Sponsorship in NBA

Data Engineering Final Group Project

Yichin Tzou

Agenda

- Introduction
- Data Preprocessing
- Data Analytics
- Data Visualization
- Model Prediction
- Recommendation

01 Introduction

Introduction

- Sports are indispensable in people's lives. NBA is one of worldwide sport industries impacting people lives in some ways like entertainment and related businesses.

- To keep growing NBA sports, sponsors of NBA need action insights which team and players are needed to get sponsorship based on the data collected and analytics to make the best sponsorship.

- Our team provide recommendations to sponsors by introducing our analytic prediction of winning team and home team effect improvement.

Business Application

- Predict which player/ team to sponsor
- Key factors influencing whether the home team will win


02
Data
Preprocessing

Solution Overview

Data	Processing	Production Database	Visualization	Modeling	Insight
NBA Website	**ETL**	**Relational database**	**Trending**	**Prediction**	**Recommendation**
Kaggle has NBA data provided by NBA website	Extraction Transformation Load Completion of data warehouse	Check Normalization Check, Transitiveness Based linked tables, provide sql queries for solution	Provide overview of winning team and players and Home team wins trends	To predict of winning team, executed the logistic regression	For sponsors, Insights and recommendation is made based on the analytics results.

Data Introduction

01	Players.csv	• Players information including their names and teams they belong to
02	Games.csv	• Games information such as home team's PTS, FG3 PCT/ whether home teams win from 2014 to 2021
03	Teams.csv	• Teams information including their names, coach, owner
04	Player_Games_details.csv	• Players and their games information such as FGM, FGA, FG3M earned
05	City.csv	• City information including city name, state, and country
06	Arena.csv	• Arena information including arena name, capacity, built year, and city where it locates

	A	B	C	D	E	F	G	H	I	J	
1	LEAGUE_ID	TEAM_ID	MIN_YEAR	MAX_YEAR	ABBREVIATION	NICKNAME	YEARFOUNDED	CITY	ARENA	ARENACAPACITY	OWN
2	0	1610612737	1949	2019	ATL	Hawks	1949	Atlanta	State Farm Arena	18729	Tony
3	0	1610612738	1946	2019	BOS	Celtics	1946	Boston	TD Garden	18624	Wyc
4	0	1610612740	2002	2019	NOP	Pelicans	2002	New Orleans	Smoothie King Center		Tom
5	0	1610612741	1966	2019	CHI	Bulls	1966	Chicago	United Center	21711	Jerry
6	0	1610612742	1980	2019	DAL	Mavericks	1980	Dallas	American Airlines Center	19200	Mark
7	0	1610612743	1976	2019	DEN	Nuggets	1976	Denver	Pepsi Center	19099	Stan
8	0	1610612745	1967	2019	HOU	Rockets	1967	Houston	Toyota Center	18104	Tilma
9	0	1610612746	1970	2019	LAC	Clippers	1970	Los Angeles	Staples Center	19060	Steve

https://www.kaggle.com/code/samtam22/nba-data-analysis/data

https://www.nba.com/statshttps://www.nba.com/stats

DATA ETL - Process & Details

Check Extracted data

Use Openrefine, Excel and Python to dropped irrelevant & redundant columns.

Data cleaning

Recoding work:
Turn data types & remove strings to analyze

Transformation
Normalize tables

Remove functional dependency &
Transitive dependency

Data Preprocessing - Python + Excel

For example-Teams data:

1. Normalize raw data ‘team’ into 3 tables including teams, arena, and city.
2. Use Excel and python to preprocess data including dealing with missing and abnormal values, creating new columns and adding additional information needed.

7223	1962936483	Jan Vesely	2011	1610612764
7224	1962936489	Brian Skinner	2011	1610612763
7225	1962936495	Damien Wilkins	2011	1610612765
7226	1962937755	Paige Marcus	2017	1610612766
7227	1962937827	Matt Matt	2017	1610612758

7228 rows × 4 columns

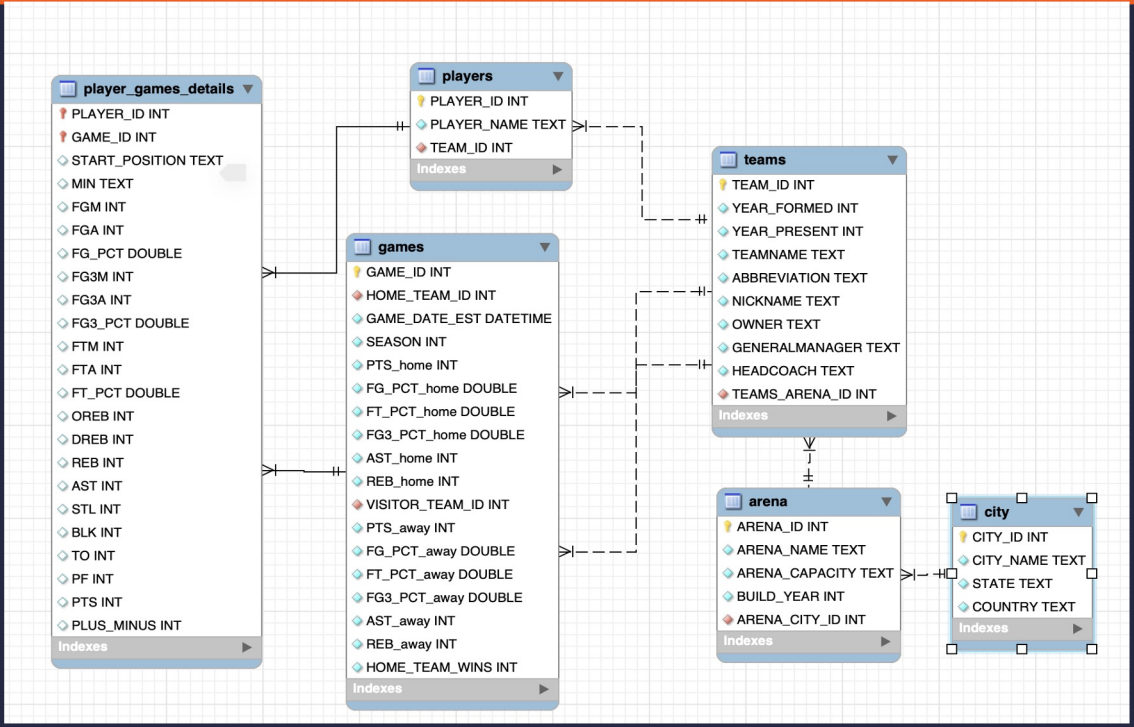
```
In [11]: player = player.drop_duplicates(subset=['PLAYER_ID'])
         player
```

Out[11]:

	PLAYER_ID	PLAYER_NAME	TEAM_ID
0	244	Dee Brown	1610612742
1	255	Grant Hill	1610612746
5	283	Lindsey Hunter	1610612741
6	406	Shaquille O'Neal	1610612738
8	436	Juwan Howard	1610612748
...	...	...	...
7223	1962936483	Jan Vesely	1610612764
7224	1962936489	Brian Skinner	1610612763
7225	1962936495	Damien Wilkins	1610612765
7226	1962937755	Paige Marcus	1610612766
7227	1962937827	Matt Matt	1610612758

1769 rows × 3 columns

Database Schema - ER diagram



1. Normalization
2. Relationships
3. Attributes

03 Data Analytics

Team Overview- Each team's arena and where it locate

Team	Arena	City	Country
Atlanta Hawks	State Farm Arena	Atlanta	United States
Boston Celtics	TD Garden	Boston	United States
Brooklyn Nets	Barclays Center	Brooklyn	United States
Charlotte Hornets	Spectrum Center	Charlotte	United States
Chicago Bulls	United Center	Chicago	United States
Cleveland Cavaliers	Rocket Mortgage FieldHouse	Cleveland	United States
Dallas Mavericks	American Airlines Center	Dallas	United States
Denver Nuggets	Ball Arena	Denver	United States
Detroit Pistons	Little Caesars Arena	Detroit	United States
Golden State Warriors	Chase Center	San Francisco	United States
Houston Rockets	Toyota Center	Houston	United States
Indiana Pacers	Bankers Life Fieldhouse	Indianapolis	United States
Los Angeles Clippers	Crypto.com Arena	Los Angeles	United States
Los Angeles Lakers	Crypto.com Arena	Los Angeles	United States
Memphis Grizzlies	FedExForum	Memphis	United States
Miami Heat	FTX Arena	Miami	United States
Milwaukee Bucks	Fiserv Forum	Milwaukee	United States
Minnesota Timberwolves	Target Center	Minnesota	United States
New Orleans Pelicans	Smoothie King Center	New Orleans	United States
New York Knicks	Madison Square Garden	New York	United States
Oklahoma City Thunder	Paycom Center	Oklahoma City	United States
Orlando Magic	Amway Center	Orlando	United States
Philadelphia 76ers	Wells Fargo Center	Philadelphia	United States
Phoenix Suns	Footprint Center	Phoenix	United States
Portland Trail Blazers	Moda Center	Portland	United States
Sacramento Kings	Golden 1 Center	Sacramento	United States
San Antonio Spurs	AT&T Center	San Antonio	United States
Toronto Raptors	Scotiabank Arena	Toronto	Canada
Utah Jazz	Vivint Arena	Salt Lake City	United States
Washington Wizards	Capital One Arena	Washington	United States

```sql
#each team's arena and where it locate
SELECT teamname as Team, arena_name as Arena, city_name as City, country as Country
FROM teams t
LEFT JOIN arena a ON t.teams_arena_id = a.arena_id
LEFT JOIN city c ON c.city_id = a.arena_city_id
ORDER BY teamname;
```

Team Overview- Number of Games Played in Each City in Season 2020

City	Number of Games
Los Angeles	100
Miami	51
Denver	49
Boston	48
Toronto	46
Milwaukee	45
Houston	45
Dallas	44
Indianapolis	43
Oklahoma City	43
Phoenix	42
Portland	42
Brooklyn	42
Salt Lake City	42
Orlando	41
Philadelphia	41
Washington	39
Memphis	39
Sacramento	38
San Antonio	38
New Orleans	38

```sql
#4) NUMBER OF GAMES PLAYED IN EACH CITY IN SEASON
SELECT
c.city_name as City,
COUNT(g.game_id) as 'Number of Games'
from
city c
left join arena a on c.city_id= a.arena_city_id
left join teams t on a.arena_id=t.teams_arena_id
left join games g on t.team_id=g.home_team_id
where g.season =2020
group by c.city_name
order by COUNT(g.game_id) desc
;
```

Team Overview- Percentage of Team Winning in Their Own Arena

Team	Percentage of Winning Games in their own arena(%)
Miami Heat	46.3415
Los Angeles Lakers	45.1220
Milwaukee Bucks	42.6829
Philadelphia 76ers	40.2439
Denver Nuggets	40.2439
Boston Celtics	39.0244
Toronto Raptors	37.8049
Los Angeles Clippers	37.8049
Houston Rockets	36.5854
Indiana Pacers	34.1463
Oklahoma City Thunder	34.1463
Utah Jazz	31.7073
Dallas Mavericks	26.8293
Portland Trail Blazers	26.8293
Brooklyn Nets	25.6098
San Antonio Spurs	24.3902
Memphis Grizzlies	24.3902
Phoenix Suns	23.1707
Orlando Magic	21.9512
Sacramento Kings	20.7317
New Orleans Pelicans	20.7317
Washington Wizards	19.5122
Chicago Bulls	18.2927
Atlanta Hawks	17.0732
Detroit Pistons	15.8537
Cleveland Cavaliers	13.4146
New York Knicks	13.4146
Golden State Warriors	12.1951
Charlotte Hornets	12.1951
Minnesota Timberwolves	9.7561

```sql
#5) PERCENTAGE OF TEAM WINNING IN THEIR OWN ARENA(HOME TEAM WINS)
select
t.teamname as Team,
round(count(g.home_team_wins)/82*100,4) as 'Percentage of Winning Games in their own arena(%)'
from
teams t
left join games g on t.team_id = g.home_team_id
where g.season=2020 and g.home_team_wins=1
group by t.team_id
order by count(g.home_team_wins) desc
;
```

Team Overview- Team Performance in Season 2020

team_id	teamname	Field Goal Made(%)	Free throw Percentage(%)	3 Points Field Goal Percentage(%)
1610612756	Phoenix Suns	47.8388	37.1847	80.2649
1610612746	Los Angeles Clippers	47.0723	38.6672	79.2477
1610612743	Denver Nuggets	47.0242	36.3625	77.0342
1610612751	Brooklyn Nets	47.0232	36.809	78.5797
1610612758	Sacramento Kings	46.9851	36.2382	75.3115
1610612744	Golden State Warriors	46.9489	37.3239	77.7527
1610612755	Philadelphia 76ers	46.7816	36.3923	74.6508
1610612749	Milwaukee Bucks	46.574	35.6903	74.2636
1610612740	New Orleans Pelicans	46.5348	35.0159	74.8769
1610612762	Utah Jazz	46.3786	36.4426	77.7645
1610612747	Los Angeles Lakers	46.2294	34.6305	74.5984
1610612738	Boston Celtics	46.2031	36.3229	78.7439
1610612764	Washington Wizards	46.1431	34.2	
1610612748	Miami Heat	46.1342	35.9	
1610612759	San Antonio Spurs	45.882	35.7	
1610612737	Atlanta Hawks	45.7757	36.2	
1610612741	Chicago Bulls	45.711	36.2	
1610612754	Indiana Pacers	45.6302	34.9	

```
#Team Performance
SELECT g1.HOME_TEAM_ID 'team_id', t.teamname,
ROUND(((AVG(g1.fg_pct_home) + AVG(g2.fg_pct_away))/2) * 100, 4) 'Field Goal Made(%)',
ROUND(((AVG(g1.fg3_pct_home) +  AVG(g2.fg3_pct_away))/2) * 100,4) 'Free throw Percentage(%)',
ROUND(((AVG(g1.ft_pct_home) + AVG(g2.ft_pct_away)) /2) * 100, 4) '3 Points Field Goal Percentage(%)'
FROM games g1 INNER JOIN games g2 ON g1.HOME_TEAM_ID = g2.VISITOR_TEAM_ID
LEFT JOIN teams t ON t.team_id = g1.home_team_id
where g.season =2020
GROUP BY g1.HOME_TEAM_ID
ORDER BY ROUND(((AVG(g1.fg_pct_home) + AVG(g2.fg_pct_away))/2) * 100, 4) DESC,
ROUND(((AVG(g1.fg3_pct_home) +  AVG(g2.fg3_pct_away))/2) * 100,4) DESC,
ROUND(((AVG(g1.ft_pct_home) + AVG(g2.ft_pct_away)) /2) * 100, 4) DESC
;
```

Player Overview- Points gain by per player in season 2020

ID	Name	Points Gained
203507	Giannis Antetokounmpo	387
203114	Khris Middleton	347
1626164	Devin Booker	322
201950	Jrue Holiday	266
202331	Paul George	237
101108	Chris Paul	227
1629027	Trae Young	206
1629028	Deandre Ayton	195
201572	Brook Lopez	182
202704	Reggie Jackson	171

```sql
#points gain by per player
select
p.player_id as ID,
p.player_name as Name,
sum(pgd.pts) as 'Points Gained'
from
players p
left join teams t on t.team_id=p.team_id
left join player_games_details pgd on p.player_id=pgd.player_id
left join games g on pgd.game_id=g.game_id
where g.season =2020
group by p.player_id
order by SUM(pgd.pts) desc
limit 10;
```

Player Overview- Top 10 Player who has best performance in field goal made in season 2020

ID	Name	Team	Field Goal Percentage
202324	Derrick Favors	New Orleans Pelicans	100
203497	Rudy Gobert	Utah Jazz	83.3
1626220	Royce O'Neale	Utah Jazz	66.7
1628993	Alize Johnson	Indiana Pacers	66.7
1628971	Bruce Brown	Detroit Pistons	63.9
203552	Seth Curry	Dallas Mavericks	61.8333
203507	Giannis Antetokoun…	Milwaukee Bucks	61.4917
1629028	Deandre Ayton	Phoenix Suns	60.9083
1629661	Cameron Johnson	Phoenix Suns	59.2545
203991	Clint Capela	Houston Rockets	58.0556

```sql
#top10 players who has best performance iin field goal in season 2019
select
p.player_id as ID,
p.player_name as Name,
t.teamname as Team,
ROUND(avg(pgd.fg_pct)*100,4) as 'Field Goal Percentage'
from
players p
left join teams t on t.team_id=p.team_id
left join player_games_details pgd on p.player_id=pgd.player_id
left join games g on pgd.game_id=g.game_id
where g.season =2020
group by p.player_id
order by  avg(pgd.fg_pct)*100 desc
limit 10;
```

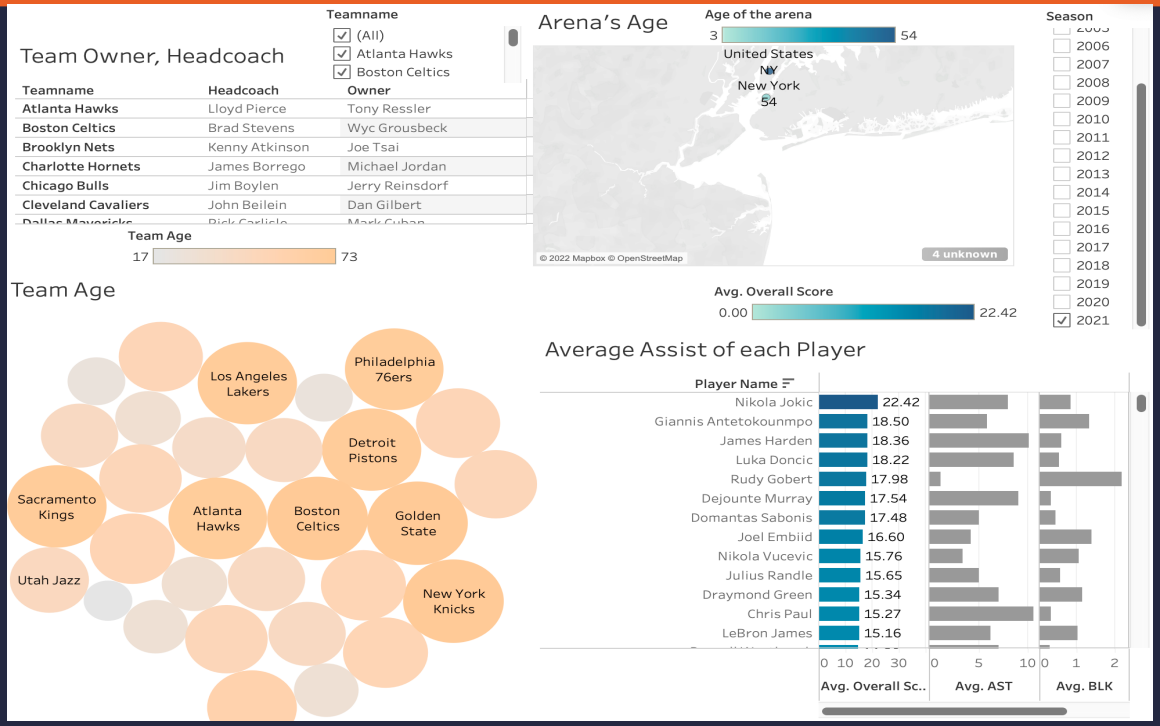
Player Overview- Top 10 Players who have Overall Best Performance in Season 2020

ID	Name	Overall Score	Field Goal Percentage(%)	3 Points Field Goal Percentage(%)	Free throw Percentage(%)
1626220	Royce O'Neale	77.8	66.7	66.7	100
203552	Seth Curry	66.1667	61.8333	61.6667	75
203903	Jordan Clarkson	61.1	50	33.3	100
1628378	Donovan Mitchell	57.0333	44.4	60	66.7
202704	Reggie Jackson	54.9458	48.9625	37.75	78.125
203954	Joel Embiid	54.9444	49.9667	38.3333	76.5333
201568	Danilo Gallinari	54.8296	47.6889	47.3556	69.4444
1629629	Cam Reddish	54.1333	53.475	46.425	62.5
1629661	Cameron Johnson	53.3364	59.2545	55.3	45.4545
201935	James Harden	51.95	42.5	33.35	80

```
select
p.player_id as ID,
p.player_name as Name,
ROUND((avg(pgd.fg_pct)+avg(pgd.fg3_pct)+avg(pgd.ft_pct))/3*100,4) as 'Overall Score',
ROUND(avg(pgd.fg_pct)*100,4) as 'Field Goal Percentage',
ROUND(avg(pgd.fg3_pct)*100,4) as '3 Points Field Goal Percentage',
ROUND(avg(pgd.ft_pct)*100,4) as 'Free throw Percentage'
from
players p
left join teams t on t.team_id=p.team_id
left join player_games_details pgd on p.player_id=pgd.player_id
left join games g on pgd.game_id=g.game_id
where g.season =2020
group by p.player_id
order by  ROUND((avg(pgd.fg_pct)+avg(pgd.fg3_pct)+avg(pgd.ft_pct))/3*100,4) desc
limit 10;
```


04
Data
Visualization

Visualization-Overview



05

Model Prediction

Who will be the champion of season 2020?

Model 1 - Feature Selection

- Field Goals
- 3 points Field Goals
- Assist
- Rebounds

```
selected_features = [
    'FG_PCT_home', 'FT_PCT_home', 'FG3_PCT_home', 'AST_home', 'REB_home',
    'FG_PCT_away', 'FT_PCT_away', 'FG3_PCT_away', 'AST_away', 'REB_away',
    ]

# check the features we selected
X = df[selected_features]
X.head()
```

	FG_PCT_home	FT_PCT_home	FG3_PCT_home	AST_home	REB_home	FG_PCT_away	FT_PCT_away	FG3_PCT_away	AST_away	REB_away
0	0.409	0.929	0.308	32.0	56.0	0.372	0.737	0.375	22.0	31.0
1	0.446	0.611	0.400	30.0	58.0	0.403	0.818	0.381	20.0	36.0
2	0.470	0.800	0.333	25.0	38.0	0.488	0.724	0.385	20.0	44.0
3	0.389	0.947	0.238	26.0	54.0	0.395	0.895	0.364	20.0	34.0
4	0.466	0.792	0.500	29.0	42.0	0.430	0.750	0.450	15.0	37.0

```
# check the targets
y = df['HOME_TEAM_WINS']
y.head()
```

```
0    1
1    1
2    0
3    1
4    1
Name: HOME_TEAM_WINS, dtype: int64
```

Model 1 - Feature Selection

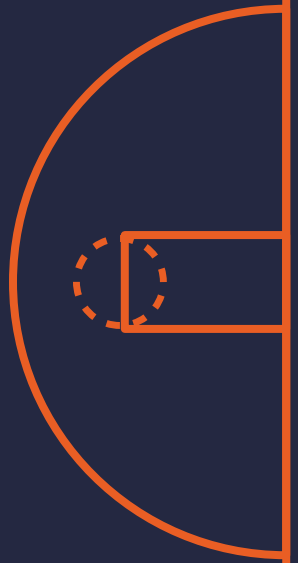
Classification problem

→ Home team win V.S loss

Logistic Regression

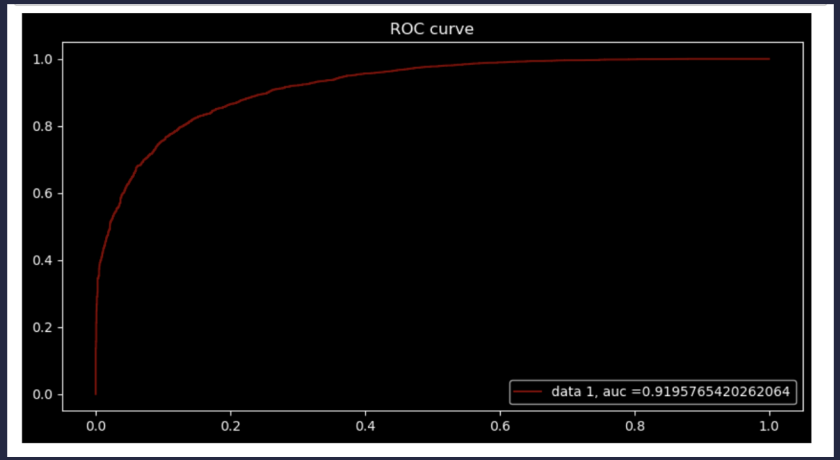
→ % home team win

- Model Accuracy:84%



Model Evaluation

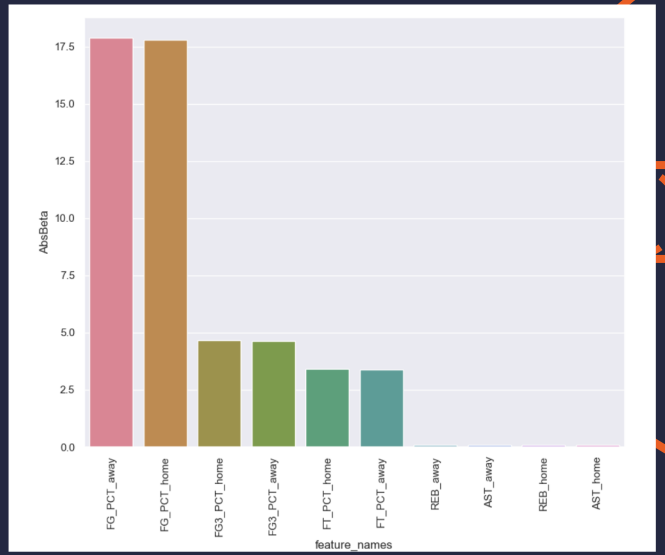
ROC curve → check the performance of a classification model.



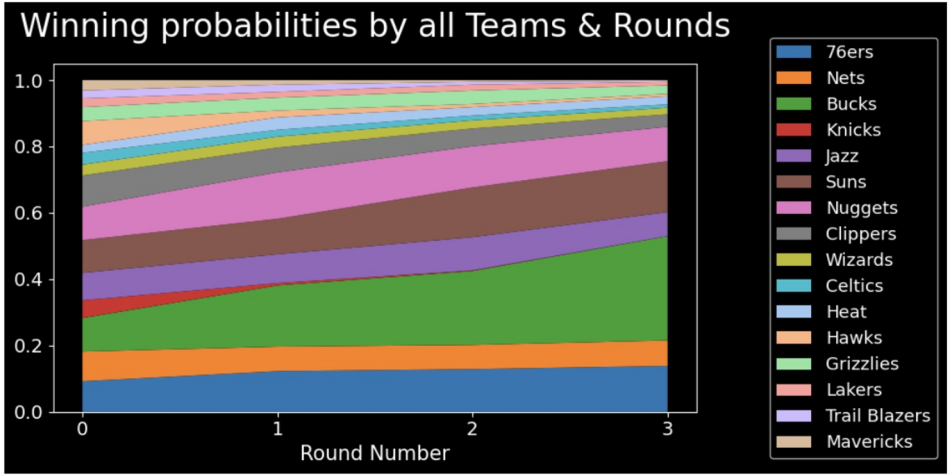
AUC = 0.9196 → model accurate

Feature Engineering

	feature_names	Beta	AbsBeta
5	FG_PCT_away	-17.885183	17.885183
0	FG_PCT_home	17.809989	17.809989
2	FG3_PCT_home	4.669203	4.669203
7	FG3_PCT_away	-4.611214	4.611214
1	FT_PCT_home	3.416628	3.416628
6	FT_PCT_away	-3.383918	3.383918
9	REB_away	-0.095112	0.095112
8	AST_away	-0.086580	0.086580
4	REB_home	0.085946	0.085946
3	AST_home	0.083341	0.083341



Model2: Winning % of all team(2020)



1. Based on overall winning probabilities, the top 3 teams are: **Bucks > Suns > Nuggets**.
2. Based on winning probabilities in the final round, the top 3 teams are: **Bucks > Suns > Nuggets = 76ers**
3. **Bucks** has a much higher overall chance of winning, if it can get through the early rounds.

06 Recommendation

Whom to invest? (player/ team)
How to maximize team performance?

Recommendation

- PLAYER -

Giannis Antetokounmpo
Derrick Favors
Royce O'Naele

- TEAM -

Bucks
Nuggets
Sun

- Sponsored Team Focus -

1. Field Goals
2. 3 Points Field Goals

- Additional insights -

-Sponsor ancient arena, Madison Squrare Garden for NY Kincks
-Have more ads in LA that held most to increase publicity

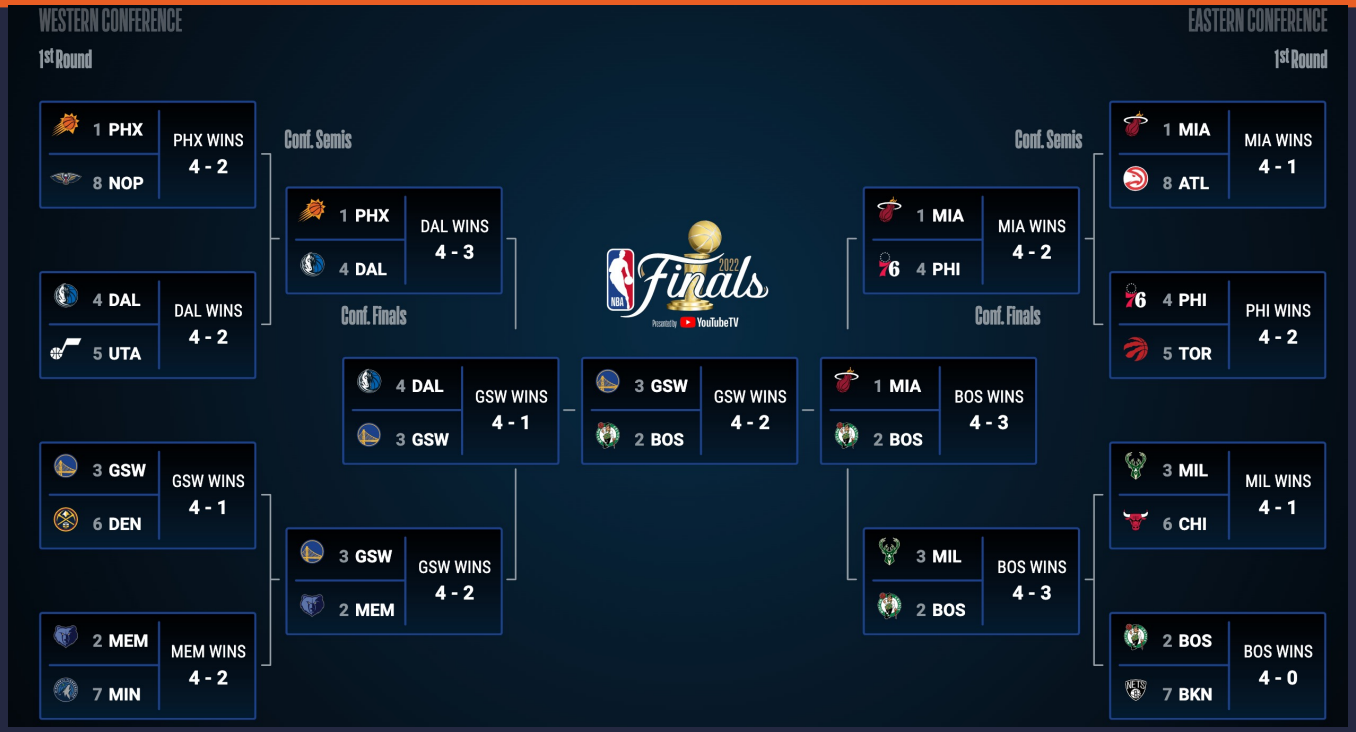
THANKS

DO YOU HAVE ANY QUESTIONS?

Appendix

Additional information

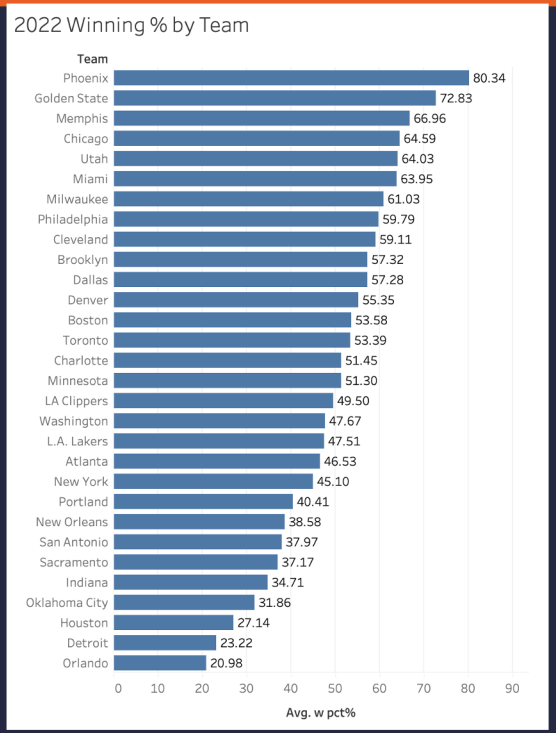
NBA PLAYOFFS

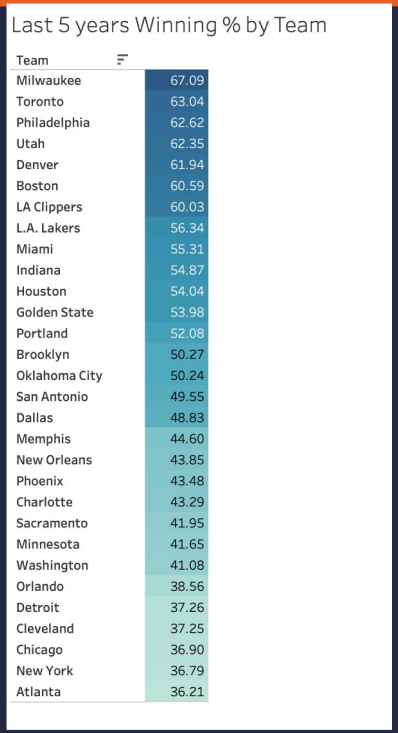


Visualization-Overview

Utah is the great team that ranked in the top 5 both in season 2022 and the last 5 years.

Phoenix and Golden State recently have great performance unlike what to expect from last 5 years ranking of winning percentage.





Visualization-Overview