Roster IQ

By: Seth Rojas | Mentored by: Mihai Surdeanu

*GitHub access upon request

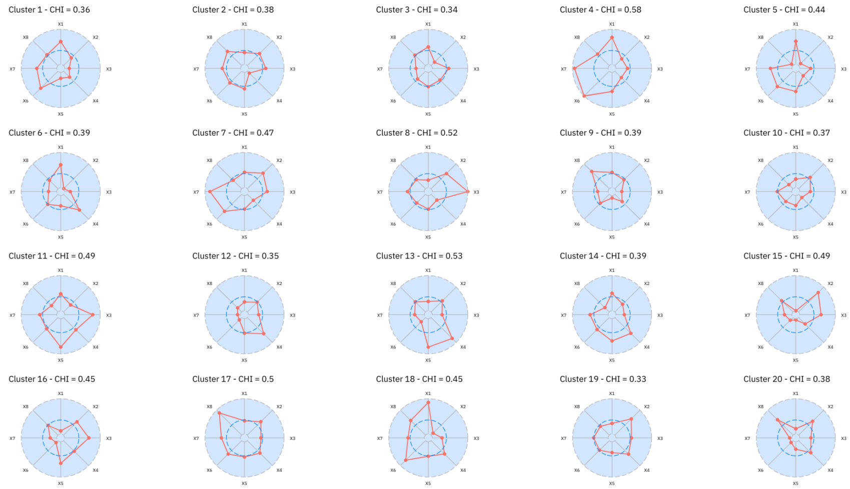
Goal

The goal of this project is to estimate and predict the impact of a Division I basketball transfer portal player on an incoming team. I initially sought to predict a player's box-plus-minus (BPM) as a direct measure of impact but found that the inherent variability and contextual dependencies made accurate prediction challenging. As a result, I am developing two alternative metrics that more reliably measure player transferability: Clustered Fit Score (CFI) and Value Over Clustered Replacement Player (VOCRP). CFI measures how well a player's statistical profile aligns with the playing style archetype of the team they are joining—defined by the team's assigned cluster—by comparing the player's statline to the average statline for their position within that cluster. VOCRP quantifies the marginal value a player provides relative to a similar player in the same system, using advanced basketball metric comparisons under the same clustered context.

Methods

I collected regular and advanced player and team season statistics from Bart Torvik; not including player weights which were webscraped from all D1 University college basketball websites. High school and Junior College player rankings were web scraped from ESPN and JUCORecruiting.

I initially trained three XGBoost models—a regression model, a multiclass classifier, and a binary classifier—to estimate a transfer player's box-plus-minus (BPM) for the upcoming season. Features included a combination of the player's profile and previous season statistics, aggregated metrics of incoming teammates, and the player's metrics relative to those teammates.



I transitioned to clustering teams using k-means based on a multi-dimensional snapshot of advanced performance, play style, and efficiency metrics. These clusters define team archetypes that allow me to classify any group of players into a repeatable system identity, significantly reducing noise and enabling structured player-to-player comparisons within a cluster. Currently, I am generating synthetic teams by aggregating player stats from the previous season, allowing me to simulate each team's projected identity for the upcoming year and apply my fit and value metrics (CFI and VOCRP). Each player is weighted according to their prior minutes and a role modifier—a scalar derived from their predicted role (via the BPM classifier) adjusted by the model's confidence score–to calibrate for individual player influence when averaging.

Results

The regression model achieved an R^2 of 0.476, a mean absolute error of 2.37, and a median error of -0.01. While it performed well for a handful of players, it struggled with outliers—typically players with poor seasons or limited opportunity—and ultimately proved too noisy to predict BPM precisely, which typically ranges from -6 to 6. I then shifted to a classifier that predicted BPM ranges (bench/rotation/starter), achieving 63% accuracy. The model correctly identified bench or replacement-level players (BPM < 0) 71% of the time but had more difficulty distinguishing between rotation-level (38% accuracy) and starter-level (60% accuracy) roles. A simplified binary classifier that predicts whether a player will post a positive BPM achieved 78% accuracy and may offer practical value for coaching staffs evaluating a player's impact potential.

Most unexpected finding

I hypothesized that I would be able to predict the box-plus-minus of a player for a next season; however, there are a multitude of unaccounted variables limiting the possibility of an accurate model. Due to the sheer noise, I am surprised these results were even this accurate. CFI and VOCRP should more comprehensively capture player transferability fit regardless.

Future Plans

I am currently generating estimates for incoming players (high schoolers and junior college transfers) who do not have previous season statistics. From here, I will be able to apply my two metrics across all transfer portal players from 2018-2024 and analyze how insightful these metrics are to team performance the next season. I hope to publish these metrics and release them to Division I basketball coaches evaluating their next best transfer portal talent.