Daniel Hyun Kyu Im – Redefining NBA Positions Using Machine Learning

im.daniel04@gmail.com

July 26, 2022

## Jupyter Notebooks

1) NBA Part 1.ipynb: Data collection & initial cleaning
2) NBA Part 2.ipynb: EDA, Modelling & Insights

## Dataset

The data was web-scraped from NBA.com & basketballreference.com. For part 2 of the project, the dataset required: NBA Datasets/'NBA_FINAL_DATA.csv'

## Report/Summary

Please refer to datadrivendaniel.wordpress.com

## Environment

List of packages in environment used:

python=3.8, pandas, matplotlib, seaborn, statsmodel, scikit-learn=0.24.1, jupyter, jupyterlab, beautifulsoup4, selenium, pyclustertend

## Project Journey

Part 1 begins with the collection of data using Beautiful Soup and Selenium (FireFox webdriver). Cleaning is done concurrently and the notebook concludes with the final CSV file that will be used for the remainder of the project.

Part 2 starts with brief EDA followed by modelling (KMeans clustering). After some modelling, there is a portion where the results of the clusters are examined and additional insights provided (ie. Random Forest model for feature importance). Lastly, to showcase an **example** of what practical implementation of this project would look like, there is a brief section on making a recommender system as well.

## NBA Statistics Glossary:

There are about 80 features used for this project, so below is a description of all statistics used (only the final 80 statistics is listed here and not the original 120, since there were unused). There are some inconsistencies when it comes to the data used (for example, there is a 3FGM_%AST but NO 2FGM_%AST) because they may have been removed from the original 120 during feature selection stage

PTS: points scored

FG%: Field goal %

3PA: 3 pointers attempted

3P%: 3 point field goal %

FT%: free throw %

OREB: offensive rebounds

REB: total rebounds

AST: assists

TOV: turnovers

STL: steals

BLK: blocks

PF: personal fouls

DD2: double doubles

TD3: triple doubles

HEIGHT: height

WEIGHT: weight

OFFRTG: offensive rating (measure of offensive output)

DEFRTG: defensive rating (measure of defensive output)

NETRTG: OFFRTG – DEFRTG

USG%: Usage %

PACE: number of possessions per 48 minutes

PIE: player impact estimate (measure of player's "impact" – similar to PER but includes defensive stats)

AST%: percentage of teammate field goals a player assisted on

AST/TO: assists/turnovers

OREB%: percentage of available offensive rebounds a player obtained

REB%: percentage of available rebounds a player obtained

TO_RATIO: number of turnovers a player averaged per 100 possessions used

TS%: true shooting %

PTS_OFF_TOV: points scored off opponent turnovers

2ND_PTS: second chance points

FASTB_PTS: points scored on fast breaks

PAINT_PTS: points scored in the paint

OPP_PTS_OFF_TOV: opponents points scored off player turnover

OPP_2ND_PTS: opponents second chance points

FGA_BLOCKED: field goal attempts blocked

PF_DRAWN: personal fouls drawn

%PTS_MidRange: percentage of points scored from midrange

%PTS_FASTB: percentage of points scores on fast break

%PTS_FT: percentage of points scored from free throws

%PTS_TO: percentage of points scored from opponent turnovers

%PTS_PAINT: percentage of points scores in the paint

2FGM_%UAST: percentage of unassisted 2-point field goals made

3FGM_%AST: percentage of assisted 3-point field goals made

3FGM_%UAST: percentage of unassisted 3-point field goals made

%FGM_UAST: percentage of unassisted field goals made

%TEAM_3PM: percentage of team's 3-point field goals made

%TEAM_TOV: percentage of team's total turnovers

%TEAM_STL: percentage of team's steals

%TEAM_BLOK: percentage of team's blocks

%TEAM_PF: percentage of team's personal fouls

%TEAM_PFDRAWN: percentage of team's personal fouls drawn

%TEAM_PTS: percentage of team's points

SCREEN_AST: the number of times an offensive player sets a screen for a teammate that directly led to a made field goal by that teammate

LOOSE_RECOVERED_OFF: loose balls recovered on offense

LOOOSE_RECOVERED_DEF: loose balls recovered on defense

LOOSE_RECOVERED_TOTAL: total loose balls recovered

%LOOSE_RECOVERED_DEF: percentage of loose balls recovered that were on defensive end

CHARGES_DRAWN: charges drawn

CONTEST_2PT: number of 2-point shots contested by player

CONTEST_3PT: number of 3-point shots contested by player

<8FT_FG%: field goal percentage for shots from within 8 feet of the basket

8-16_FGA: field goal attempts from between 8-16 feet away from the basket

8-16_FG%: field goal percentage from between 8-16 feet away from the basket

16-24_FGM: field goals made from between 16-24 feet away from the basket

16-24_FG%: field goal percentage from between 16-24 feet away from the basket

DRIVE_PTS: points scores from drives to the basket

DRIVE_FG%: field goal percentage for drives to the basket

CATCHSHOOT_PTS: points score from catch and shoot shots

CATCHSHOOT_FG%: field goal percentage from catch and shoot shots

PULLUP_PTS: points score from pullups

PULLUP_FG%: field goal percentage for pullup shots

PAINTTOUCH_PTS: points scored by player on touches in the paint

PAINTTOUCH_FG%: field goal percentage by player on touches in the paint

POSTTOUCH_PTS: points scored by player on touches in the post

POSTTOUCH_FG%: field goal percentage by player on touches in the post

ELBOWTOUCH_PTS: points scored by player on touches in the elbow

ELBOWTOUCH_FG%: field goal percentage by player on touches in the elbow