

# Stat 3274 Final Project Progress Report

Baker Dean & Caleb Ramsey

## Research Question

How can we create a player rating system for men's and women's college basketball that accounts for both player performance and conference strength, allowing staff to compare athletes across divisions and evaluate transfer portal talent more effectively?

## Abstract

In this new age of college basketball, evaluating the talent of players across the nation is more important than ever given the new transfer portal and NIL rules. Player movement is at an all-time high with teams often turning over more than 75% of their teams every offseason and with players able to be paid for their services coaches and general managers must be aware of the true value of a player, how much to offer him/her and how to allocate all their financial resources. With many players choosing to move between the Power 5 conferences (ACC, Big Ten, Big 12, Big East, and SEC) and the many smaller conferences, evaluating the skill and fit of a player is not as simple as simply pulling up stats and sorting by points per game. With such a wide variety of levels of talent, determining how to account for the level of competition they faced becomes an important factor alongside the performance of the player.

This project seeks to accurately handle this issue by creating a video-game style player ratings system that will assign every player in the nation a single number that represents the level of competition they faced and their individual performance. We will determine level of competition by analyzing the conference the player played and rate their stats using a model based on previously existing but potentially flawed rankings. We will apply our ratings to the 2024-25 season to evaluate its accuracy since we have full data for that season but the end-goal is a tool that can be used to rank and project player performance when scouting transfer portal players.

## Methods

### Data Collection

#### Men's

For the men's side of the project, collecting the data was a combination of web-scraping and dataset manipulation. HDI is a company that provides a rating system for college basketball players using a variety of advanced statistics. After accessing their data via a Google Sheet that the Virginia Tech Men's Basketball team provided, the HDI ratings and stats were downloaded to a csv that could be imported into this R project. Since part of the goal includes simplify the equation that makes player ratings, the `cbbdata` package (Weatherman, 2024) was used to scrape data from barttorvik.com for the corresponding seasons. After obtaining both of these datasets the two were joined using a customized function to assign HDI ratings to almost all of the players in the barttorvik dataset. Rows that contained NA values and players that only appeared in one of the datasets were removed and special care was taken to handle difficult cases like multiple players with the same name.

```
head(mbb_raw_printing,5)
```

```
##           Name mpg  ppg rpg apg spg bpg  2%  3% ft% tov a_to efg  ts
## 1      A.J. Hoggard 26  9.6 2.6 4.6 1.2 0.2 39.0 27.1 73.4 2.2 2.08 39.5 44.4
## 2      A.J. Lopez  31 14.8 2.3 1.1 0.8 0.1 52.6 39.1 88.3 1.5 0.76 54.9 60.0
## 3 A.J. Staton-McCray 22  7.3 2.4 1.0 1.1 0.6 48.5 32.3 82.9 0.8 1.15 48.5 52.0
## 4      A.J. Wills  18  4.8 0.9 1.6 0.3 0.1 37.7 41.8 70.6 1.1 1.44 49.6 52.6
## 5      AC Bryant  24 14.0 4.4 1.6 1.4 0.1 50.6 40.7 67.7 1.2 1.29 54.9 57.9
```

This is the data we will use to assign the men’s ratings.

## Women’s

Data came from ESPN via the `wehoop` package. We used “`load_wbb_player_box()`” to collect the 2025 WBB game-by-game box scores and aggregated them into season totals. The 2025 schedule was collected with `load_wbb_schedule` and mapped each team to a conference using available team/conference fields. Manual overrides were added for teams with missing or amiguous conference labels. Minutes listed as “MM:SS” were normalized to decimal minutes, and percentage fields with zero denominators were set to NA. We filtered to Division I teams and enforced a minimum threshold (Minutes = 200, Games Played = 10).

```
head(wbb_raw_printing,5)
```

```
##      player_name mpg  ppg rpg apg spg bpg  2%  3% ft% tov a_to efg  ts
## 1    A'Niya Young 26  5.0 4.5 0.5 1.0 0.0 40.0  0.0 33.3 1.5 0.33 30.8 32.0
## 2 A'Vyonna Kinsey 32 24.0 5.0 2.0 4.0 1.0 30.0 60.0 75.0 3.0 0.67 42.0 44.8
## 3 A'riel Jackson 28  6.9 2.9 2.1 1.1 0.1 39.6 22.2 78.4 1.9 1.07 37.7 42.7
## 4      AC Markham 21  3.6 4.6 1.4 0.5 0.4 37.7 27.9 66.7 1.7 0.82 39.2 41.8
## 5      AJ Marotte 35 11.4 3.4 2.4 0.5 0.4 44.0 29.4 82.6 2.5 0.94 44.1 46.5
```

This is the data we will use to assign the women’s ratings.

## Data Manipulation

After gathering the HDI stats for the men, we trained a random forest model on the data to predict each player’s rating from per-game and efficiency statistics: on ppg (points per game), rpg (rebounds per game), apg (assists per game), spg (steals per game), bpg (blocks per game), two\_pct (2 point shot percentage), three\_pct (3 point shot percentage), ft\_pct (free throw percentage), tov (turnovers per game), ast\_to (assist to turnover ratio), efg (effective field goal percentage), ts (true shooting percentage), and ftr (free throw rate). We then applied this model to the men’s and women’s data to assign each player a “statistics score”.

On the women’s side, ESPN’s women’s data do not reliably include player positions and no stable scraping was found for positional labels, we generated our own position buckets using a rule-based approach. Players were grouped into guards, wings, or centers, using height, usage-rate, assist-rate, and rebound profile. Guards were defined by high assist rates and perimeter shooting volume, wings by balanced scoring and rebounding profiles, and centers by high rebound and low three-point attempt rates. These buckets were not created to be perfect positional labels; rather, they ensured players were evaluated with role-appropriate scoring formulas rather than a single one-size-fits-all approach. This prevented guards from being penalized for low rebounding and centers from being penalized for low assist rates as their roles within a teams context are very different. We also computed per-game, per-40, and advanced metrics (such as TS%, eFG%, and tendency rates) and converted them into percentile scores. We then used the aforementioned role weights and applied shrinkage and penalties based on minutes, role, and rotation rank. Finally, we combined the position-adjusted stat-score with team and opponent strength-of-schedule and conference strength z-scores, and rescaled the composite to a 2K-style 40-99 rating. This yields a women’s “statistics score” that incorporates both box-score production and context.

Men's:

```
head(men_4099,5)
```

##	Name	School	Rating
## 1	Ryan Kalkbrenner	Creighton	99.0
## 2	Johni Broome	Auburn	98.6
## 3	Cooper Flagg	Duke	98.2
## 4	Hunter Dickinson	Kansas	97.6
## 5	Collin Murray-Boyles	South Carolina	97.5

Women's:

```
head(women_4099,5)
```

##	Name	School	Rating
## 1	Paige Bueckers	UConn Huskies	99.0
## 2	Ally Becki	Ball State Cardinals	98.1
## 3	Katie Dinnebier	Drake Bulldogs	97.5
## 4	Serena Sundell	Kansas State Wildcats	97.1
## 5	Sarah Strong	UConn Huskies	96.9

Looking at the outputs it is obvious that we are not finished with the project since players from Ball State and Duquesne are not the top players in the nation but instead putting up excellent stats on bad teams against poor competition. This justifies the need for some form of conference weighting which we will accomplish using strength of schedule and conference performance numbers.

## Conference Strength Modeling

### Men's

For men's basketball, we estimated team strength directly from scoring margins using ESPN game results via the `hoopR::load_mbb_schedule()` function. We restricted to D1 matchups (conf IDs: 1-32) and constructed a matrix where each row encoded a game, with a +1 for the home team, -1 for the away team, and a shared intercept column. We then solved a least-squares system:  $(A^T(A) + \lambda(I))\beta = A^T(y)$  where  $y$  is the home-away score margin and  $\lambda = 10^{-2}$  is a small penalty. The resulting coefficients were centered with mean zero and then standardized to a z-score, yielding a latent team strength measure (`team_z`) for each team.

To quantify strength-of-schedule (SOS), we paired each game with its opponent and averaged opponent `team_z` values weighted by a small home/away factor (0.9 for a home games and 1.1 for road games). This produced a raw SOS measure for each team, which we standardized to `sos_z`. Finally, we aggregated to the conference level by averaging `team_z` and `sos_z` within each conference and converting these conference means to z-scores (`conf_strength_z` and `conf_sos_z`). These two standardized conference-level quantities form inputs to the conference index.

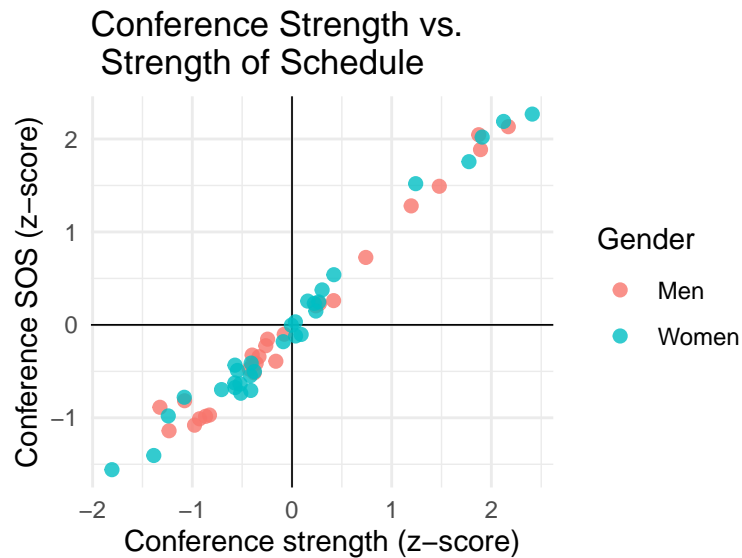
We combined conference-level strength and SOS into a single conference index:  $\text{ConfIndex} = 0.5(\text{Z\_strength}) + 0.5(\text{Z\_sos})$ . We then used that for our conference adjustment rating:  $\text{FinalRating} = \text{BaseRating} + w(\text{ConfIndex})$  where  $w$  is a tuning parameter controlling adjustment strength.

### Women's

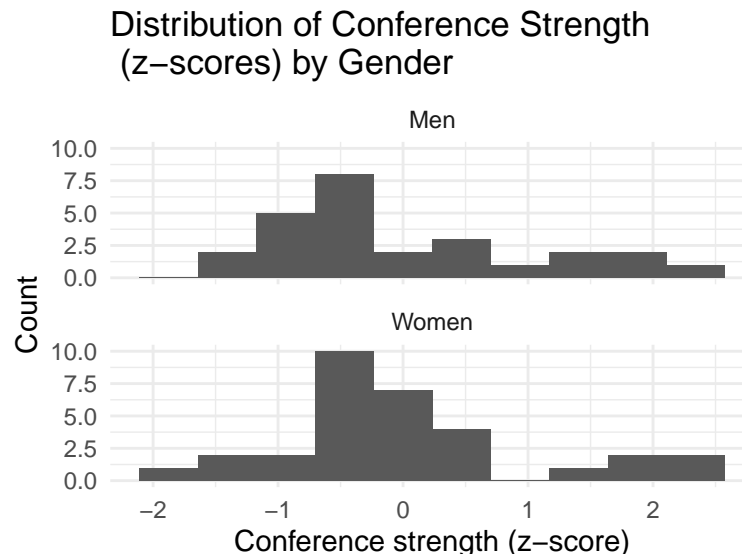
For women's basketball, we constructed team and conference strength using ESPN WBB schedules and final scores. From the schedule data, we built a game-level table with home/away team IDs and final scores, then computed score margin as home points minus away points. Using this margin data, we fit a least squares model with +1 for the home team, -1 for the away team, and a shared intercept term, mirroring the men's approach. The resulting team coefficients were centered around mean zero and standardized to obtain a team-strength z-score for each team (`team_z_map`).

We then computed strength-of-schedule (SOS) by pairing each game with its opponent and averaging opponent team-strength z-scores, using a weight of 0.9 for home games and 1.1 for road games to reflect the added difficulty of road atmospheres. This produced a raw SOS value per team, which we standardized to `sos_z`. To move from teams to conferences, we linked each team to a `conference_name` from the player table, averaged team-strength values within each conference, and standardized those conference means to obtain a conference-strength z-score `conf_z`. Finally, we merged `team_z`, `sos_z`, and `conf_z` back into the women's player table (`players_enriched_avg`) and flagged Power 5 teams (`conf_tier_pow5`) so that team strength, schedule, and conference tier could all be incorporated into the final women's rating formula.

```
print(conf_scatterplot)
```



```
print(conf_hist)
```



The histograms indicate that men's conferences vary more widely in strength, while women's conferences are more tightly clustered with only a few elite outliers.

The scatterplot confirms a strong positive relationship between conference strength and schedule difficulty for both genders. Stronger conferences consistently play harder schedules, while weaker conferences play softer

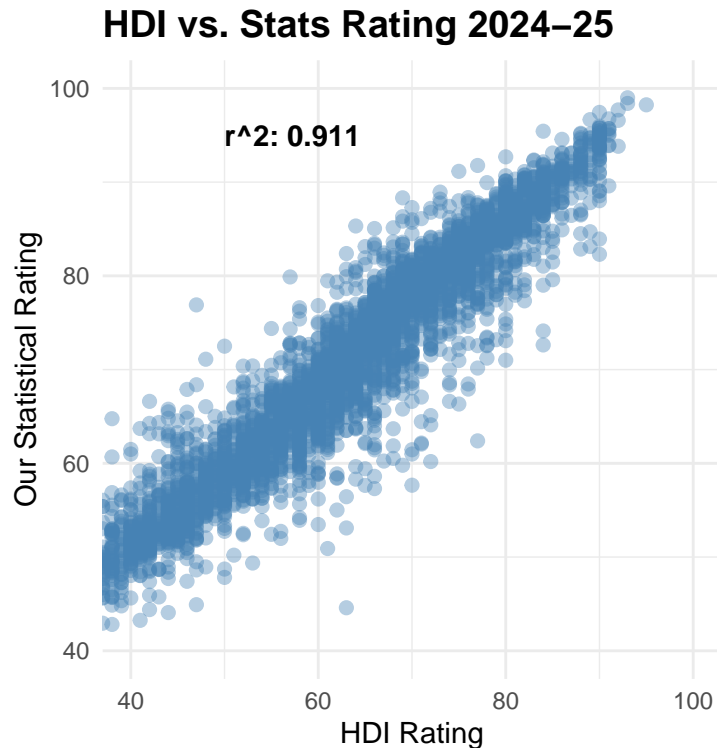
ones.

These findings demonstrate that conference environment has a meaningful impact on player's stats and supports the need for conference-adjusted ratings in our model.

## Analysis

To evaluate how well our statistical model replicates the original HDI ratings, we compared our model's predictions against the actual HDI ratings for the men's basketball players:

```
print(plot)
```



Since our ratings were scaled to be 40-99 (similar to NBA2K, a basketball video game) and have not been adjusted for conference weight yet, this shape is perfectly reasonable and an indication that the statistical portion of the code is working. There are only a few outliers and the narrowing of the graph at the top is to be expected given a fewer number of elite players. A 91%  $r^2$  value indicates that the stats we used in the model account for 91% of the variability in the ratings which is a great report given that we could not scrape many of the more advanced statistics for the women's players.

## Potential Limitations (so far)

### Mens to womens model translation

One of the problems we ran into was applying a model trained on men's data to the women's game, the original outputs were too low (top players were in the 70s) due to the different paces and styles so this led us to scaling the ratings.

### Adjusting for conferences on HDI ratings

As we have begun to work on adjusting for conferences on the HDI ratings it is difficult to bring down top mid-major players without destroying the middle tier players from lower conferences so we are still working

on a way to balance this out.

### **Heavy reliance on box-score stats**

Both the men and women’s models are constrained to box-score variables because play-by-play and advanced tracking data are unavailable for the full dataset. This limits the model’s ability to capture defensive impact, screening, spacing, off-ball value, and other impacts players may have that aren’t captured in the traditional box-score.

### **Margin-based team strength models assume linearity**

Our team strength estimates rely on least-squares systems that use scoring margins. Margin-based ratings assume linear, symmetric behavior and may overweight blowouts or undervalue teams that eek out close games, especially close wins versus good teams.

### **Small conferences harder to adjust for**

Conferences with fewer teams or limited non-conference scheduling leads to unstable conference z-scores that make estimating their strength less reliable.

### **Transferability to future seasons untested**

The model is tuned and validated on the 2024-25 season. It’s stable across single seasons, but tests across multiple seasons, roster turnover patterns, and NIL-induced shifts are not yet validated.

## **References**

Weatherman A (2024). `cbbdata`: API for College Basketball Data. R package version 0.3.0.9000, <https://cbbdata.aweatherman.com/>

Kartes, Weston. MBB Player Database 2024. HD Intelligence. <https://www.hdintelligence.com/mbb-player-database-2024>

Kartes, Weston. MBB Player Database 2025. HD Intelligence. <https://www.hdintelligence.com/mbb-player-database-2025>

Torvik, Bart. NCAAM Player Stats 2024. barttorvik. <https://barttorvik.com/playerstat.php?year=2024>

Torvik, Bart. NCAAM Player Stats 2025. barttorvik. <https://barttorvik.com/playerstat.php?year=2025>

ESPN Women’s college basketball data (via `wehoop`)

ESPN 2025 schedules and boxscores (via `wehoop`)

Hird, K. `wehoop`: Women’s Basketball Data for R: <https://github.com/sportsdataverse/wehoop>