Supervised Learning Report

Caleb Carlyle

April 2025

1 Context

NBA2k is a video game series that offers a simulation-style experience for users to play as a created player or their favorite existing players and teams. A point of fascination with each iteration of the game is the overall rating that each player receives, which is a combination of dozens of attributes. The large number of attribute ratings, designed to make a player perform like their real-world counterpart, offers a unique format of data to begin to investigate real-world play styles. For a few years, there has been a narrative that the NBA is becoming "positionless" and players are no longer fitting the traditional 5-position mold. Using data from the most recent NBA2k video game, NBA2k25, offers the opportunity to analyze positionality using more direct methods (the carefully selected attribute ratings) rather than trying to extract latent skills from real NBA box score data. The goal of this analysis is to find if NBA players are still fitting in the standard positions, what latent groupings may exist that differ from those positions, and to identify players who are anomalous under these frameworks that seek to categorize players.

2 Models

Several supervised classification models were trained to predict a player's position based on their NBA2k25 attributes. Model selection and hyperparameter tuning were critical components of this analysis due to the complexity and high dimensionality of the dataset. Each model class required specific attention to its hyperparameters in order to optimize performance while avoiding overfitting.

The Multinomial Logistic Regression model was tuned with a maximum number of iterations set to 50 to ensure convergence given the size of the dataset, and the 'lbfgs' solver was used for efficient optimization in multi-class settings. For the Regularized Logistic Regression model, the penalty parameter was set to L1 regularization (lasso) to promote sparsity in the model coefficients, helping to mitigate the risk of overfitting in a feature-rich dataset. The regularization strength was controlled by setting $C = 1$, balancing the trade-off between model fit and coefficient penalization.

The Random Forest Classifier required careful tuning of multiple hyperparameters to control model complexity. The number of trees in the forest was set to 100, providing a robust ensemble size. The maximum depth was limited to 15 to prevent overfitting on highly specific patterns in the training data. The maximum number of features considered at each split was set to the logarithm of the total number of features ('log2'), ensuring diversity among trees while maintaining computational efficiency. The minimum number of samples required to split an internal node was left at the default value of 2.

For the K-Nearest Neighbors (KNN) Classifier, the number of neighbors (k) was set to 20 to smooth predictions and reduce variance. The model used distance-based weighting so that closer neighbors had a higher influence on the predicted class, allowing the model to better capture local structure in the feature space.

Model	**Key Hyperparameters**	**Chosen Values**
Multinomial Logistic Regression	Solver, Max Iterations	lbfgs, 50
Regularized Logistic Regression	Penalty, C, Max Iterations	L1, 1, 50
Random Forest Classifier	Max Depth, Max Features, Estimators	15, log2, 100
K-Nearest Neighbors	Number of Neighbors, Weights	20, Distance

Table 1: Chosen Hyperparameters for Final Models

3 Model Comparison and Selection

Below are the classification reports for each of the models, showing fairly strong performance across models, particularly in predicting Point Guards and Centers, with more ambiguity in the middle positions.

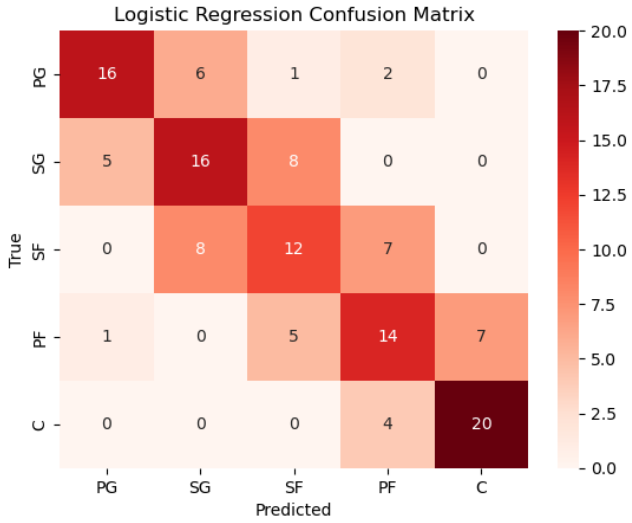

Figure 1: Classification Report for Logistic Regression

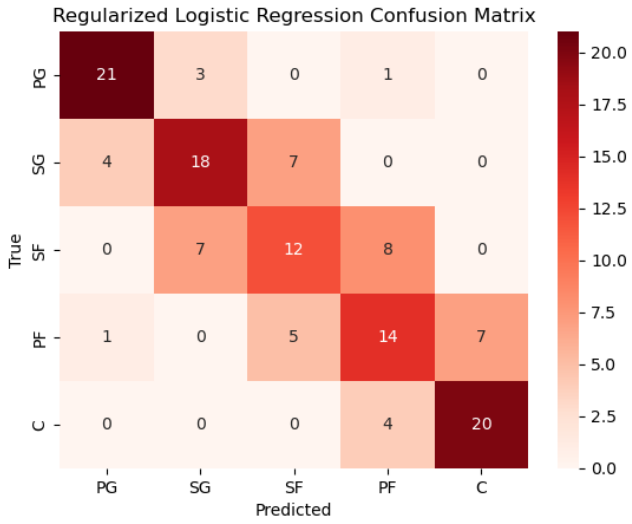

Figure 2: Classification Report for Regularized Logistic Regression

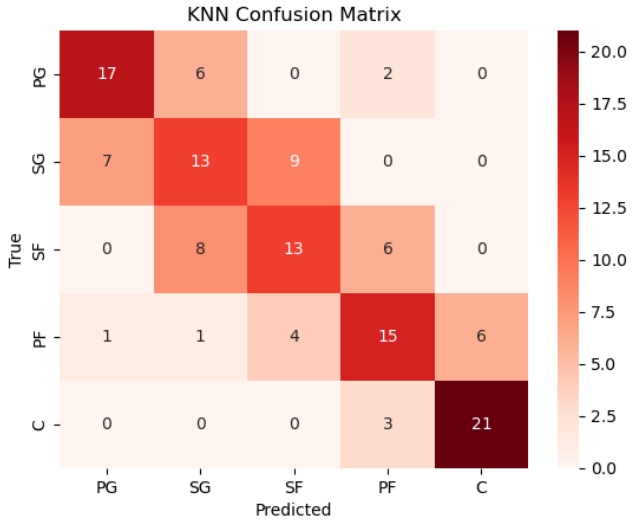

Figure 3: Classification Report for KNN Classifier

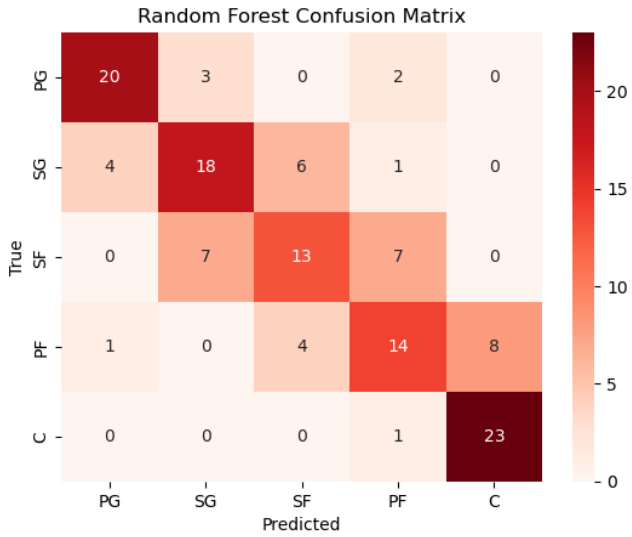

Figure 4: Classification Report for Random Forest Classifier

Across all models, there were clear trends in performance and behavior. The Random Forest Classifier outperformed linear models such as Logistic Regression, likely due to its ability to model non-linear relationships between player attributes and position. The Regularized Logistic Regression model performed better than the unregularized version, suggesting that the standard Logistic Re-

gression model may have been overfitting due to the high-dimensional feature space. The K-Nearest Neighbors Classifier performed reasonably well, but its accuracy trailed behind the Random Forest, even with the increased number of neighbors to limit overfitting.

Model	**Evaluation Metrics**
Multinomial Logistic Regression	Accuracy=0.591 , AUC=0.885
Regularized Logistic Regression	Accuracy=0.636 , AUC=0.892
Random Forest Classifier	Accuracy=0.667 , AUC=0.905
K-Nearest Neighbors	Accuracy=0.636 , AUC=0.896

Table 2: Accuracy and AUC for each model

Accuracy and AUC were chosen as the primary evaluation metrics for this data set. The data are balanced, with approximately equal class sizes for each position, so there is no need to worry about undersampling or other methods and metrics to deal with class imbalance.

Overall, the Random Forest model was selected as the final model due to its slightly improved evaluation metrics and lack of a much higher computational cost. One challenge encountered during modeling was the significant overlap between player positions within the attribute space, which reflects the real-world trend of the NBA moving towards positionless basketball. Hyperparameter tuning was also essential across all models to avoid overfitting while maintaining interpretability.

4 Explainability and Interpretability

To further interpret the Random Forest Classifier, SHAP (SHapley Additive exPlanations) values were computed to identify the most impactful features contributing to position prediction. The SHAP analysis revealed that Rebounding, Wingspan, Standing Dunk, Height, and Playmanking were the most influential attributes for predicting a player's position. The summary plot indicated that positions are largely estimated through physically-linked attributes like rebounding, dunking, and height. The exception to this comes to be playmaking and ball handling, skills that are not directly linked to a player's physical attributes. These are skills that are generally developed by smaller players in order to compensate for their size however.

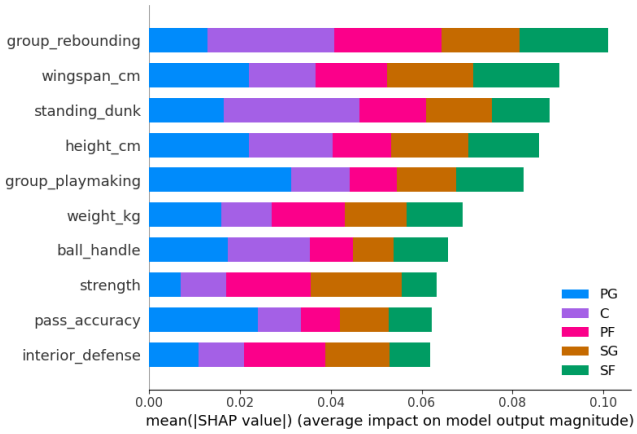

Figure 5: SHAP Summary Plot for Random Forest Classifier

5 Final Takeaways

This analysis demonstrates that supervised learning methods can effectively capture structure in NBA2k25 player attributes that relate to position, while also highlighting the growing ambiguity of traditional positional categories. The Random Forest Classifier achieved the highest overall performance in terms of accuracy and AUC, reinforcing its strength as a flexible, non-parametric model capable of handling non-linear relationships between features and class labels. However, even the best-performing models exhibited modest accuracy, suggesting that the distinctions between positions are becoming increasingly blurred, especially among versatile modern players.

The SHAP analysis of the Random Forest model revealed that attributes such as Rebounding, Wingspan, Standing Dunk, Height, and Playmanking were most influential in determining a player's position. Guards were typically differentiated by playmaking and handling skills, while forwards and centers were more dependent on dunking and rebounding attributes. Players with strong ratings across traditionally contrasting skill sets often confused the model, reflecting a real-world shift towards hybrid player archetypes that defy conventional labels.

Ultimately, this analysis supports the narrative that the NBA is evolving into a more positionless style of play. While player positions still provide a useful heuristic for organizing and thinking about basketball roles, the sharp boundaries that historically separated them are eroding. Machine learning models trained on curated attribute data from NBA2k25 provide a unique and data-driven lens through which to explore this transformation. Future work could extend this analysis by clustering players into latent archetypes independent of position labels or by analyzing changes in positionality over time using data from previous NBA2k editions.