Basketball Data Madness Challenge 2020

Jen Sheng Wong, Jin Fu Ooi, Kai Xuan Shau, Wen Hoong Ling

Research Question

Based on players' performance loads, can we determine factors that result in effective training sessions as well as predicting the quality of future training sessions?

Our Assumptions

- We assume the best predictor for a Guard's performance is `assists_turnover_ratio`. We then constructed a predictor, `assists_turnover_ratio_pct`, which is the percentage of "assist turnover ratio" by Wolverines against opponent.
- Assume best predictor for a Wing is `three_points_pct` and the best predictor for a Post is `rebounds`. Percentage predictors are also created.
- Assume training sessions for the particular game happened at most 4 days before the game. A game with all its relevant training sessions will all be assigned the same `unique_session`.

How We Arrived at `players_df`

1. Performed a left join on `game_df` and `catapult_df` based on date, so that every entry in `catapult_df` has `game_df` information.
2. `good_perf` is our response. Data is separated into 3 segments based on position Wing, Guard and Post. For `guard_df`, assign `good_perf = 1` when `assists_turnover_ratio_pct` is greater than 75th quantile. Applied same method to both `wing_df` and `post_df` based on their percentage predictors.
3. Combined `guard_df`, `post_df`, and `wing_df`.
4. Removed performance during an actual game.
5. Normalized catapult data based on players.

Position: Guard

Zavier Simpson

3-Point Pct (%) : 31.8

Rebounds Per Game : 4.233

Assist Per Game : 6.067

** Average ratings from the past 3 seasons

Source: University of Michigan “2019-2020 Men’s Basketball Roster”
https://mgoblue.com/sports/mens-basketball/roster/zavier-simpson/20388

Position: Post

Jon Teske

3-Point Pct (%) : 18.2

Rebounds Per Game : 5.667

Assist Per Game : 0.800

** Average ratings from the past 3 seasons

Source: Source: University of Michigan “2019-2020 Men’s Basketball Roster” https://mgoblue.com/sports/mens-basketball/roster/jon-teske/20389

Position: Wing

Isaiah Livers

3-Point Pct (%) : 39.7

Rebounds Per Game : 3.400

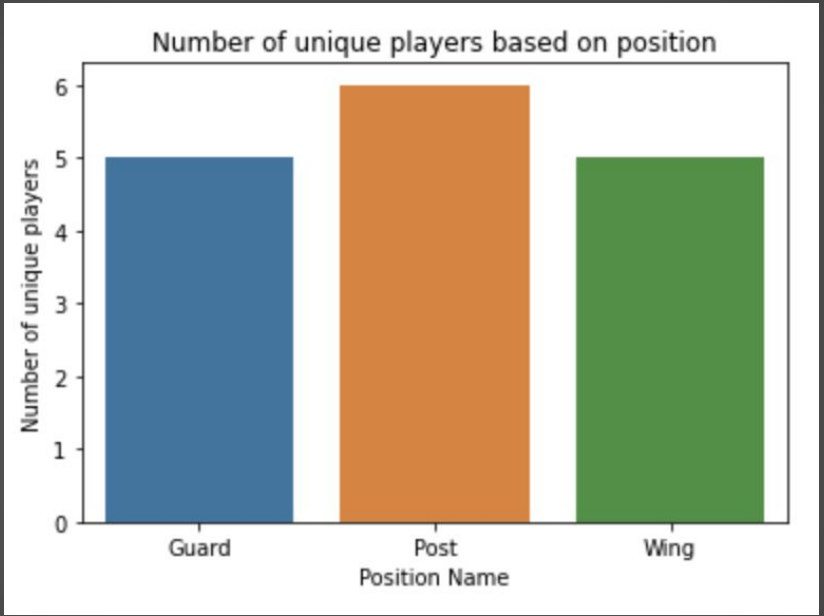
Assist Per Game : 0.7333

** Average ratings from the past 3 seasons

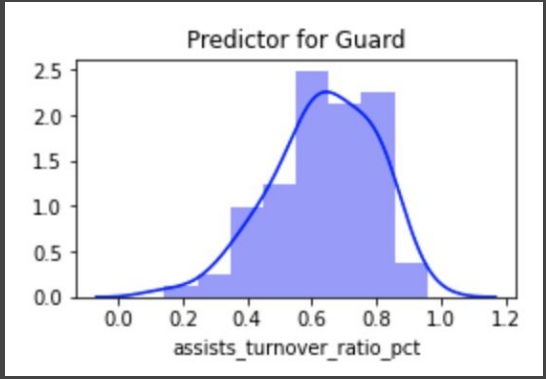
Source: Source: University of Michigan “2019-2020 Men’s Basketball Roster” https://mgoblue.com/sports/mens-basketball/roster/isaiah-livers/20385

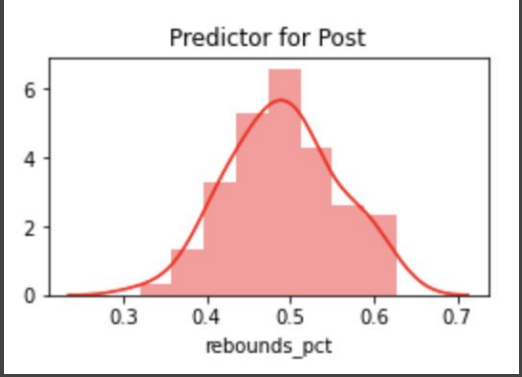
Exploratory Data Analysis

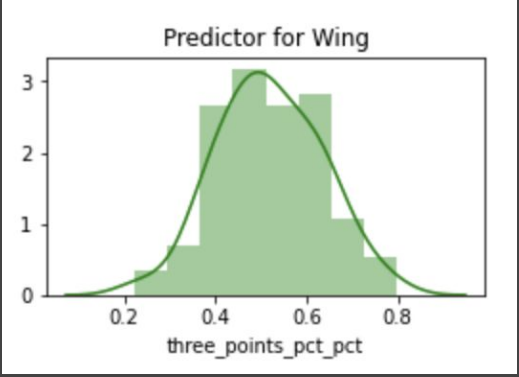
The dataset consists of 6 Posts, 5 different Guards and Wings.



Exploratory Data Analysis



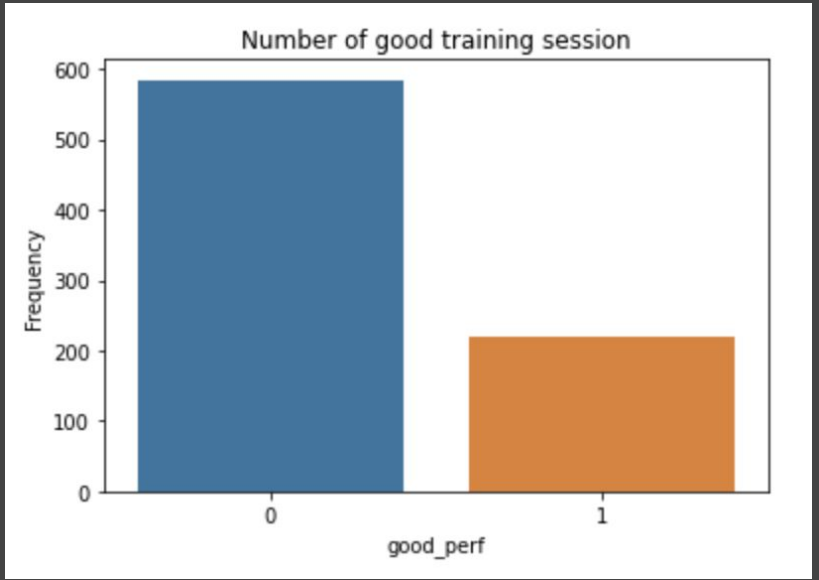




The predictor for each position is relatively normal, indicating less or no presence of outliers.

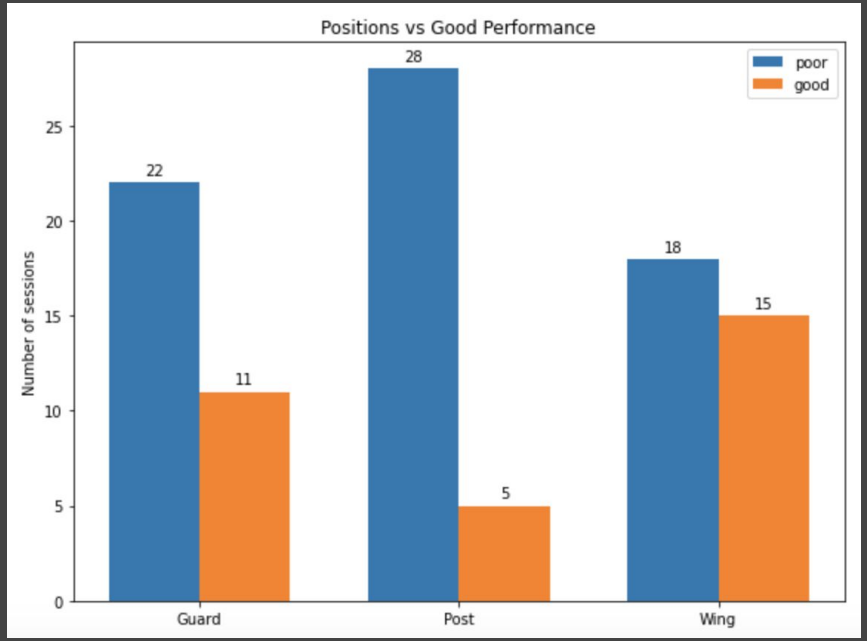
Exploratory Data Analysis

Our data frame has close to 600 poor and 200 good training sessions



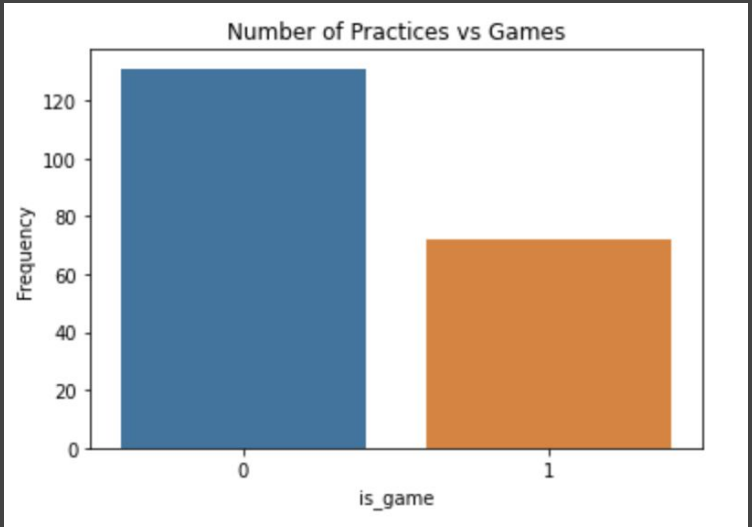
Exploratory Data Analysis

The plot shows the number of poor and good performance based on positions.



Exploratory Data Analysis

Our data frame has around 120 training sessions and 60 games



Our Model

- Our models used Lightgbm, a decision tree algorithm
- K-Fold cross validation, k = 5

Position	Guard	Post	Wing
Average AUC ROC	0.677	0.553	0.567

Our Model

We are predicting the quality of training sessions for each position based on the performance loads of the players during training sessions.

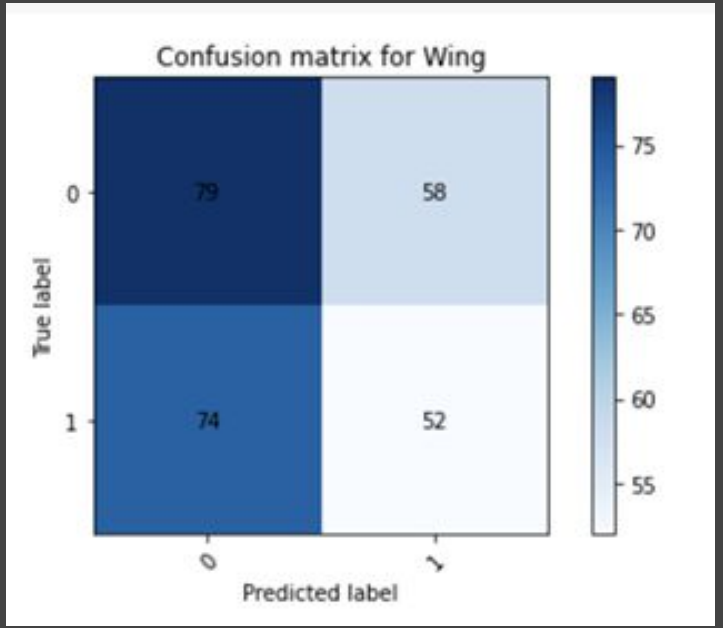
0 (bad quality) - If that position performs poorly in the game after that particular training session

1 (good quality) - If that position performs well (predictor above the 75th percentile) in the game after that particular training session

* We defined performance of players in the game based on game dynamics
  (Post - Rebound per game ; Guard - Assist per game ; Wing - 3 pts %)

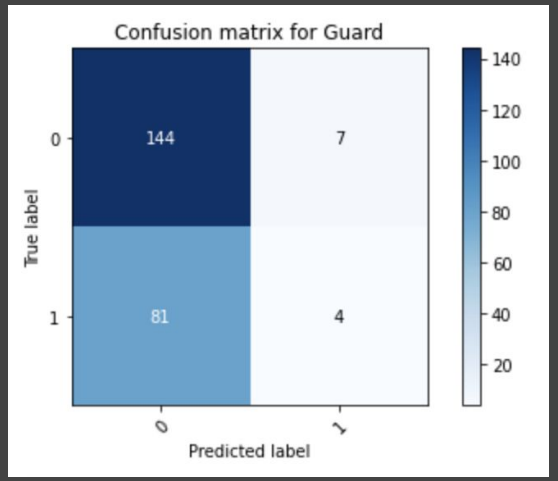
Our Findings

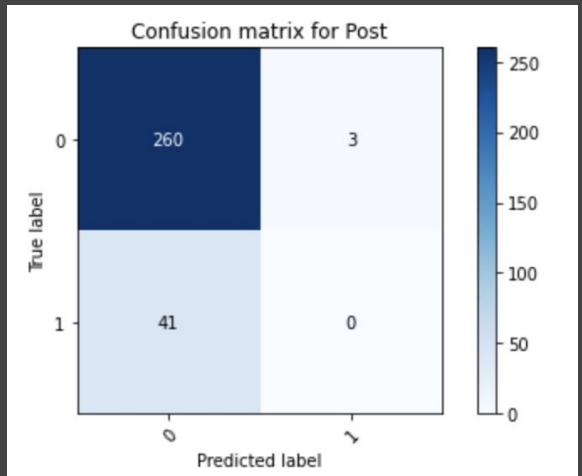
Based on our confusion matrix, the prediction for wing is not as good and this might imply performance loads do not really affect the performance of wing player (dynamic of 3 points pct).



Our Findings

From the model for post and guard, we can definitely see a better prediction compared to wing position.



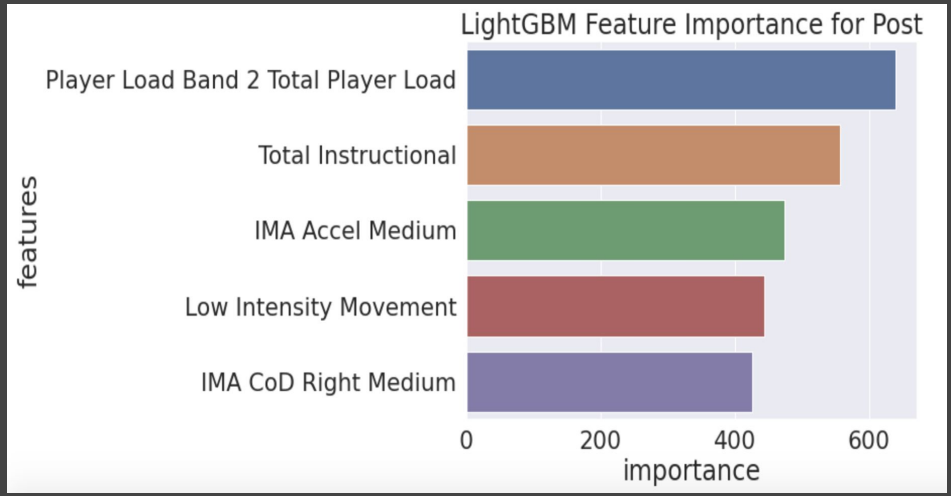


However, note that in our data there are more records with poor training than that with good training and so the prediction of poor training is better.

Our Findings

The most important feature for Post:
Total Player Load in Band 2

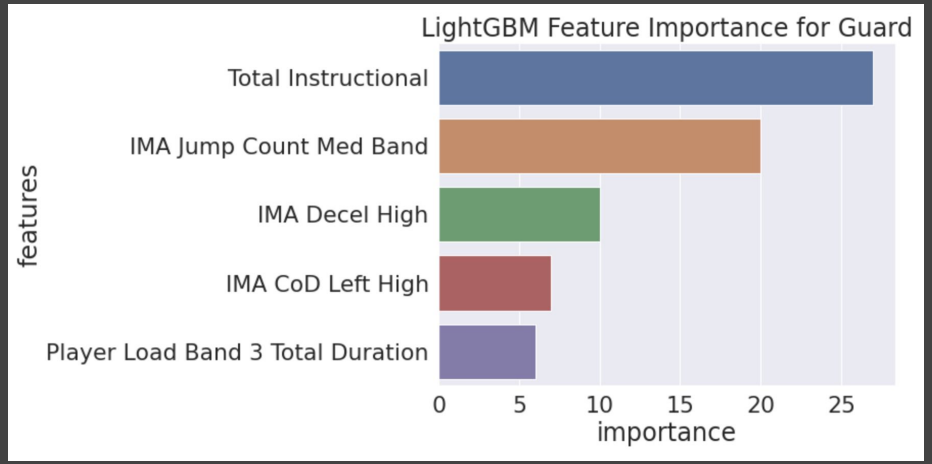
Compare to other position, post player is the closest to the rim and needs to possess strength and ability to catch a rebound. Thus, it is crucial to stay alert and react to opponent's move at any direction. This is coherent from our findings but it is worth mentioning that a high acceleration is not required to achieve such goal.



Our Findings

The most important feature for Guard:
Total Instructional Time

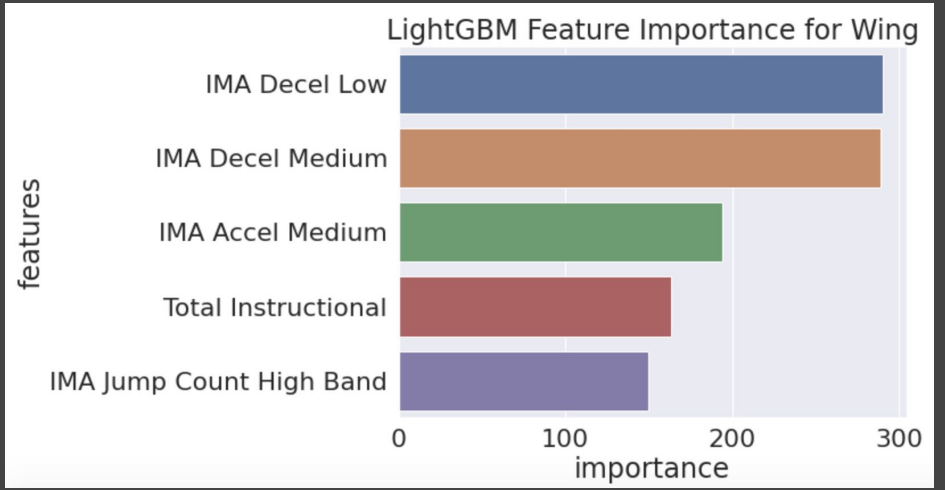
As a guard player aka “coach on the floor”, he needs to know precisely how the game plan works and strike an offensive play at the right time. Hence, it is crucial for him to spend more time on the floor to have a clear grasp on the court situation in order to create scoring opportunities for the team.



Our Findings

The most important feature for Wing:
Deceleration (Low & Medium)

The result makes sense because as a wing player, he needs to slow down in a stable pace to make an accurate shot from far especially 3pts shot. However, our model shows that decel slow and decel medium are equally important, suggesting that the classification based on IMA rubic might not be exactly precise.



Our Findings

- All positions exhibit different important features which implies the performance loads contribute differently to quality of training sessions for different positions in the team.
- Although not perfect, our methodology shows that we can predict if a training session is good or bad.

Model Application

Based on the feature importance, training sessions can be tailored for each positions.

Work on your specific training sessions, boys!

Our Conclusion

- With the performance loads of the players, we can predict if the training sessions suit the players for each position.
- If the training sessions predicted to be bad, we recommend coaches to focus on specific factors for different position, which could lead to useful training sessions.

Our Challenges

- Our models heavily depend on our assumptions such that we only look into certain game statistic for each position.
- Lack of data, only, 40 unique_sessions since some games occur at different dates - cannot join with catapult dataframe
- It would be better to account for opponent's performance - sometimes opponents are too strong


Thank You
DUKE
3
5
13
50
MICHIGAN
WOLVERINES
NCAA