Ranking NBA statistics using XGBoost Classification

Gene Zaleski
Data Mining II - Spring 2022
https://github.com/genezaleski/classify_nba_stats

Table of Contents

Why evaluate NBA Statistics like this?

[JJ Redick] The Kings just traded their best player. Yes Fox averages more points, but all of the **advanced stats**, it's been Tyrese, and the 2nd best player has been Harrison. That is a fact. I don't give a f*ck about how many points a guy averages, that doesn't mean sh*t.

4,671 points • **799 comments** submitted 2 months ago by JoeBiden2020FTW to r/nba

JJ was getting heated, he really likes Haliburton and was getting tired of all the casual fans quoting PPG at him.

https://youtu.be/LKirW6CndOk?t=175

more

Where can I find **advanced stats** besides BBref?

35 points • **45 comments** submitted 2 days ago by Winter_Practice5678 to r/nba

BBref has win shares, VORP, and offensive/defensive BPM, but I've found myself in friendly arguments with redditors who'll pull out stats like RAPTOR, DARKO, LEBRON, RAPM, etc.

more

[OC] Using **advanced stats** to compare Devin Booker to the three MVP frontrunners

78 points • **222 comments** submitted 1 month ago by s1mpleGOAT to r/nba **Original Content**

Player	GP	EPM	LEBRON	RAPTOR	RPM	RAPM	WS/48	BPM	Average Rank
Nikola Jokic	67	8.5	7.58	14.3	11.54	3.24	0.295	13.7	1.00

more

How much impact have **advanced stats** had on this seasons MVP race and is defensive prowess not considered much when determining the MVP?

13 points • **62 comments** submitted 8 days ago by throwawayfor_now to r/nba

Disclaimer: I am by no means an expert at being an NBA analyst and my knowledge about topics probably isn't the best or accurate. That being said that is probably the case for a majority of this sub so give me a break.

more

Advanced Stats Visualized for MVP Candidates - Jokic head and shoulders above the competition

19 points • **124 comments** submitted 1 month ago by dairyman777 to r/nba

https://streamable.com/5hi6ij

Nikola Jokic's PER this season is by far the best all time and he is leading nearly every advanced stat in the league.

1,960 points • **375 comments** submitted 3 months ago by metrodome93 to r/nba

https://www.basketball-reference.com/leaders/per_season.html

https://www.basketball-reference.com/leagues/NBA_2022_advanced.html

more

Zach Lowe: "It's beyond stupid that we frame the MVP race as a two-man race… Pick any advanced stat you want, Jokic leads the NBA in every single one… Nikola Jokic belongs in the race."

6,576 points • **1,007 comments** submitted 4 months ago by GuyCarbonneauGOAT to r/nba

https://streamable.com/qybaxw

Advanced Stats Hall of Shame

20 comments submitted 1 hour ago by IamNotStupidUR to r/nba

This is something that I always found funny when the perception of a players contributions are out of line with their advanced stats contributions. I bring this up because I took a look at Rodman's advanced stats and while his

more

Historical MVP based on **advanced stats**.

0 points • **44 comments** submitted 14 days ago by jamoalki to r/nba

With all the talk about Jokic running away as the clear MVP if it were only based on advanced stats, who would be past winners?

[Kevin Pelton] The NBA **advanced stats** that matter most in the MVP race

0 points • **78 comments** submitted 20 days ago by RN1337 to r/nba

https://www.espn.com/nba/insider/story/_/id/33701327/nba-mailbag-advanced-stats-say-mvp-battle-nikola-jokic-joel-embiid-giannis-antetokounmpo

NoDunks on youtube did an MVP scoreboard for the three main candidates for the following categories - Record, Stats, **Advanced Stats**, defense, highlights, head to head and Narratives.

- NBA discourse has become saturated with "Advanced Stats" attempting to quantify a player/team's overall performance.
- These stats can be interpreted in many different ways, so there is a lot of noise about how they can be accurately used.
- *Why should you care*?
  - **If you are involved in any sports discussion, you likely will see these stats cited in comparisons or rankings.**
  - **If you care more about the analytics aspect of these stats, this is a test of the models professional data scientists have created.**
  - **Data literacy is very important!**
- *What is original about this research?*
  - **I am essentially creating a new statistic, which evaluating the effectiveness of other statistics.**

Left - Example of how frequently these statistics are cited in Reddit NBA discourse

Why evaluate NBA statistics like this? - Problem Statement/Approach

- ***Problem:*** *In an effort to sort through all the noise regarding nba statistics, I wanted to rate each of these statistics.*
- ***Approach:***
    - *Since these stats are used so frequently to decide who is the best at X, rating these stats requires knowing <u>which stats can accurately predict who is the best at X</u>.*
    - *Using XGBoost classification, we can use any statistic to classify achievements of measures of success in the NBA, then rank said statistic by their classification accuracy.*

Data Collection and Cleaning

- https://www.nba.com/stats/players/advanced/ contains nba advanced stats for both players and teams dating back to 1996.
- Their API allows for requests to be made directly to endpoints retrieving advanced and basic stats for both players and teams.
- You can download stats as csv for every year available by specifying it in a url via requests python library

This process was modelled after the procedure found here:
https://towardsdatascience.com/how-scraping-nba-stats-is-cooler-than-michael-jordan-49d7562ce3ef

Data Collection and Cleaning (Example)

- Iterate over all years, replace year in endpoint, and convert returned json to csv

```
yearString = str(ii) + "-" + str(upperYear)
currURL = url.replace("2021-22",yearString)
if not exists("/home/gene/Documents/DataMiningII/Project/getNBAcom/teamAdvancedCSV/"+ yearString +"_regular.csv"):
    response = requests.get(currURL, headers=header)
    response_json = response.json()
    frame = pd.DataFrame(response_json['resultSets'][0]['rowSet'])
    frame.columns = response_json['resultSets'][0]['headers']
    frame.to_csv("/home/gene/Documents/DataMiningII/Project/getNBAcom/teamAdvancedCSV/"+ yearString +"_regular.csv",sep=",",header=frame.columns)
```

Data Collection and Cleaning

- The next step of the process was to combine all scraped data into master datasets, one for each of these four categories: player, player playoffs, team, team playoffs.
- For each of the above categories:
    - CSV for each year were combined to join advanced and regular stat CSVs by columns, yielding close to 100 unique stats.
    - Assign new “Year” column to maintain order of stats.
    - Combine all CSV for each year into one.

Assigning Labels for measures of success

```
#!/usr/bin/sh
file=$1
filteredFile=$file"_filtered"

#Get correctly formatted columns
awk '{print $1 " " $2 "," $4 " " $5}' $file > $filteredFile
# replace newlines with commas
sed -zi 's/\n/,/g;s/,$/\n/' $filteredFile
# replace duplicate commas
sed -i "s+,\s++g" $filteredFile
# new line on years starting with 2
sed -i "s+,2+\n2+g" $filteredFile
# new line on years starting with 1
sed -i "s+,1+\n1+g" $filteredFile

# destroy the evidence
rm $file
mv $filteredFile $file
```

```
2022,LeBron James,Giannis Antetokounmpo,Stephen Curry,DeMar DeRozan,Nikola Jokic,Luka Doncic,Darius Garland,Chris Paul,Jimmy Butler,Donovan Mitch
ell,Fred VanVleet,Jarrett Allen,Joel Embiid,Ja Morant,Jayson Tatum,Trae Young,Andrew Wiggins,Devin Booker,Karl-Anthony Towns,Zach LaVine,Dejounte
 Murray,Khris Middleton,LaMelo Ball,Rudy Gobert
```

- Basketball is often hard to quantify, as there are many variables contributing to various outcomes.
- To attempt to account for this, we can look at multiple different measures of success to see how accurate statistics classify multiple contexts.
  - Players were classified as All Stars, MVPs, or Finals MVPs.
  - Teams were classified into Finals winners and losers.
- I didn't find any good sites to scrape this information from, so I just copied lists from ESPN.com and used awk, sed, etc. to format my this data into lists of names and years in CSV format.

Assigning Labels for measures of success

- Once lists of names and years for the players and teams were formatted, I could assign labels of 1 and 0 for matching columns when a player or team achieved said success.
- Assigned labels where indices of CSV matched Names & Years.

```
with open(champpath,'r') as allStarsFile:
    for line in allStarsFile:
        allStars = line.split(",")
        year = int(allStars[0].strip())
        allStars = allStars[1:]
        for player in allStars:
            nidx = teamRegularSeason[teamRegularSeason['TEAM_NAME']==player.strip()].index.values
            yidx = teamRegularSeason[teamRegularSeason['YEAR']==year].index.values
            nidx1 = teamPlayoffs[teamPlayoffs['TEAM_NAME']==player.strip()].index.values
            yidx1 = teamPlayoffs[teamPlayoffs['YEAR']==year].index.values
            regIdx = np.intersect1d(nidx,yidx)
            playoffIdx = np.intersect1d(nidx1,yidx1)
            if regIdx.size > 0:
                teamRegularSeason['WIN'][regIdx[0]] = 1
            if playoffIdx.size > 0:
                teamPlayoffs['WIN'][playoffIdx[0]] = 1
```

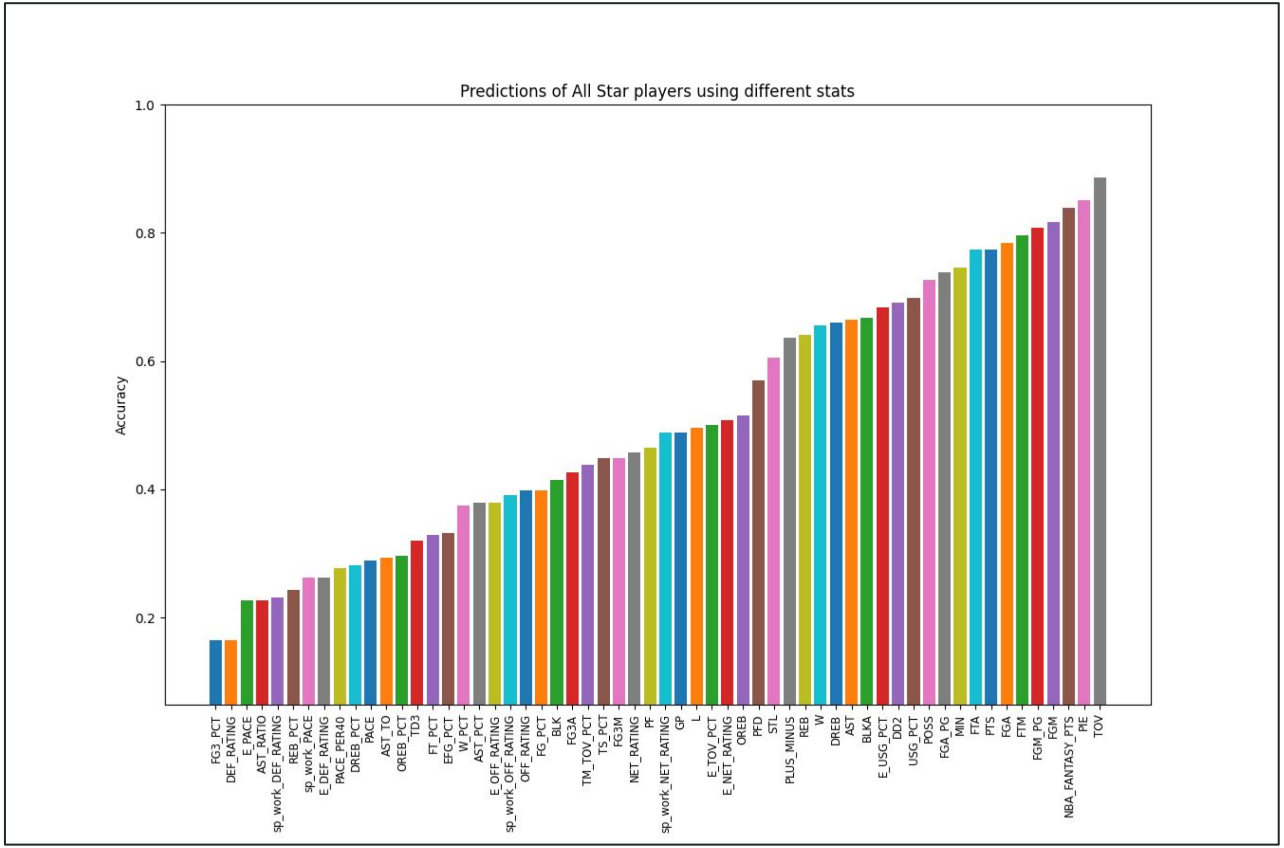
Data Classification

- With clean, labelled data, classification is now possible.
- Because there are limited amounts of “True” entries in my testing data (i.e. only 25/~12300 players in the data are labelled MVP), SMOTE was utilized to increase the number of entries labelled “True” in the data.
- Used XGBoost in Python.
    - Why XGBoost?
        - Data is highly structured.
        - Dataset is small(ish)
        - XGBoost is the fastest and most accurate Classification technique for structured data.
- Iterated over all stats, fit XGBoost with said stat and labels, then compared the accuracy!

Data Classification (example)

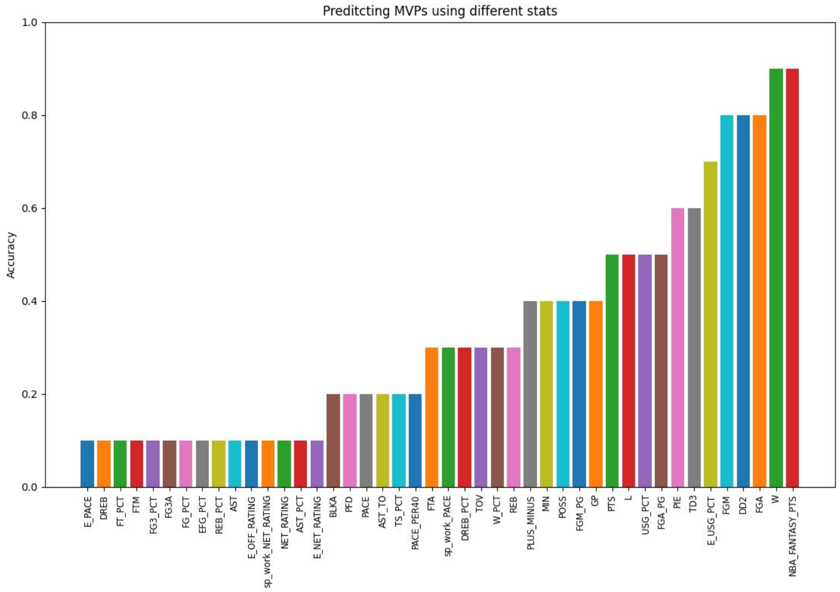
```
y = data[label]
oversample = BorderlineSMOTE()
for columnName,columnData in data.iteritems():
    if columnName in drops:
        continue
    elif "Unnamed" in columnName:
        continue
    elif columnName == label:
        continue

    print(columnName)
    X = data[columnName]
    X_train,X_test,y_train,y_test = train_test_split(X,y,random_state=42,stratify=y,test_size=0.3)

    X_train,y_train = oversample.fit_resample(X_train.to_frame(),y_train)

    xg = XGBClassifier()
    xg.fit(X_train.squeeze(),y_train)
    predictions = xg.predict(X_test)

    out = classification_report(y_test,predictions,output_dict=True)
    accuracies.append(out['accuracy'])
    stats.append(columnName)
```

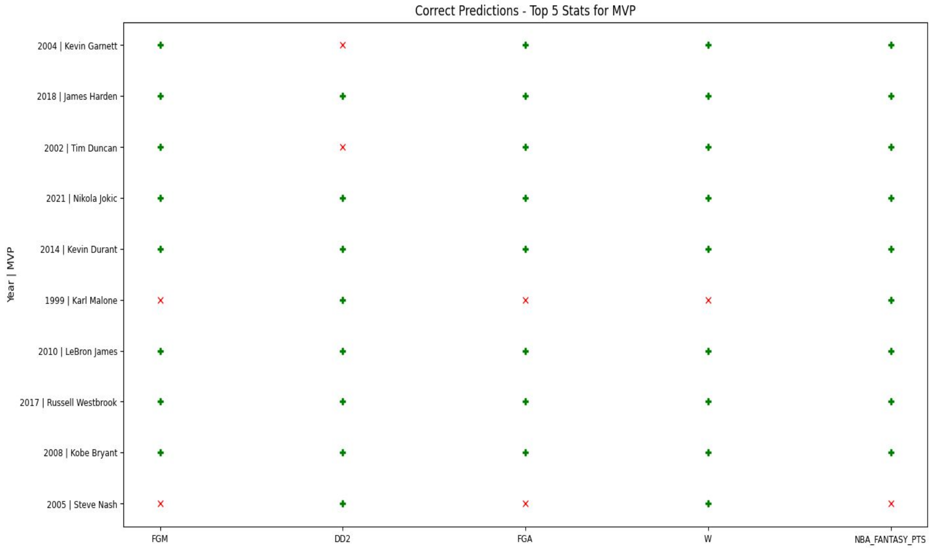
Results - Classifying All-Stars



- To yield a legitimate accuracy, predictions were only evaluated for True Positives, False Negatives.
- Certain stats were not considered due to them being duplicates (PIE_RANK is not evaluated, only PIE.)

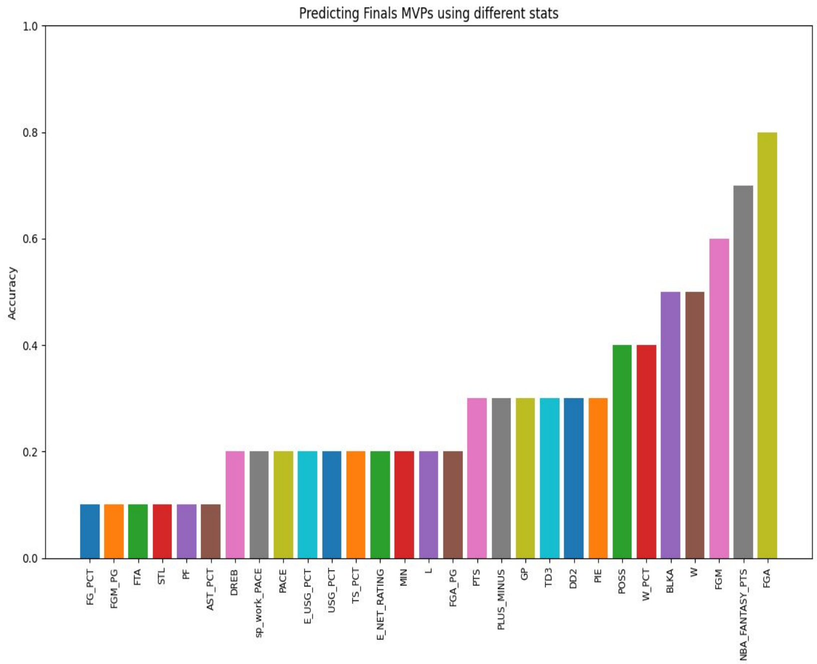
Results - Classifying MVPs

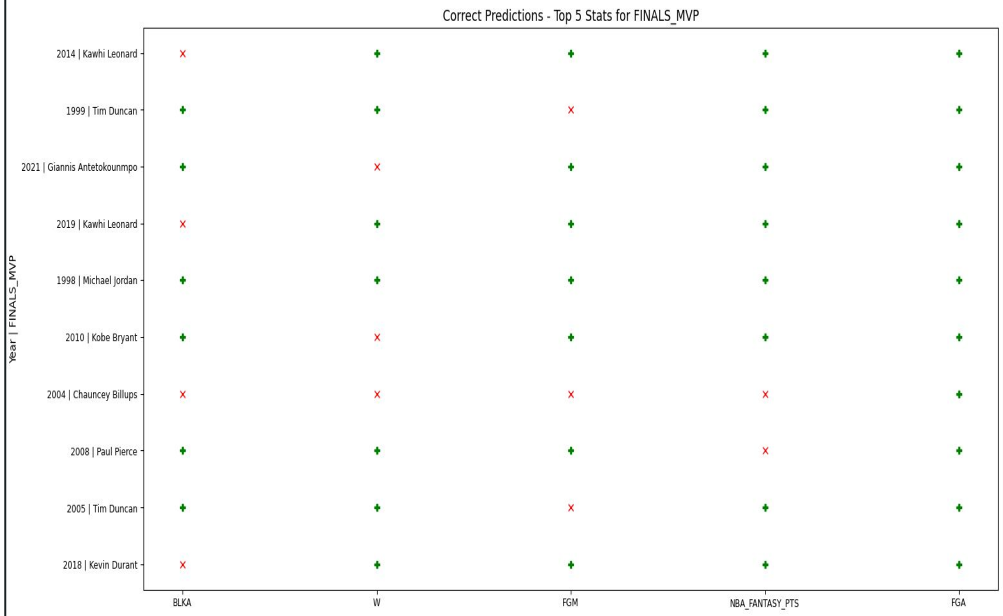




Not Pictured:
sp_work_OFF_RATING,E_DEF_RATING,DEF_RATING,sp_work_DEF_RATING,AST_PCT,AST_RATIO,OREB_PCT,TM_TOV_PCT,E_TOV_PCT,FG3M,OREB,STL,BLK,PF

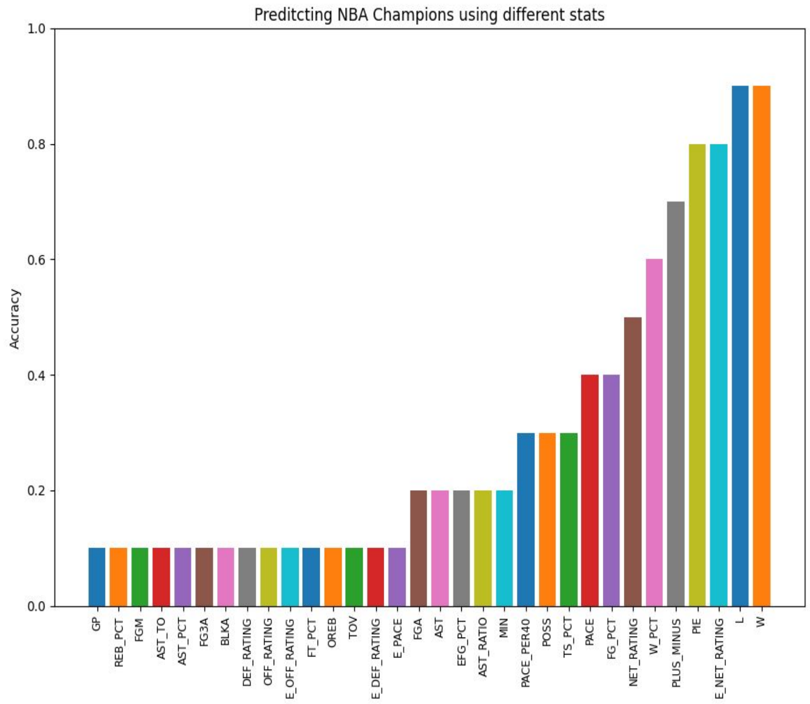
Results - Classifying Finals MVPs

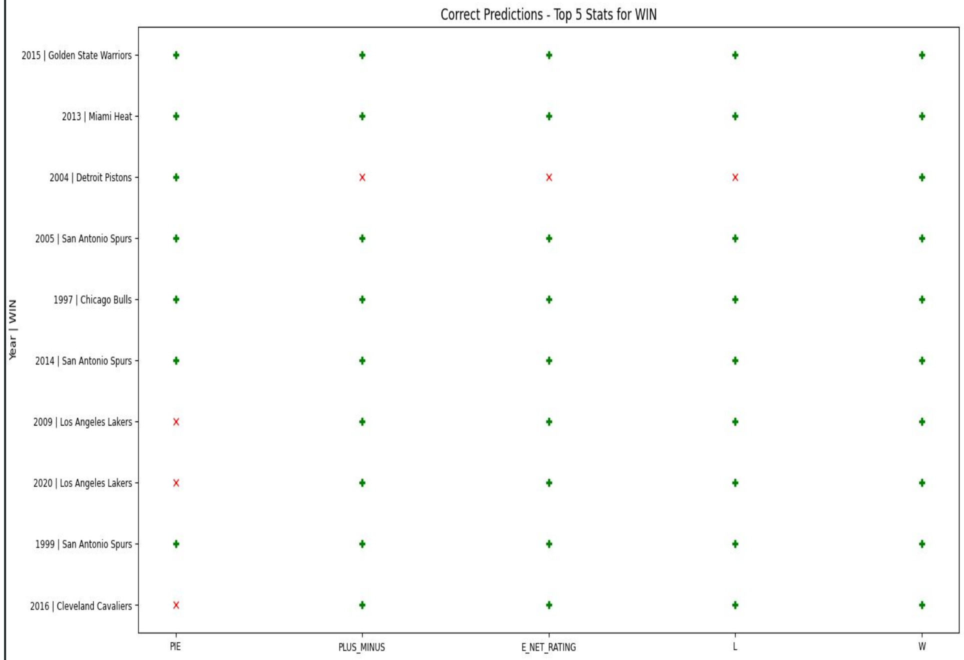




Not Pictured:
E_OFF_RATING,OFF_RATING,sp_work_OFF_RATING,E_DEF_RATING,DEF_RATING,sp_work_DEF_RATING,NET_RATING,sp_work_NET_RATING,OREB_PCT,REB_PCT,TM_TOV_PCT,E_TOV_PCT,EFG_PCT,CFID,FG3M,FG3A,FG3_PCT,FTM,FT_PCT,OREB,REB,AST,TOV,BLK,PFD

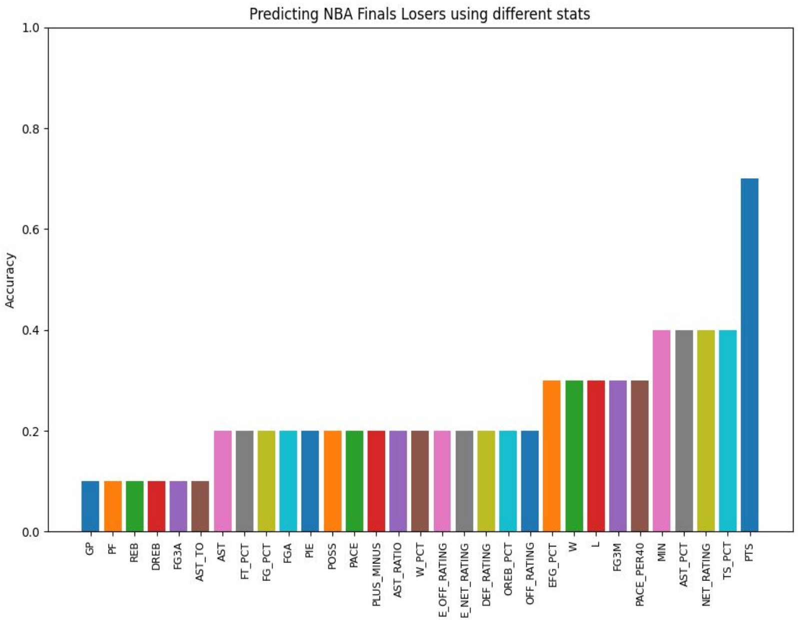
Results - Classifying NBA Champions

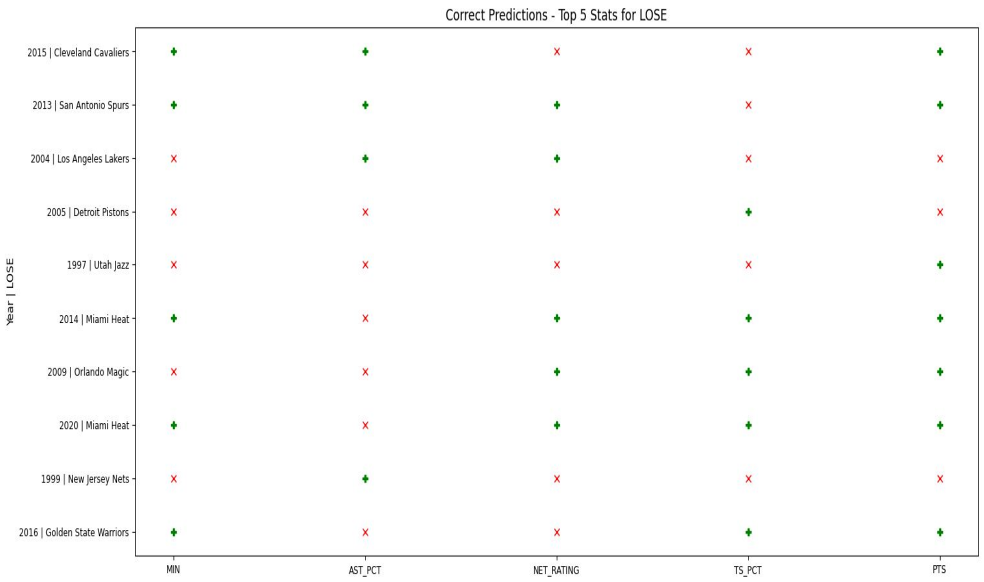




Not Pictured: DREB_PCT,TM_TOV_PCT,FG3M,FG3_PCT,FTM,FTA,OREB,DREB,REB,STL,BLK,PF,PFD,PTS

Results - Classifying NBA Finals Losers





Not Pictured:
E_DEF_RATING,AST_TO,DREB_PCT,REB_PCT,TM_TOV_PCT,E_PACE,FGM,FGA,FG3A,FG3_PCT,FTM,FTA,OREB,TOV,STL,BLK,BLKA,PFD

Conclusion

- A majority of stats by themselves cannot accurately predict a measure of success in the NBA.
- Many that can are basic counting stats, or combinations of said stats (i.e. fantasy points)
- “Advanced” Statistics that do perform well are ranked highly for a reason. Stats such as PIE, Net Rating, etc. have been curated by data scientists for this purpose, but still are not an end-all-be-all for NBA rankings and comparisons.