Can Machine Learning be used to select the NBA All-Star teams?

Joshua Adebayo

1 Introduction

In this document, I will discuss my approach, results, analysis and motivations behind answering the research question 'Can Machine Learning be used to select the NBA All-Star (AS) teams?'

I have trained a multi-variate logistical regression to decide whether a player should make it onto the All-Star team given their statistics throughout the year. By using machine learning (ML) it should remove any bias and only the best players get selected to participate

2 Background

2.1 All-Star Game History

The first annual NBA All-Star game was in March of 1951, thought of by NBA President Maurice Podoloff, and publicity director Haskell Cohen as an attempt to increase the league's popularity by displaying the best players compete like never before [1]. Sports writers were told to vote for 10 players from each conference and those with the most votes were selected to play. The event was very successful, a crowd of over 10,000 came to watch the All-Star game, much higher than the season's average attendance of 3,500 [3]. 72 years later in 2023, the event is still going strong, with 4.3 million viewers worldwide.

From 1980 the All-Star team selection was decided by fan votes but in 2017, the voting system changed to a (50%, 25%, 25%) split of fan votes, player votes, and trusted media personnel. This change was done since players who play in large cities, for teams with a large following such as the New York Knicks, LA Lakers, and Golden State Warriors could get carried into the All-Star game because of their team's fan base. In 2016, some known mediocre players crept into the big game which forced NBA Commissioner Adam Silver to come out with this new system.

2.2 Why does selection matter?

Having a fair way to select players from a fan's perspective, only matters from an entertainment standpoint. Seeing the league's premier players go head-to-head is going to be high-level basketball. From the player's side, however, making the AS team is very profitable. There have been cash prizes of up to $100,000 given out by the league for just attending the event and additionally, some players have received substantial bonuses from their respective teams for being selected. After last year's game Jaylen Brown, Julius Randle, Domas Sabonis, and Jrue Holiday received $ 1.5 million, $ 1.3 million $ 1.1 million, and $ 345k [2] from their teams on top of their payment from the league. Additionally, the exposure from being invited to the game brings a lot of endorsements for players typically with brand deals for footwear and apparel. An example of this is each year, the All-Star players with Nike get to model the All-Star colorway for their popular basketball trainers.

Although the new voting system minimizes the bias from fans, NBA players are human too, nothing is stopping the players from just voting for their friends/teammates. Having a way to quantitatively pick the best players for the AS game removes bias completely. For fans, they would be able to truly watch the best compete and for the players, they will be able to be fairly compensated for their talents and hard work.

3 The Data

I got my dataset from kaggle, a trusted website https://www.kaggle.com/datasets/owenrocchi/nba-advanced-stats-20022022. The dataset had 12211 entries from 2002 - 2022 but I chose to use the data from 2017-2020 for continuity of the voting system used this reduced the number of entries to 4135 with 141 All Stars and 3994 others. The dataset had 29 columns initially but I chose to drop 7 of them (Year, Team, player name,' Unnamed: 0', Position, and games played) to leave only the parameters related to player performance. The remaining parameters were as follows:

3.1 Parameters

Minutes played, PER is player efficiency rating, TS% is true shooting percentage, 3PAr is 3-point attempt rate, FTr is free throw attempt rate, ORB% is offensive rebound percentage, DRB% is defensive rebound percentage, TRB% is total rebound percentage, AST% is assist percentage, STL% is steal percentage of total steals, BLK% is block percentage of total blocks, TOV% is turnover percentage, USG% is usage percentage, OWS is offensive win shares, DWS is defensive win shares, WS is win shares, WS/48 is win shares per 48 minutes, OBPM is offensive box plus/minus, DBPM is defensive box plus/minus, BPM is box plus/minus, and VORP is the value over replacement player.

4 Dimension Reduction

Before fitting a model, it was clear that the number of parameters must be greatly reduced. I chose to perform Principle Component Analysis (PCA). PCA identifies linear combinations of the dataset that account for the greatest amount of variance in the data. I could've used t-SNE or UMAP for dimension reduction however, those methods lose the relationship between the clusters after the reduction, moreover, they would also require tuning of a hyperparameter. Thus, those methods would only be good for visualizing the data, in my study however, I was interested in how many parameters would be optimal in the model.

First I standardized all the parameters by subtracting the mean and scaling to unit variance. This ensured that each feature contributed equally to the analysis. Figure 1 illustrates that the first 3 components account for 60% of the variance in the data, each parameter after that accounting for negligible amounts of variance. Visualizing the remaining 3 parameters, Figure 2, shows that there are 2 clear clusters, All-Star and Non-All-Star players.

5 Modelling

I chose to train a multivariate logistical model, using the 3 components chosen from the PCA. This is because Multivariate logistic regression can incorporate multiple independent variables simultaneously, allowing for a more comprehensive assessment of the factors influencing the outcome variable

The goal of multivariate logistic regression is to estimate the odds of the outcome occurring for a given combination of independent variables. The odds are defined as the ratio of the probability of the outcome occurring to the probability of it not occurring.

An alternative way to model the data would be with a Naive Bayesian model as they are also robust with several parameters and can handle missing data by imputing the missing values using the mean or median of the corresponding feature. However, that wasn't required for my dataset.

6 Results and Analysis

From using a 70% train, 30% test split on the dataset, my results were positive but fairly worrying. First I plotted a ROC curve and calculated the AUC Figure 3, since the AUC was 0.979 it suggests that the trained model was really good. Secondly, I calculated the accuracy (0.978) precision, recall, and F-1 score (all 0.65). A low precision value indicates that the classifier is returning a high proportion of irrelevant results compared to relevant ones, false positives. The low recall can be explained by the large imbalance in the number of All-Star and Non-All-Star players. Since the F-1 score is the combination of all 3 previous metrics a low precision and recall would result in a

low F-1 score. Because of this, I decided to retrain my model but have front and backcourt players separated.

The AUC for both models was marginally better Figures 4, 5. The precision also improved when separated into 2 models up to 90% for frontcourt players and 66% for backcourt players. This increase shows that when using a classifier, it would be best to separate players by their position.

Maybe switching to a Naive Bayesian model would've improved performance. Since I decided to model front and backcourt players separately, the dataset of each model was reduced making each mode susceptible to overfitting. Naive Bayes models make a strong assumption about the independence of the features, which can help to prevent the model from overfitting to the training data.

7 Conclusion

In summary, from my investigation, I can see machine learning being used in selecting the NBA All-Star teams fairly so truly the best players can be selected to participate in the historical event and be compensated well. Also as fans, we'd be able to see fiercely contested games which no doubt be a spectacle and could further increase the viewership of the event.

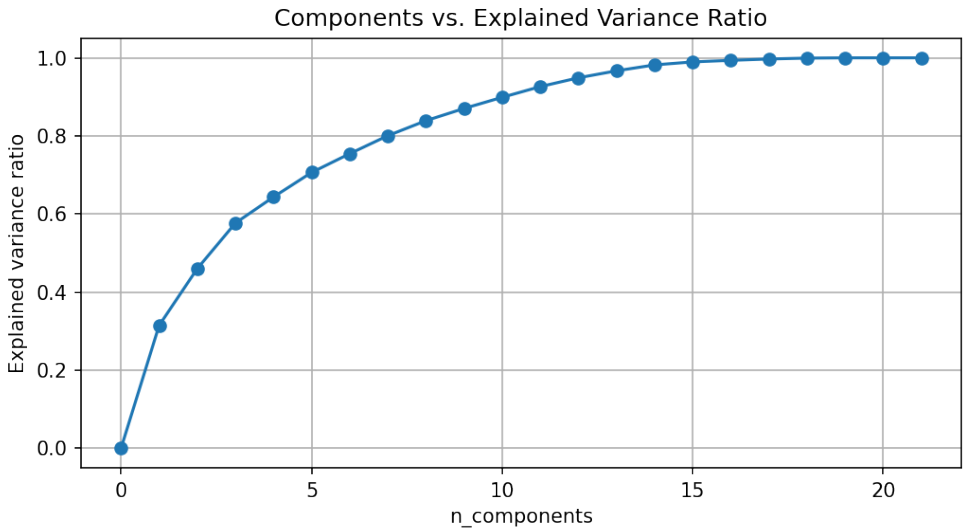

Figure 1: Explained Variance

References

[1] In *1951 NBA All-Star recap*. 2021.

[2] Michael Kaskey-Blomain. 2023 nba all-star game: These four players earn bonus cash thanks to contract clauses. *CBS Sports*, 2023.

[3] Paurush Omar. Nba all-star games: A look at past highlights, origin and evolution.

8 Appendix

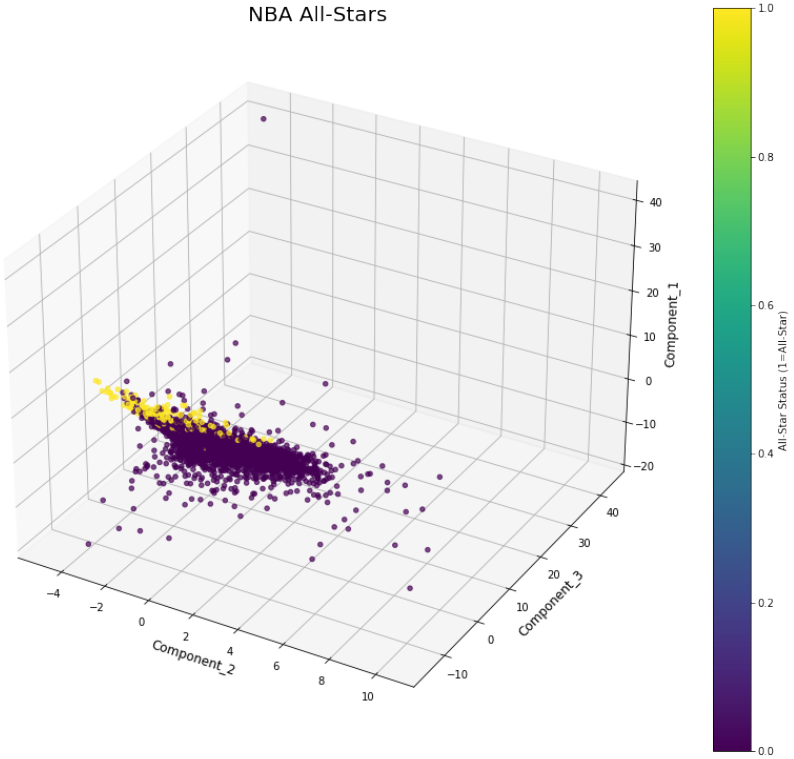

Figure 2: Visualisation of 3 components in relation to All-Star selection

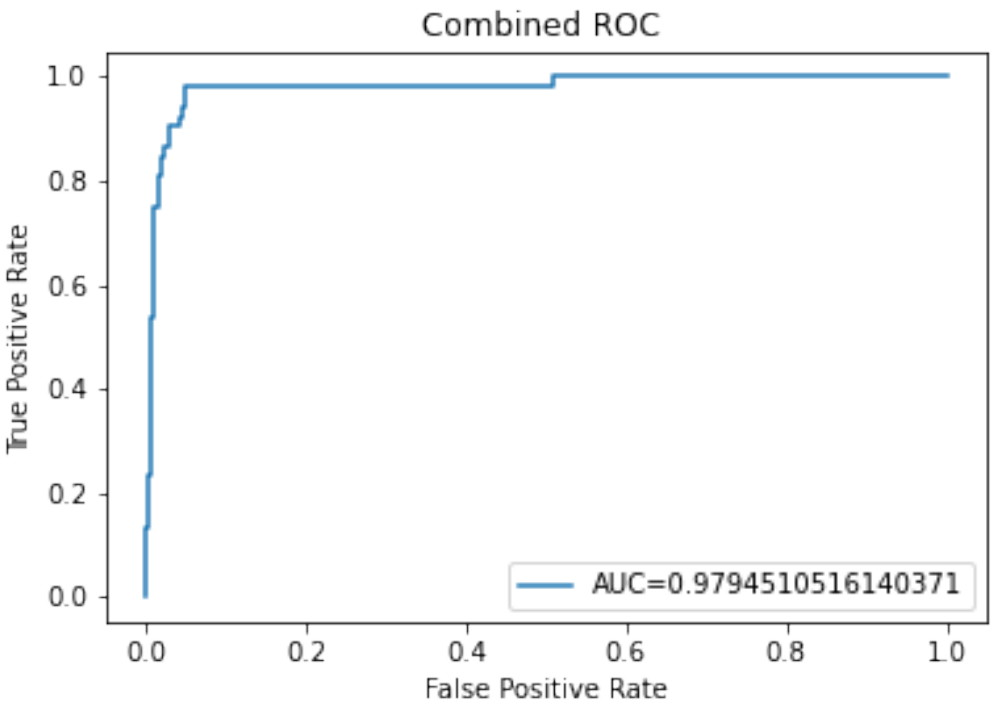

Figure 3: ROC of the initial model

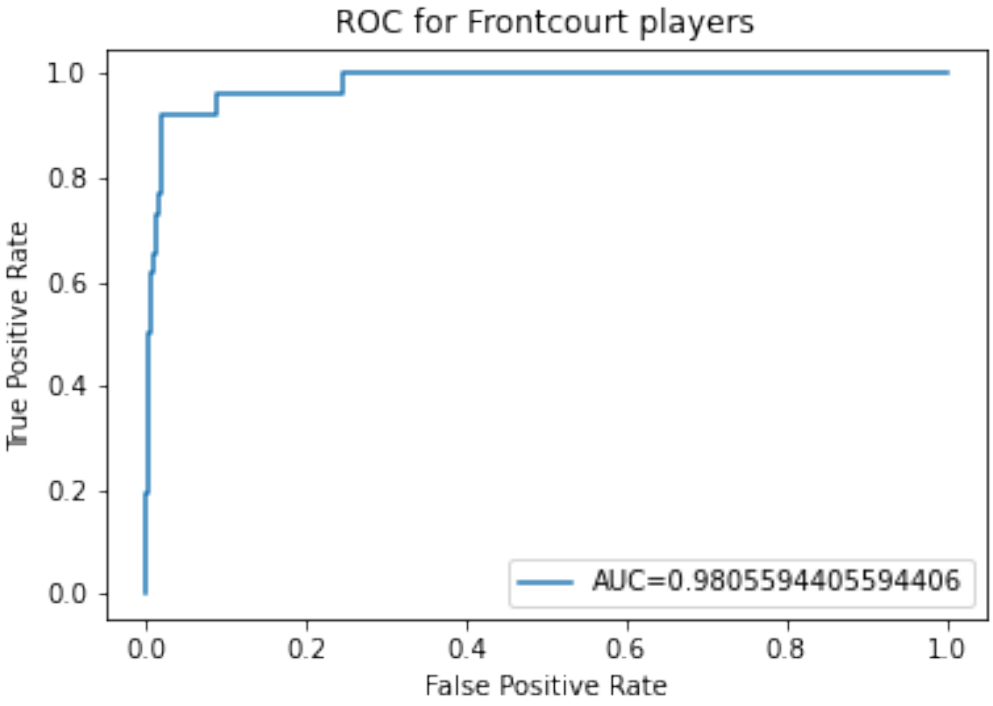

Figure 4: ROC of the model for just frontcourt players

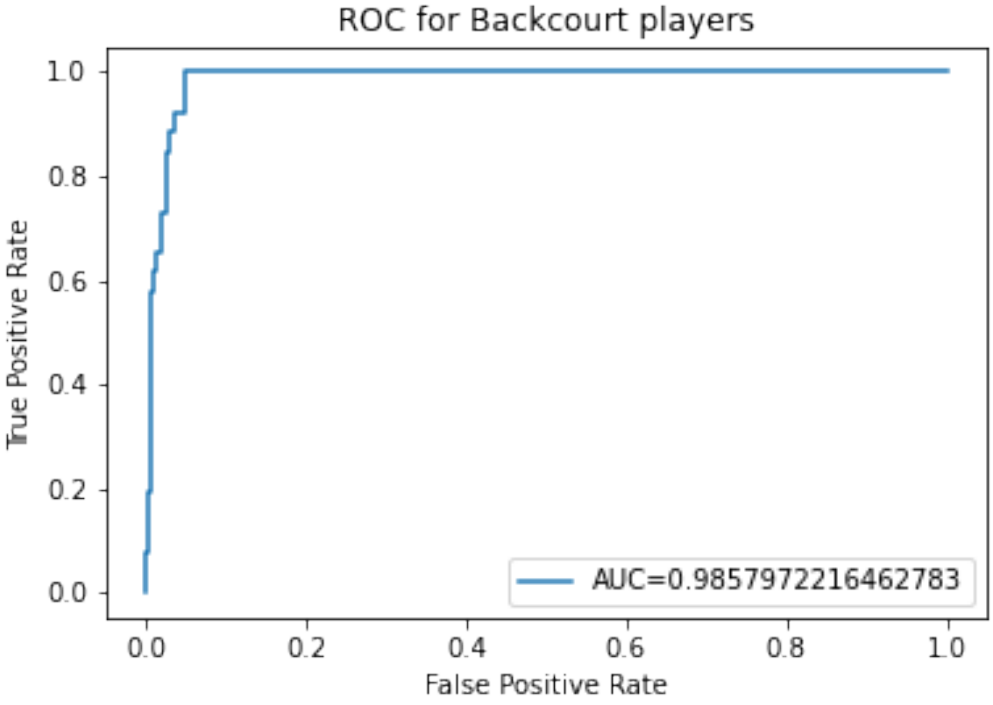

Figure 5: ROC of the model for just backcourt players