CS 4112 Project Proposal

Gabriel Oyarekhua
College of Computing and Softeare Engineering
Kennesaw State University
Kennesaw, USA
goyarekh@students.kennesaw.edu

***Abstract*—With the rise of positionless basketball in the NBA, the traditional five positions on the court (PG, SG, SF, PF, C) are no longer descriptive enough for determining the on-court value and skillset of a player. Using various statistical methods and data mining techniques, this project aims to identify data-driven archetypes of players that exist beyond the traditional 5 archetypes and how they have impacted teams over time. By segmenting players into clusters using performance data drawn from NBA statistics over time, we can better understand the evolution of different player archetypes and outliers that may exist.**

I. Introduction

Basketball has become a positionless game. We are continuing to see players enter the league that defy the traditional norms of the position they are meant to play. As a result, the conventional five-position framework is no longer sufficient for describing player roles and on-court value. This project aims to identify 3 questions: (1) What player archetypes exist beyond the traditional positional labels? (2) How have these archetypes evolved over time? and (3) How do these archetypes contribute to or hinder team success? By analyzing historical performance data and grouping players based on statistical similarities, this study aims to provide a clearer understanding of player evolution in the NBA and identify the types of players that contribute most to successful teams.

II. Data Source

Main datasource for this project will be the NBA Api. This api is an open source abstraction on the main stats.nba.com, making it easier to query and work with NBA data in python. Data is partitioned to include both static, up to date player and team data that can be returned in the form of multiple formats such as JSON, DataFrames, or dictionaries making this library extremely easy to work with. There will be 2 main endpoints being used to start the project:

- **Leauge Dash Player Bio Stats:** This endpoint shows base player metrics such as their traditional position, weight and height. This information is critical in order to feed the algorithms important biometric player information
- **Leauge Dash Player Stats:** This is the endpoint to retrieve all player information over the source of an entire season. These basic statistics will serve as a baseline for clustering and outlier detection

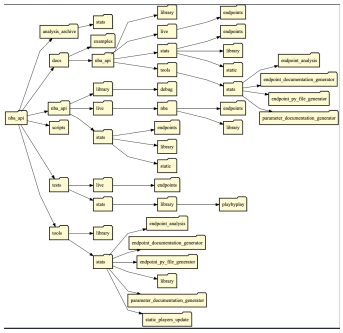

Fig. 1. Nba API Structure

More advanced endpoints are readily available and new features can be implemented iteratively in order to improve the performance and reliability of algorithms implemented

III. Planned Techniques

The completion of this project depends on a comprehensive understanding of multiple algorithms taught throughout the semester. The methodology is structured into three distinct phases: dimensionality reduction, anomaly detection, and clustering. Each method serves a specific purpose in moving the analysis from broad performance metrics to actionable roster insights.

A. Dimensionality Reduction

With the amount of features contributing to player statistics, dimensionality reduction will be important in order to improve the accuracy of other data mining methods use throughout the lifecycle of the project. Dimensionality reduction will also lead to more efficient processing of the features overall

B. Anomaly Detection (PCA)

Anomaly detection will be used to categorize players that may be extremely talented or lacking, as insights gathered from outliers can be of use when looking at team structure, compatibility, and the overall skill increase or decrease of players overtime.

C. Clustering

The accuracy and interpretability of clustering techniques used in this project are the backbone and most important component of this project. Clustering algorithms will take this data and assign categories based on the features fed into the algorithm, allowing for data-defined categories of players.

IV. Projected Timeline

This project will closely follow the KDD process defined in week 1 of the course. This allows for a comprehensive pipeline throroughly injesting, evaluating, and interpreting data in a way that makes the best use of the chosen dataset

1) **Preprocesing:** Both Player and team Statistics will be ingested year over year before bring combined and organized into 2 comprehensive datasets. One of players organized team over the past 10-15 years (determined by overall data availability), and the other of team data organized by year spanning the same length as player data. Next, Exploratory Data Analysis will be performed to gain a better understanding. In this phase schema will be unified, data will be checked for a consistent structure, and null values will be handled accordingly. Both univariate and multivariate exploratory data analysis techniques will be performed to visualize and make sence of the data ingested from the NBA API
2) **Transformation / Dimensionality Reduction:** Once the data has been cleaned, organized, and evaluated, the PCA algorithm will be used in attempt to reduce the dimensionality of the dataset while preserving useful information in the clustering phase
3) **Clustering:** Different clustering methods will be used and compared throughout the course in order to gain the most insightful cluster of player data on individual seasons. Once metrics are fine tuned, the best clustering algorithm will be run on all seasons in order to show the overall data trend and cluster movement.
4) **Anomaly Detection:** While Anomoly detection will be used in the beginning stages (EDA), more finetuned techniques learned in the course will be used to further understand clustered data in order to interpret the data.
5) **Team Data Integration:** players belonging to the same team in the same year will be grouped together and assigned to clusters created in order to get a deeper insight into the construction of teams and determine if there are patterns relating roster construction beyond traditional archetypes and team success

Milestones

The project is seperated into 3 deliverables with clear expectations by each point. However, iteration is to be expected as we are learning on the fly and new features may be added

a) **M2 (Data Cleaning and EDA):** Clean, scaled, and organized dataset in csv fromat with clear visuals showing comprehesnsive understanding of both player and team data
b) **M3 (PCA and Initial Clustering):** Reduced dataset along with initial clusters using initial cleaned data. Also, evaluation document of current state of clusters, issues found, and plan going forward.
c) **M4 (Cluster Optimization and Team Integration):** Fine tuning clusters (if errors found) followed by an interpretation the data using domain knowledge. By the end of this stage visualizations of clusters, team evaluation, and anomalies should be complete along with the report.

V. Project Structure

The initial proposal structure is just a reference point and is subject to change throughout the course of the semester. Currently, the planned structure is as follows:

- **.Env & .GitIgnore:** These files will store all secrets and configurations required to query the api as well as information not being pushed to the remote repository
- **Data:** Here is where the data will be stored. Data will be split into raw and clean data in order to reproduce results. The raw folder will hold the data coming straight from the api while the clean data will contain the processed and normalized data ready to be fed to the data mining algorithm
- **Notebooks:** Here will be where the scripts will be ran. Notebooks allow for visualizations to be shown and models output to be interpreted as opposed to a normal python file.
- **Scripts:** The scripts will hold all the logic and error handling for processing the data and running the data mining algorithms.
- **Models:** This folder will store all the models for ease of access throughout the later stages of the project
- **Reports:** Here is where all visualizations will be output to in order to further save results in order to use in the final report.

VI. Anticipated Challenges

Perhaps the biggest challenge will be adjesting for era. With the style of basketball being completely different now than even ten to fifteen years ago, comprehensive research will have to be done to mitigate these challenges. Another huge challenge is learning on the fly while completing this project. Communication and community is crucial to the completion of this project. Going to the CCSE tutoring lab, office hours, and online forums will be crucial to the completion of this project. Some Python libraries that will be required to complete this project are bound to be new, so being able to learn new technology while completing this will be a unique but interesting challenge to tackle, further increasing problem-solving and development skills.

VII. Conclusion

In conclusion, this project provides a unique opportunity to merge the domain knowledge with the techniques and concepts learned through the course of the semester. By building a full pipeline, this project will be a repeatable framework, using both data mining and software engineering skills to demonstrate how players have evolved to this positionless era.