**Predicting NBA All-Star Players from Early-Career Traits**

Alayna Arnolie

5/2/2025

## Introduction

Basketball scouting and draft decisions are often based on early-career traits like physical attributes and collegiate performance. However, predicting whether a player will ultimately reach elite status in the NBA remains a challenge. In this project, I aimed to explore whether early-career traits, such as draft round, height, weight, and college attended, could help predict future NBA All-Star selections.

I chose this topic because I've been very involved in basketball both as a player and a fan for most of my life. It felt like a topic I could stay engaged with while also testing an interesting hypothesis: that early measurable attributes alone might offer clues about long-term success.

Using real-world data from the Men's Professional Basketball Dataset on Kaggle, I merged draft records, player biographical data, and a record of All-Star selections. I trained both a decision tree classifier and a logistic regression model to predict whether a player would eventually become an All-Star based only on those early-career traits.

While both models produced test accuracies around 74%, the results were shaped heavily by the class imbalance—far more players never become All-Stars than those who do. Although this accuracy may initially seem promising, further analysis revealed that the models performed significantly better on the majority class. Still, exploring different models, adjusting decision thresholds, and incorporating visualizations helped provide a fuller picture of what can and can't be predicted from early traits alone.

## Methods

### Data Description

I used the Men's Professional Basketball Dataset from Open Source Sports on Kaggle. This dataset contains multiple CSV files with real, historical NBA player data. For my project, I used the following:

- basketball_draft.csv: player draft details (round, year, team, etc.)
- basketball_master.csv: height, weight, college, and demographic data
- basketball_player_allstar.csv: a list of players who made All-Star appearances

After loading the data into pandas, I merged the draft and master files on a common player identifier (playerID/bioID). I then joined this with the All-Star dataset to label whether each player had ever been an All-Star, creating a new binary column called is_all_star.

**Preprocessing**

To keep the analysis realistic, I only used traits that would be known at the time of the draft. My features included height, weight, draft round, and college. These were chosen because they are publicly available before a player's NBA debut and are often heavily considered by scouts and front offices. Rows with missing values were dropped. Draft round was converted to numeric.

The college column contained too many unique values to be useful as is, so I simplified it by keeping only the ten most common colleges and grouping the rest as "Other." This cleaned column was label-encoded for modeling.

The label column (is_all_star) was created by checking whether the player's ID was present in the All-Star dataset.

**Modeling**

I trained two models: a decision tree classifier and a logistic regression model. Both were built using scikit-learn. The dataset was split into 80% training and 20% testing. The decision tree was chosen for its simplicity and interpretability, which makes it easy to understand which features are being used to make decisions. Logistic regression was included to allow exploration of threshold adjustments using predicted probabilities and to provide a probabilistic baseline model.

Model performance was evaluated using accuracy, confusion matrices, and classification reports. I also created three visualizations: a confusion matrix (Figure 1), a class distribution bar chart (Figure 2), and a scatterplot showing draft year vs. draft round by All-Star status (Figure 3).

**Results**

The decision tree classifier achieved a test accuracy of 0.74. However, it heavily favored predicting the majority class (Not All-Star). The confusion matrix (Figure 1) shows that the model correctly classified most non-All-Stars but failed to identify many true All-Stars. This highlights how class imbalance limited the model's effectiveness.

The logistic regression model gave similar results but allowed adjustment of the decision threshold. At a 0.3 threshold, recall for All-Stars improved slightly, meaning the model was better at identifying players who did become All-Stars. However, this came at the cost of lower precision and decreased overall accuracy. The classification report (Figure 2) revealed that precision and recall were both above 0.8 for non-All-Stars but remained below 0.3 for All-Stars. This trade-off shows the challenge of predicting rare outcomes.

The class distribution bar chart (Figure 3) emphasizes the imbalance in the dataset, and the scatterplot (Figure 4) shows that most All-Stars were selected in earlier rounds and earlier years. This suggests that draft position and timing play a role in long-term success.

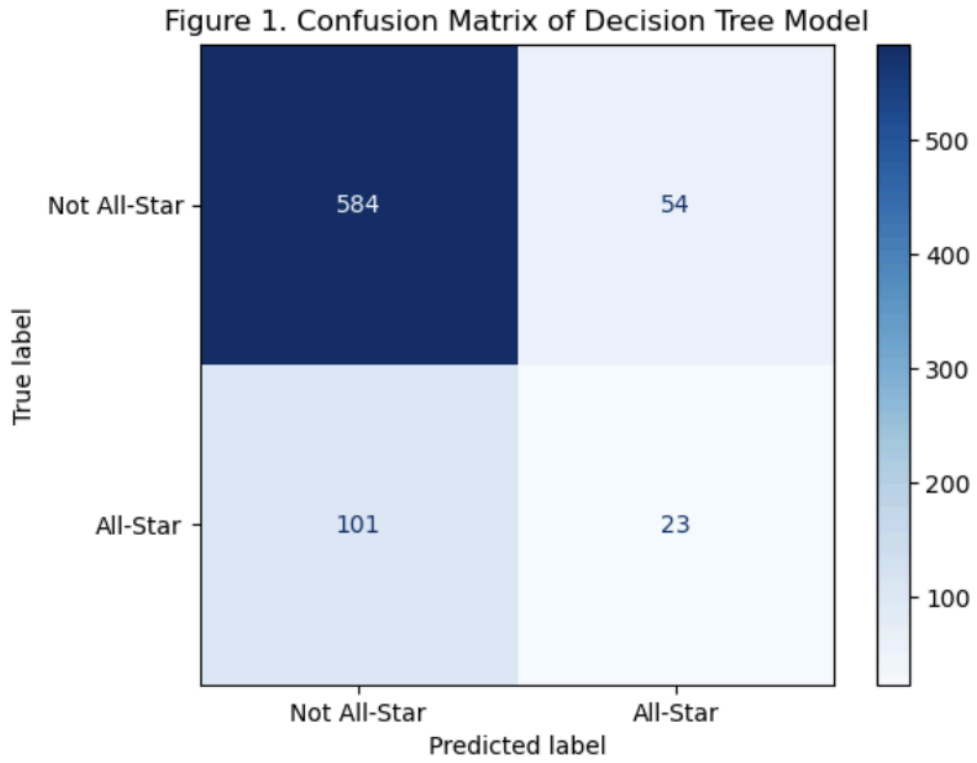


***Figure 1.*** *Confusion Matrix output from Decision Tree model.*

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.85 | 0.92 | 0.88 | 638 |
| 1 | 0.30 | 0.19 | 0.23 | 124 |
| accuracy | | | 0.80 | 762 |
| macro avg | 0.58 | 0.55 | 0.56 | 762 |
| weighted avg | 0.76 | 0.80 | 0.78 | 762 |

***Figure 2.*** *Classification report showing precision, recall, and F1-scores.*

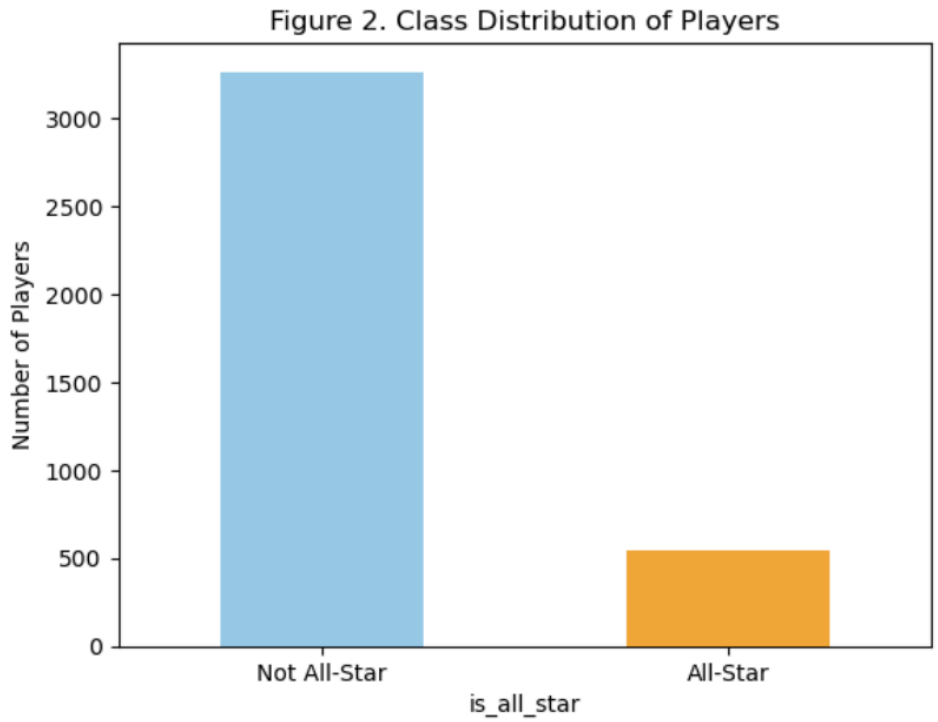


***Figure 3***. *Bar chart distribution. Most players are not All-Stars.*

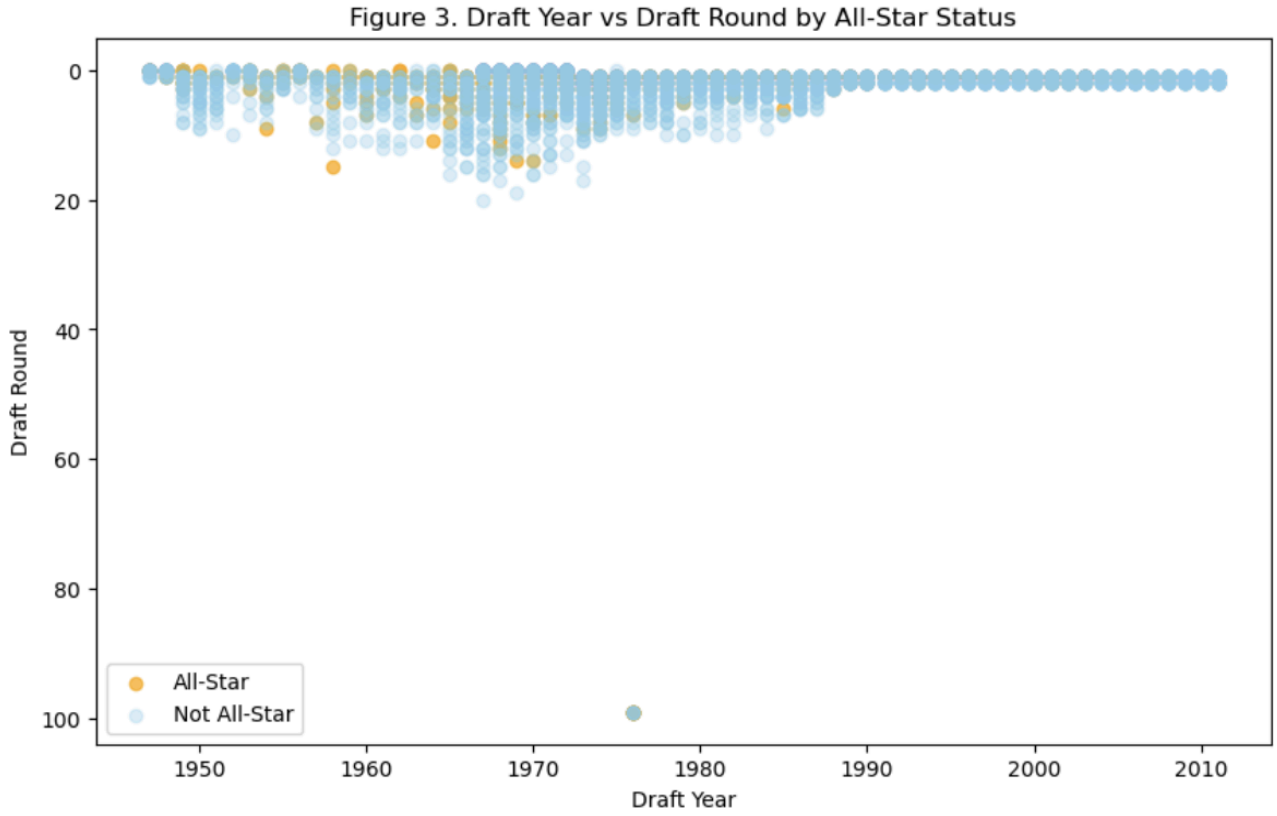


***Figure 4***. *Scatterplot showing Draft Year vs Round, colored by All-Star status.*

**Discussion**

This project showed that early-career traits like draft round, height, and college may offer some predictive value, but they are not enough alone to reliably predict NBA All-Star status. Both models struggled with class imbalance, leading to high accuracy for non-All-Stars but low precision and recall for All-Stars.

Even with threshold adjustments, the logistic regression model failed to achieve balanced performance. While recall for All-Stars increased slightly at a lower threshold, it came with a drop in precision and more false positives. This demonstrates the difficulty of detecting rare classes and highlights the need for a more robust feature set.

In future work, I would include performance statistics such as college points per game, efficiency ratings, or minutes played to offer more meaningful predictors. I would also explore applying class weighting to give the minority class more influence. Ensemble models like Random Forests could help improve generalization while maintaining interpretability.

Overall, this project helped me understand the limitations of surface-level traits and the importance of balancing accuracy, fairness, and model complexity in real-world prediction tasks.

## Citations

Lahman, Sean. *Men's Professional Basketball Dataset*. Kaggle, https://www.kaggle.com/datasets/open-source-sports/mens-professional-basketball. Accessed 4 May 2025.