Final Report

MET CS 777

Ting-Yen Hwang

October 15, 2024

Research Background

Basketball has always been my favorite sport and the National Basketball Association (NBA) has become one of the most successful sports leagues around the world. While the game is evolving itself, technology also plays a big role in this situation. Basketball has become a data-driven sport, General Managers and Head Coaches uses different data to predict or evaluate a player's performance and with no surprise Machine Learning is the key to it. While the new season is upon us and knowing the strong connection between my favorite sport basketball and data, it would be really interesting for me to have a project about it.

Objective

There are two objective in this project:
(1) Predict player's stats for the new 2024-2025 season.
(2) Create a Model to predict voting results for NBA Most Valuable Player Award.

Data Sources

(1) Basketball Reference
(2) Kaggle

Data Introduction

There's too many standard or advanced data for NBA players, for this project, I would be focusing on following stats:

(1) Age: Player's age for the season
(2) G: Games Played
(3) MP:Average Minutes Played per game
(4) 3pt%: 3 Pointer Shooting Percentage
(5) TRB: Total Rebounds
(6) FG%: Field Goal Percentage
(7) AST: Assists per Game
(8) STL: Steals per Game

(9) BLK: Blocks per Game
(10)TOV: Turnover per Game
(11) PTS: Points per Game
(12) WS: Total Win Share in a Season
(13) WS/48: Win Share per 48 mins

We will also be focusing on the VARIATION for each category year by year in last 3 seasons.

Methodology

Predict player's stats for the new 2024-2025 season
(1) Preprocessing player's data.
(2) Create training data featuring player's last 3 season performance and the trend in each season.
(3) Select a model and tuning.
(4) Predict a player's performance for the new 2024-2025 season by using the model.

Create a model to predict MVP Awards
(1) Preprocessing previous MVP Awards voting results data
(2) Select a model and tuning.
(3) Based on the predicted stats for the new season we created, predict the new season's MVP award winner.

Data Preprocessing

(1) Delete Duplicate Data
(2) Clean data that is too old (1980~)

```
from pyspark.sql import functions as F

# Our data would start form 1979-1980 season.
start_season = '1946-1947'
end_season = '1978-1979'

# Filter the DataFrame to remove rows between the defined seasons
players_df = players_df.filter(~((F.col('Season') >= start_season) & (F.col('Season') <= end_season)))
adv_players_df = adv_players_df.filter(~((F.col('Season') >= start_season) & (F.col('Season') <= end_season)))
```

(3) Create Variation for Training Data

```
from pyspark.sql import functions as F


for col in ['G', 'MP', '3P%', 'TRB', 'FG%', 'FT%', 'AST', 'STL', 'BLK', 'TOV', 'PTS', 'WS', 'WS/48']:
    veteran_data = veteran_data.withColumn(f"{col}_Percent_Change_1",
                                           F.when(F.col(f"Prev_{col}_2") != 0,
                                                  ((F.col(f"Prev_{col}_1") - F.col(f"Prev_{col}_2")) / F.col(f"Prev_{col}_2")) * 100)
                                           .otherwise(0))


for col in ['G', 'MP', '3P%', 'TRB', 'FG%', 'FT%', 'AST', 'STL', 'BLK', 'TOV', 'PTS', 'WS', 'WS/48']:
    veteran_data = veteran_data.withColumn(f"{col}_Percent_Change_2",
                                           F.when(F.col(f"Prev_{col}_3") != 0,
                                                  ((F.col(f"Prev_{col}_2") - F.col(f"Prev_{col}_3")) / F.col(f"Prev_{col}_3")) * 100)
                                           .otherwise(0))
```

Model Performance

(1) Random Forest

	G	MP	3p%	TRB	FG%	FT%	AST	STL	BLK	TOV	PTS	WS	WS/
RSME	14.9	5.06	0.11	1.27	0.03	0.07	0.95	0.27	0.27	0.44	3.16	0.55	0.03
R2	0.11	0.60	0.50	0.76	0.53	0.50	0.78	0.64	0.75	0.7	0.73	0.55	0.48

(2) Linear Regression

	G	MP	3p%	TRB	FG%	FT%	AST	STL	BLK	TOV	PTS	WS	WS/
RSME	15.1	5.11	0.11	1.18	0.04	0.071	0.86	0.25	0.24	0.42	3.04	2.1	0.04
R2	0.08	0.6	0.45	0.79	0.56	0.53	0.82	0.68	0.8	0.71	0.75	0.56	0.48

Prediction Results

	MP	3p%	TRB	FG%	FT%	AST	STL	BLK	TOV	PTS	WS
Nikola Jokic	34.2	32%	10.1	54.8%	81.2%	6.8	1.2	0.8	3.02	25.6	11.18
G. Antetokoumpo	34.9	26.4%	10.3	56.4%	68.2%	5.7	1.2	0.9	3.16	25.17	10.20
Luka Doncic	36.3	32.8%	8.2	48.4%	79%	7.6	1.5	0.6	3.44	25.14	10.04
Joel Embiid	33.9	33.2%	10	51%	82.6%	5.8	1.1	1.4	3.22	24.8	8.31

	MP	3p%	TRB	FG%	FT%	AST	STL	BLK	TOV	PTS	WS
S. G-Alexander	34.9	33.7%	5.6	50.7%	83.2%	6.5	1.6	0.7	2.5	24.3	10.86
Kevin Durant	35.1	33.9%	6.9	50.3%	82.8%	5.2	0.9	1.1	2.87	24.1	7.24
Donovan Mitchell	35.2	35.2%	5.1	45.9%	83.5%	5.9	1.6	0.4	2.83	23.8	7.37
Anthony Davis	35.2	27.5%	10.5	54.4%	81.7%	3.5	1	1.8	2.26	23.2	8.78
Jayson Tatum	35.1	33.8%	7.6	46.3%	82%	4.5	0.9	0.5	2.46	23.2	7.96

Some of the results didn't go well and it's hard for Machine to predict.
Games Played (G) is hard to do that is because there's unpredictable factors such as injuries, Trades, Suspension, NBA lockdown, COVID etc..

Win Share (WS) and Win Share per 48 min (WS/48) is not easy because it is related to how your team performs, if your affiliated team has a outstanding season with higher winning percentage, a player's stats would be higher. Since we didn't include a team's standings, it is much more difficult to predict.

I've also dig deep in the data, player's that has less playing time and limited chance to perform is quite difficult to predict as well. On the contrary, All-Star players and MVP-caliber players have much more consistent statistics.

Predict MVP Awards

What is Share?
Share is a stats for us to count the voting results. In the MVP voting, each voters has 5 votes, No.1 ~No.5. If you vote for a player No.1, the player would receive 5 points, 4 points for No.2 and so on.

N = amount of voters

Share = Total Score / 5N (Maximum points a player can get)

Train-Test Split
From 2000~2023, I made '04, '08, '12, '16 and '20 as test data, other years in train data.
Results
Rank represents the actual rank of win share a player had in the voting results.
The list is displayed sequentially based on the model prediction.

Rank	Player
4	LeBron James
2	Chris Paul
1	Kobe Bryant
6	Amar'e Stoudemire
5	Dwight Howard
9	Steve Nash
3	Kevin Garnett
11	Dirk Nowitzki
7	Tim Duncan
14T	Paul Pierce
10	Manu Ginóbili
12	Deron Williams
14T	Carlos Boozer
14T	Rasheed Wallace
14T	Antawn Jamison
13	Carmelo Anthony
8	Tracy McGrady

Rank	Player
1	Stephen Curry
5	Kevin Durant
9	James Harden
2	Kawhi Leonard
4	Russell Westbrook
3	LeBron James
6	Chris Paul
7	Draymond Green
10	Kyle Lowry
8	Damian Lillard

Rank	Player
1	Kevin Garnett
2	Tim Duncan
6	Shaquille O'Neal
4	Peja Stojaković
10T	Sam Cassell
8	Jason Kidd
14T	Yao Ming
10T	Baron Davis
7	Ben Wallace
13	Andrei Kirilenko
5	Kobe Bryant
10T	Dirk Nowitzki
3	Jermaine O'Neal
14T	Michael Redd
9	LeBron James
	Carmelo Anthony

Rank	Player
1	Giannis Antetokou...
3	James Harden
8	Damian Lillard
6	Anthony Davis
4	Luka Dončić
5	Kawhi Leonard
2	LeBron James
9	Nikola Jokić
11	Jimmy Butler
7	Chris Paul
12	Jayson Tatum
10	Pascal Siakam

Rank	Player
1	LeBron James
2	Kevin Durant
3	Chris Paul
7	Dwight Howard
6	Kevin Love
8	Rajon Rondo
9	Steve Nash
11	Derrick Rose
4	Kobe Bryant
10	Dwyane Wade
12T	Dirk Nowitzki
15	Joe Johnson
5	Tony Parker
12T	Russell Westbrook
14	Tim Duncan

The results were pretty good. 4 of the 5 years I made as test data the model correctly predicted the MVP winner. The only year that the model failed is 2008, model predicted Cleveland Cavaliers Forward LeBron James as the MVP instead of the actual winner Los Angeles Lakers Kobe Bryant (R.I.P.!).
As I looked back on that season, I'd believe that the Team Standings played a part on that. The winner Kobe Bryant, who played for the Lakers, had a 57W25L .695% season and was the first seed of the western conference. However, Cleveland Cavaliers which LeBron played of only had a 45W37L season.

To Combine these two models I made, I fed the MVP model the stats produced from my previous model on the 2024-2025 season.

My Prediction on NBA 2024-2025 season MVP voting results:

Player Reference	Season	Player	Age
jokicni01	2024.0-2025.0	Nikola Jokić	29.0
antetgi01	2024.0-2025.0	Giannis Antetokou...	30.0
gilgesh01	2024.0-2025.0	Shai Gilgeous-Ale...	26.0
embiijo01	2024.0-2025.0	Joel Embiid	30.0
doncilu01	2024.0-2025.0	Luka Dončić	25.0
davisan02	2024.0-2025.0	Anthony Davis	31.0
allenja01	2024.0-2025.0	Jarrett Allen	26.0
sabondo01	2024.0-2025.0	Domantas Sabonis	28.0
hardeja01	2024.0-2025.0	James Harden	35.0
portemi01	2024.0-2025.0	Michael Porter Jr.	26.0
jacksja02	2024.0-2025.0	Jaren Jackson Jr.	25.0
tatumja01	2024.0-2025.0	Jayson Tatum	26.0
harrito02	2024.0-2025.0	Tobias Harris	32.0
markkla01	2024.0-2025.0	Lauri Markkanen	27.0
murraja01	2024.0-2025.0	Jamal Murray	27.0
quickim01	2024.0-2025.0	Immanuel Quickley	25.0
whitede01	2024.0-2025.0	Derrick White	30.0
irvinky01	2024.0-2025.0	Kyrie Irving	32.0
sextoco01	2024.0-2025.0	Collin Sexton	26.0
halibty01	2024.0-2025.0	Tyrese Haliburton	24.0

(1) Nikola Jokic (Denver Nuggets)
(2) Giannis Antetokounmpo (Milwaukee Bucks)
(3) Shai Gilgeous-Alexander (Oklahoma City Thunder)
(4) Joel Embiid (Philadelphia 76ers)
(5) Luka Doncic (Dallas Mavericks)

Obviously the season hasn't started yet so there's no way we could know the results, but I did look up the Vegas Odds on who will be the NBA MVP next season. The results:

(1) Luka Doncic (Dallas Mavericks)
(2) Nikola Jokic (Denver Nuggets)
(3) Shai Gilgeous-Alexander (Oklahoma City Thunder)
(4) Giannis Antetokounmpo (Milwaukee Bucks)
(5) Joel Embiid (Philadelphia 76ers)

All five of the top favorites were correctly predicted!

Conclusion

I believe that predicting a player's performance is a difficult and complex task. Not only does it require a solid understanding of machine learning models, but also a deeper comprehension of what the player's data represents. In addition, there are many external, uncontrollable factors that can impact accuracy. However, it is precisely this uncertainty that makes this project particularly interesting.

I think there's still considerable room for improvement in the overall prediction results. More time needs to be spent fine-tuning parameters, selecting features, and choosing the right prediction models. Additionally, the model should be trained differently for various types of players. The model I've trained performs better in predicting the performance of key players in the league, but struggles more with role players whose playing time is inconsistent, which is an area that needs significant improvement.

Regarding the model for predicting MVP, I believe the results are quite good. However, since MVP voters typically take both team success and individual performance into account, I believe that incorporating team records as an additional factor could make the model even better.