Unsupervised Player Archetypes and Anomaly Detection in NBA2k25

Caleb Carlyle

April 16, 2025

1 Problem Context and Research Question

The unsupervised section of this project seeks to use the data found in NBA2k25 to accomplish two goals. The first goal is to identify latent clusters of players and compare them to the predefined positions for each player. In the narrative of an increasingly positionless game, it is interesting to see whether new "positions" or archetypes are emerging, contrary to the traditional five-position format. The second goal of this analysis is to identify anomalous players when evaluating the clusters. After identifying anomalous players, it is of interest to determine what attributes separate them from their classification counterparts. These goals motivate the usage of clustering and anomaly detection techniques that do not rely on labels but instead focus on the constructs that underlie the data.

2 Methods and Workflow

2.1 Dimensionality Reduction with UMAP

UMAP was applied to project player attributes into 2D space for clustering and visualization. The plot of the UMAP projections is colored by known position to show the significant mixing, particularly among the Small Forwards and Power Forwards. This further motivates the idea that the traditional positions may not be the most informative groupings, and there may be a different number of underlying clusters, instead of the traditional 5 positions.

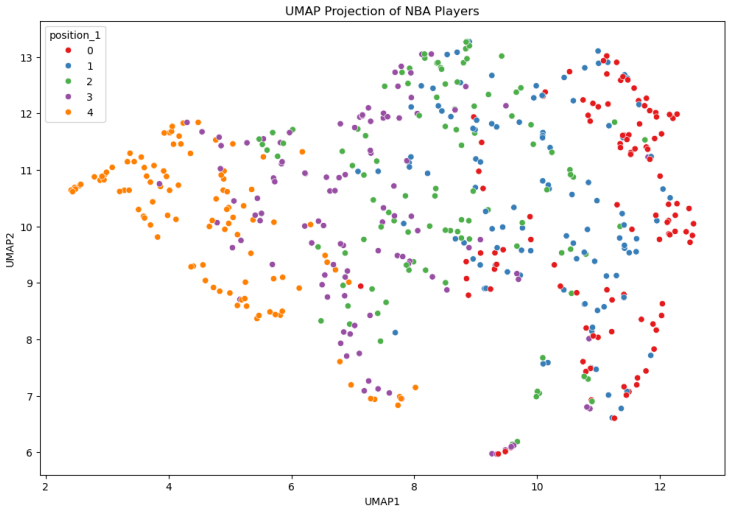

Figure 1: UMAP Projection of NBA Players Colored by Known Positions

2.2 Clustering Analysis

KMeans was selected to cluster the players. To evaluate the number of clusters to be used, WCSS and Silhouette scores were used. The WCSS plot does not have a strong "elbow", but 3 appears to be a decent selection. 3 clusters also seems appropriate based on the Silhouette Scores as well, with a steep dropoff after 4 clusters. The WCSS for the chosen model was 17280 and the Silhouette Score was 0.1797.

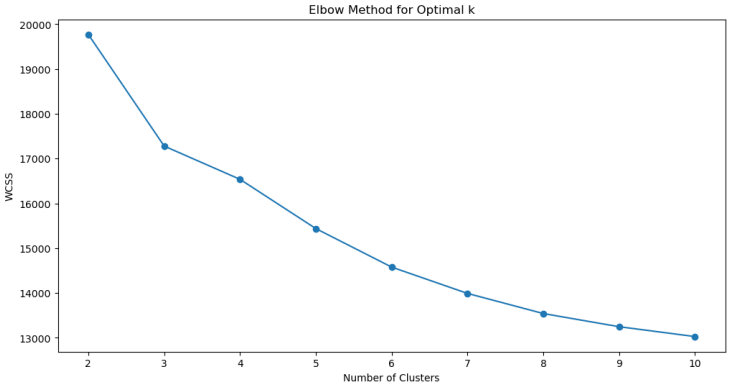

Figure 2: Elbow Plot for K Selection

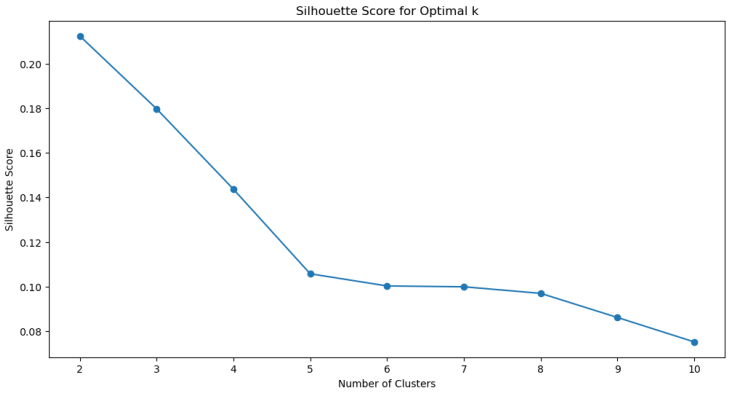

Figure 3: Silhouette Scores for Different K Values

After identifying 3 clusters, I used a Random Forest Classifier to obtain feature importances for the classification into the three clusters. Summaries of the most influential attributes are found below in 1.

Feature	**Importance Value**
Ball Handling	0.0796
Speed With Ball	0.0571
Group Playmaking	0.0538
Group Inside Scoring	0.0486
Height(cm)	0.0479

Table 1: Top 5 Most Important Classification Features

To visualize the clusters that were found, the same UMAP plot from earlier is included (4) to show the separation between clusters. Using the 3-cluster KMeans, it is clear that there is greater separation than there is using the original labeled positions. There is still some overlap between clusters which drives the interest in anomaly detection.

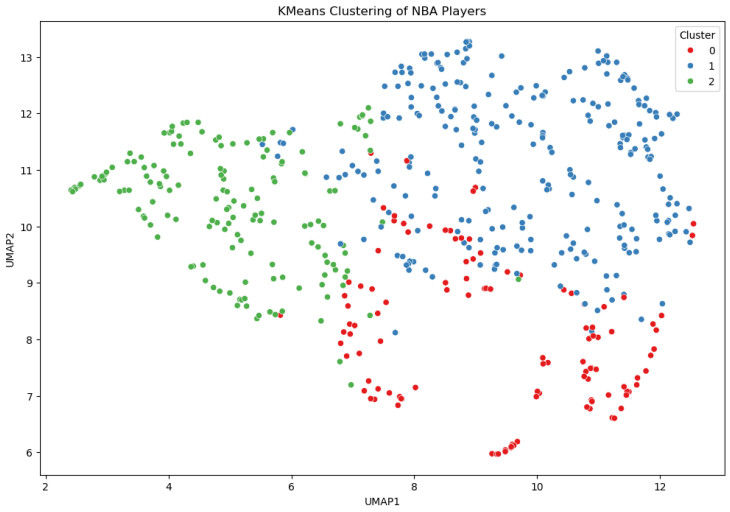

Figure 4: UMAP Projection with K-Means Cluster Assignments

2.3 Anomaly Detection with Isolation Forest

An Isolation Forest model was fit on the data to identify outlier players based on their attribute profiles. The Isolation Forest identified 6 players to be clustered in an anomalous fashion. Those players are:

- Anthony Davis
- DeAndre Jordan
- Giannis Antetokounmpo
- Joel Embiid
- LeBron James
- Nikola Jokic

These players are interesting because, with the exception of DeAndre Jordan, they are considered some of the most elite players in the NBA. Given this context, it is not surprising that they do not cluster neatly within the 3-cluster framework. Feature importance can give a better indication of what drives their anomalous classification.

2.4 Feature Importance

The figure below shows the most important features in the determining the anomaly scores.

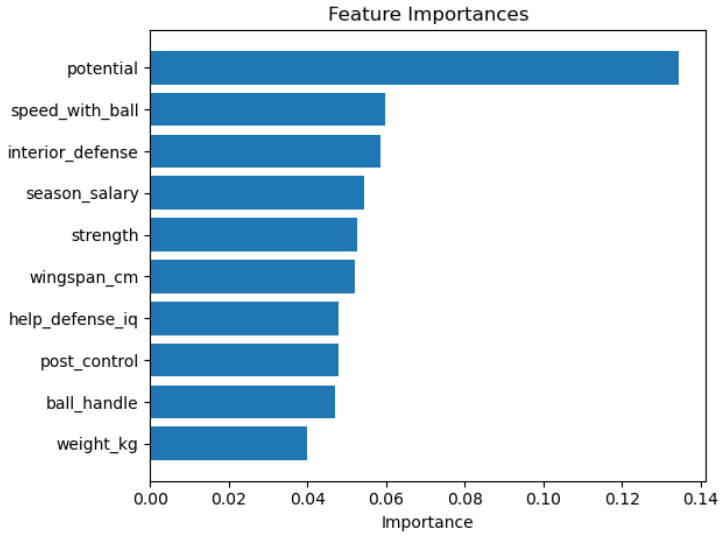

Figure 5: Isolation Forest Feature Importance for Anomalous Players

Based on the feature importances, the potential rating is by far the most influential feature. This makes sense with the players who were identified as anomalous because the majority of them had extremely high potential ratings, as visible in 2. DeAndre Jordan exhibits a lower potential rating so to see the reason he is considered an anomaly, his shap values were explored. It seems that he is an anomaly among anomalies since his potential is not even in his top 10 shap values. Instead, his combination of horrible speed, ball handling, and playmaking contributes to him not fitting any of the clusters well.

Player	**Potential**
A. Davis	96
D. Jordan	75
G. Antetokounmpo	98
J. Embiid	98
L. James	99
N. Jokic	99

Table 2: Anomaly Potential Ratings

3 Comparison to Supervised Models

The unsupervised approach revealed nuanced player groupings that do not always align with the strict positional labels used in the original dataset. While

the supervised models forced players into predefined roles, the clustering analysis suggested transitional or hybrid roles are increasingly common.

Anomaly detection also surfaced influential players who may not fit traditional molds but could have outsized impact in gameplay or strategic matchups.

4 Conclusion

This project demonstrates the utility of unsupervised learning in sports analytics. Clustering identified emergent player archetypes, and anomaly detection aided in understanding what separates players from these molds. It is interesting that the traditional 5-position grouping is not supported by unsupervised learning, and the KMeans clustering created improved separation between clusters than the original positions.

These findings align with the broader narrative of an increasingly positionless NBA, where player roles are defined more by skillsets than by rigid positional labels. The emergence of three primary clusters suggests that player archetypes may be better represented along functional or stylistic lines rather than conventional court assignments. Features such as playmaking, ball handling, and physical attributes like height and rebounding were critical in driving these clusters, reinforcing the idea that modern basketball favors hybrid skill profiles.

Furthermore, the use of Isolation Forest for anomaly detection brought attention to exceptional players whose unique combinations of attributes defy typical classification. The fact that many of these outliers are elite players (e.g., Jokic, Giannis, LeBron) illustrates how statistical anomalies can reflect innovative or dominant playstyles that challenge traditional frameworks. Even the case of DeAndre Jordan, who was flagged for the opposite reason—extreme deficiencies in areas like speed and playmaking—underscores how the model captures divergence in both directions of performance.

In sum, this analysis offers a data-driven lens to understand how player roles are evolving in professional basketball. It not only reveals latent structures in player attributes but also highlights those who lie beyond the boundaries of standard classification. Future extensions of this work could incorporate longitudinal data from past game editions to track the evolution of player archetypes over time or integrate in-game performance statistics to validate the utility of these clusters in real-world contexts.