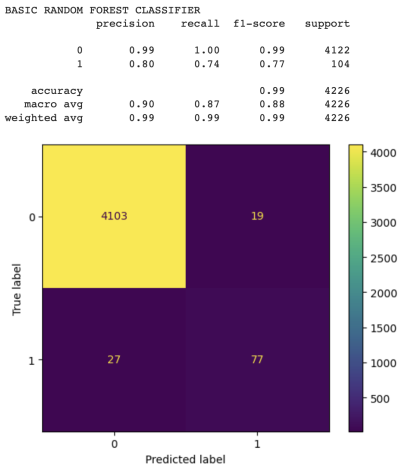
Preprocessing

Most of my data was numerical, but I had a few features such as ‘team name’ and ‘position’ which would cause issues when modeling. I dropped the team column and one-hot encoded the position column. I scaled the data with a standard scaler, and later I used SMOTE and Tomek Links methods to deal with the imbalance between all-stars and non-all-stars.

Modeling

I test four different types of classification models: Random Forest, Logistic Regression, Naive-Bayes, and Support Vector Machines. I used a mix of grid searches and random searches to help with hyperparameter tuning, but ultimate the Random Forest Classifier with default parameters and 300 estimators performed the best.

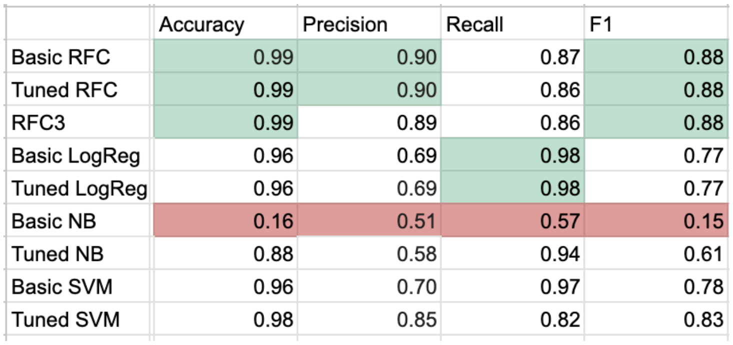
In general, the Random Forest Classifiers outperformed the other models. The first RFC I tried successfully classified one more record than the other forests.



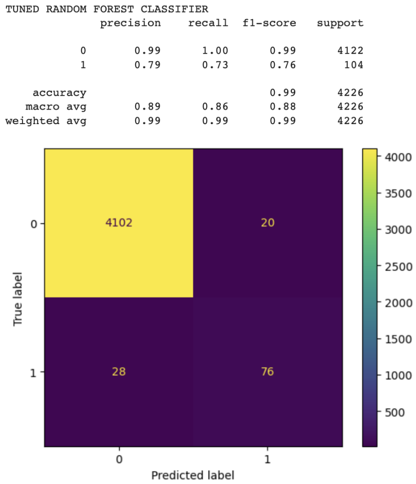
```
BASIC RANDOM FOREST CLASSIFIER
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      4122
           1       0.80      0.74      0.77       104

    accuracy                           0.99      4226
   macro avg       0.90      0.87      0.88      4226
weighted avg       0.99      0.99      0.99      4226
```



	Accuracy	Precision	Recall	F1
Basic RFC	0.99	0.90	0.87	0.88
Tuned RFC	0.99	0.90	0.86	0.88
RFC3	0.99	0.89	0.86	0.88
Basic LogReg	0.96	0.69	0.98	0.77
Tuned LogReg	0.96	0.69	0.98	0.77
Basic NB	0.16	0.51	0.57	0.15
Tuned NB	0.88	0.58	0.94	0.61
Basic SVM	0.96	0.70	0.97	0.78
Tuned SVM	0.98	0.85	0.82	0.83



```
TUNED RANDOM FOREST CLASSIFIER
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      4122
           1       0.79      0.73      0.76       104

    accuracy                           0.99      4226
   macro avg       0.89      0.86      0.88      4226
weighted avg       0.99      0.99      0.99      4226
```

Performance in the 2023 Season

When the model was applied to the 2023 regular season data, it predicted 16 players as All-Stars. Thirteen of the sixteen predictions were correct. My model’s three false positives for the 2023 season were James Harden, Jimmy Butler, and Anthony Davis, all of whom received some press coverage for being “snubbed” since they were not selected as All-Stars this season.

The model undervalued certain popular players who were injured most of the season. For example, Kevin Durant and Zion Williamson were indeed selected as all-stars despite only playing 47 and 29 games (of the possible 82), respectively. It is worth noting that I ran my model off of the entire season’s date, when voters had to make their choices mid-way through the season. Thus, the timing of Durant’s and Williamson’s missed games could be a factor. In the future, the NBA plans to require players to play for 20 minutes in at least 65 games to be eligible for All-Star honors.

Future Research

It would be interesting to see if data from older seasons help or hinder the model. The game has changed over the years, and, notably, the three-point shot has become a staple for successful players while mid-range jumpers have become less and less effective.

In addition, the selection process for All-Stars has changed over the years. Some years, the restrictions on positions are different, and fans were not always allowed to vote.

Finally, there might be other data that could improve my model. For example, one would expect players with larger numbers of social media followers to receive more votes, and large sponsorship deals could be a signal for all-star selections. My model only analyzed a single player-season of a statistics at a time, but many voters are likely biased by past seasons. Did Kevin Durant make the All-Star team in 2023 (despite playing very few games) because the basketball community was already well familiar with his achievement from past seasons? These kinds of factors could help improve model accuracy.