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 ups 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 webscraping 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 developed by Andrew Weatherman 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 assing 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. I pulled 2025 WBB game-by-game box scores with `load_wbb_player_box`, then aggregated to season totals per athlete. I pulled the 2025 schedule with `load_wbb_schedule` and mapped each team to a conference using available team/conference fields. I also allowed manual overrides via a text file for cases where a conference couldn't be established. I filtered to D1 conferences only and enforced basic minimums (mins = 200, min_games = 10). Output is a cleaned player table with player totals and conference labels. Some assumptions to note are: ESPN IDs are stable within season. Minutes may appear as "MM:SS", so I normalized to decimal minutes. If a percentage denominator is zero, the percentage is set to NA.

```
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 a players rating based 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".

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
```

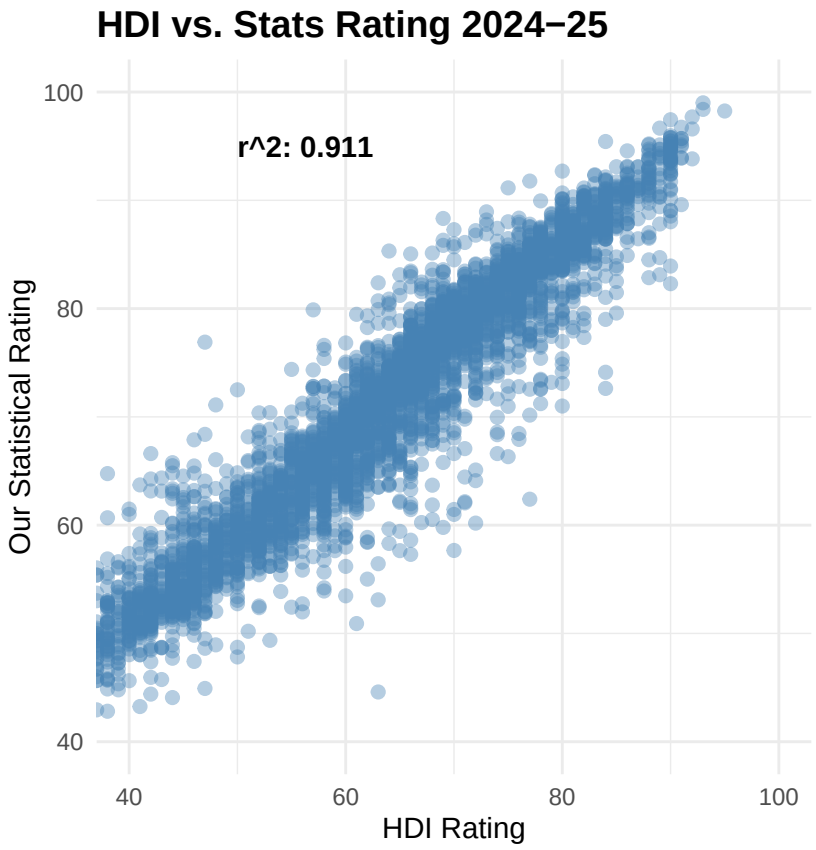
```
## 1  Paige Bueckers         UConn Huskies   99.0
## 2      Ally Becki  Ball State Cardinals   98.1
## 3 Katie Dinnebier        Drake Bulldogs   97.5
## 4  Serena Sundell Kansas State Wildcats   97.1
## 5    Sarah Strong         UConn Huskies   96.9
```

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.

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)

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.

As we have begun to work on adjusting for conferences 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.

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