Project 3

Octavio Villalaz

11/19/2024

CSE-632

Abstract

This project analyzes NBA player performance and archetypes using a dataset spanning multiple seasons. Key objectives included clustering players based on performance and physical attributes, uncovering associations between draft details and performance, and detecting outliers. Results highlight player archetypes, relationships between draft positions and success, and standout players who deviate from norms.

Project: NBA Player Archetypes and Performance Analysis

Problem Description

The project focuses on identifying patterns, archetypes, and anomalies among NBA players using a dataset spanning multiple seasons. The goal is to analyze performance metrics, physical attributes, and draft details to derive actionable insights.

Dataset Description:

- **Format**: CSV file named all_seasons.csv.
- **Dimensions**: 22 columns and 12500 rows, each representing a player's season performance.
- **Key Attributes**:
  - Player Name, Team Abbreviation, Age, Player Height, Player Weight
  - Performance metrics: Points, Rebounds, Assists
  - Draft details: Draft Round, Draft Year

Example:

player_name	pts	reb	ast	draft_year	draft_round
Michael Jordan	35.6	6.4	5.7	1984	1
Shaquille O'Neal	29.7	11.2	3.5	1992	1

Software Tools

Tools:

- Python libraries:
  - pandas for data manipulation.
  - sklearn for clustering and outlier detection.
  - matplotlib for visualization.

- Microsoft Word for documentation and result presentation.

Algorithms:

- **KMeans Clustering**: Grouped players into archetypes based on performance and physical attributes.
- **Association Rule Mining (Apriori)**: Discovered links between draft rounds/years and performance metrics.
- **Isolation Forest**: Detected outliers based on performance and physical characteristics.

Preprocessing

- ***Data Cleaning***: Columns like pts, reb, ast, player_height, and player_weight occasionally had missing values. These rows were dropped for analyses where complete data was required. Column **draft_round** was initially stored as text and was converted to numeric for analysis. Invalid or non-numeric values like the column Undrafted were treated as NaN and excluded. **draft_year** was converted to numeric to calculate correlations with performance metrics.
- ***Feature Standardization***: To account for the differences in scales, attributes such as pts, reb, and ast (in the tens), as well as player_height (in centimeters) and player_weight (in kilograms), were standardized using z-scores. This ensured that unit variations did not skew analyses like clustering or outlier detection.
- ***Data Transformation***: Categorical attributes such as draft_round and draft_year were one-hot encoded to generate binary columns (e.g., draft_round_1, draft_round_2), enabling a structured representation for association rule mining. Performance metrics like pts and reb were converted into binary flags (high or low) based on a median split, facilitating the identification of patterns, such as "players drafted in round 1 often score above the median points."
- ***Dimensionality Reduction***: Correlation analysis was employed to identify and address feature redundancy before conducting outlier detection and clustering. Features like player_height and player_weight were retained without transformation, as they showed weak correlations with performance metrics, ensuring they remained relevant for analysis.

This preprocessing ensured the dataset was clean, consistent, and properly formatted for clustering and outlier detection. Let me know if you'd like this incorporated into the Word document or need further elaboration!

Data Mining Process

This project's data mining process was meant to generate useful insights by analyzing NBA players' performance and traits. It used three main techniques: clustering for player archetypes and outlier identification to identify extraordinary players. Each procedure was carefully applied to achieve reliable findings.

Clustering was employed to group players into archetypes based on their performance metrics and physical attributes. The dataset was first preprocessed by selecting relevant features, such as points, rebounds, assists, height, and weight. These attributes were standardized to ensure that differences in scale did not skew the clustering results. The optimal number of clusters was determined using the Elbow Method, which indicated that three clusters would best represent the data. Using the KMeans algorithm, players were grouped into three distinct categories: scoring specialists, rebounding/defensive specialists, and balanced all-rounders. These archetypes were analyzed to reveal insights into player roles and how they contribute to team dynamics. For example, scoring specialists were identified as players excelling in points but average in other metrics, while rebounding specialists were taller players with dominant rebound stats but lower scoring.

Outlier detection sought to identify players who strayed considerably from expected performance or physical standards. The Isolation Forest method was employed, with contamination set to 5%, to identify players who stood out based on their points, rebounds, assists, height, and weight. Preprocessing involved standardizing these traits to ensure equitable evaluation across measurements. The study highlighted noteworthy performers such as Michael Jordan, recognized for his remarkable scoring and all-around ability, and Allen Iverson, who scored well despite being shorter than normal. These outliers reflect distinct player profiles that challenge standard archetypes, giving teams opportunities to capitalize on their exceptional abilities.

Results

Figure 1 demonstrates that the correlation between draft_round and performance metrics shows a negative relationship, with points (pts) having a stronger negative correlation (-0.27) than rebounds (reb, -0.23). In terms of scoring and rebounding measures, this suggests that players chosen earlier in the draft (lower draft_round values) typically perform better. On the other hand, individuals selected later in the draft have a lower chance of excelling in these domains. The draft year appears to have little bearing on a player's performance measures in this dataset, as evidenced by the weak correlation between draft_year and both pts (0.01) and reb (-0.04).

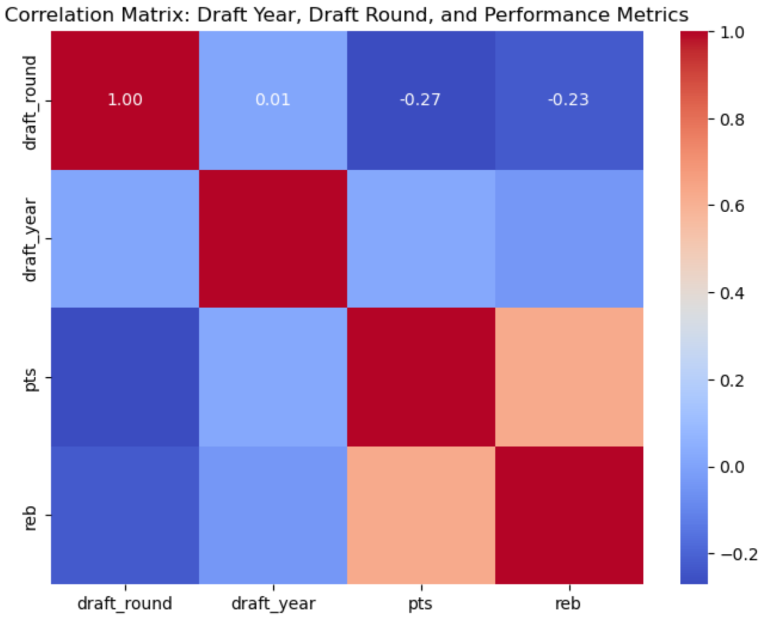

Figure 1: Correlation Matrix

Figure 2's scatter plot illustrates the correlation between NBA players' scoring and rebounding, emphasizing elite performers with green stars. Joel Embiid is an example of an excellent all-around player who excels in all measures and makes a substantial offensive and defensive contribution to his team. James Harden and Kobe Bryant, two scoring specialists, are notable for their excellent scoring but mediocre rebounding, highlighting their roles as the team's main offensive alternatives. Conversely, players who specialize on defense and board management, such as Dennis Rodman and Ben Wallace, have remarkable rebounding skills despite contributing little in the way of points. While the standout players occupy different sections of the plot according to their specialty, the majority of players cluster in the lower-left quadrant, signifying average contributors with modest scoring and rebounds.

With guards (like Harden and Iverson) leading the scoring category and forwards or centers (like Embiid and Drummond) dominating the rebounding category, the distribution shows distinct positional trends. With strategic ramifications for team-building, this approach emphasizes the existence of performance archetypes like scoring guards, rebounding experts, and well-rounded players. For instance, defensive-minded teams may target players like Rodman, while offensive-minded teams may favor scorers like Harden. The image draws attention to how uncommon exceptional players are and stresses how crucial it is to recognize and utilize distinct skill sets in order to assemble a competitive and well-rounded squad

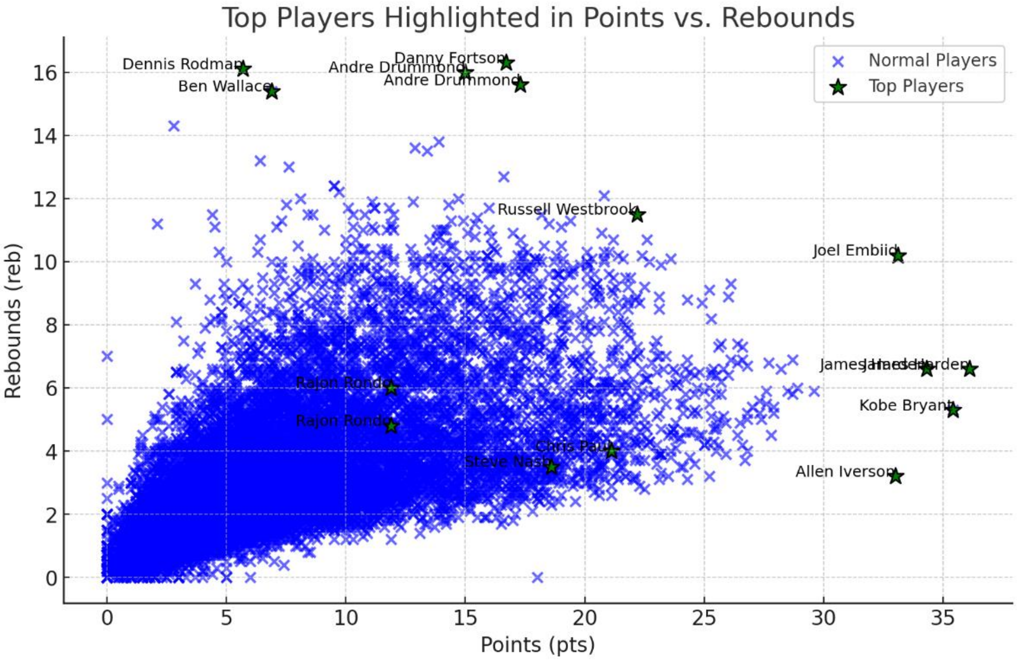

Figure 2: Top Players Highlighted in Points vs. Rebounds

Although correlation and clustering studies are useful tools on their own, combining them can yield a more comprehensive picture of player performance and draft tactics. While correlation exposes underlying statistical patterns, clustering identifies outliers and player archetypes. When combined, these approaches can help teams, scouts, and analysts make well-informed decisions. For deeper insights, future research should concentrate on improving interpretability, adding more data, and utilizing cutting-edge machine learning algorithms.