

NBA MVP Prediction



Crystal Xu, Diandian Yuan, Jack Chen, Jiaqi Feng, Zoey Zhou

Agenda

01

Problem Statement

02

Data Profile

03

Data Preprocessing

04

Model Selection

05

Conclusions



01

Problem Statement

- Business problem
- Project purpose

Problem Statement

Business Problem:

- The NBA Most Valuable Player (MVP) award has been awarded since the 1955-56 NBA season, with one player with the most perfect behaviors each regular season. For a NBA player, becoming the MVP is not just an affirmation of his ability, his commercial value and popularity will also be greatly increased. But for those companies that want to sign a contract with an MVP and reach a partnership, if they can discover the commercial value of the player before he becomes an MVP, they will save a lot of time, money and energy, and can get more interests as well.

Project Purpose:

- **Our purpose is to predict the next NBA Most Valuable Player** using Machine Learning and Statistical Data from NBA players between 1980 and 2021. And the result could be valuable for those who would like to acquire and maintain partnerships with the most promising NBA stars ahead of time so that they can seek greater benefit and development for themselves.

02

Data Profile

- Data sources and descriptions
- Definition of features & variable
- Clustering & EDA
- Interesting findings



Data Sources

All data are scrapped from Basketball-Reference website



Data	Description	Columns/Rows	Source
Classic Stats per Game	Includes each player's average statistics per game for each season. Has columns like average minutes played, field goal stats, points per game..	30 columns X All players X 1980-2022 Seasons	http://www.basketball-reference.com/leagues/NBA_1980_per_game.html
Advanced Stats per Game	Includes some calculated stats for each player per game, like player efficiency rating, true shooting percentage, 3-point attempt rate	27 columns X All players X 1980-2022 Seasons	https://www.basketball-reference.com/leagues/NBA_1980_advanced.html
MVP Trophy Historic Winner	Includes the winner of MVP for each season and their stats per game and shooting. Because of the change of MVP voting rules in 1980, we will use the data after 1980	18 columns X 1955-2021 Seasons	https://www.basketball-reference.com/awards/mvp.html
MVP Trophy Voting Results	Includes the voting results like points won and final shares won	20 columns X All Candidates X 1980-2021 Seasons	https://www.basketball-reference.com/awards/awards_1980.html
Team Standings	Team stats for each season, including the rank of teams, wins and losses stats and whether the team got in the playoffs	7 columns X each Team X 1980-2021 Seasons	https://www.basketball-reference.com/leagues/NBA_1980_standings.html

Feature Definition & Variables

Prediction columns:

- **MVP**: True for MVP winner each season
- **MVP_share**: A percentage of points won in MVP voting for each MVP candidate each season
- **ShareYN**: True if the player is an MVP candidate, eligible for MVP voting

MVP Share Calculation:

- $$MVP\ Share = (MVP\ points\ for\ particular\ player) / (Total\ MVP\ points)$$

2020-2021 NBA Awards Voting (Top 3 Candidates)

Player	1st Place Votes (10 points)	2nd Place Votes (7 points)	3rd Place Votes (5 points)	4th Place Votes (3 points)	5th Place Votes (1 point)	Total Points	MVP Share
Nikola Jokic	91	8	1	0	0	971	$971 / 1010 = 0.961$
Joel Embiid	1	62	23	8	3	586	$586 / 1010 = 0.580$
Stephen Curry	5	23	32	23	13	453	$453 / 1010 = 0.449$

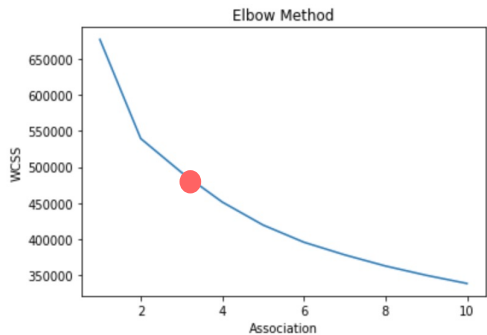
Feature Summary:

- **67 Features in total before feature engineering:**
 - **General**: Season, Decade
 - **Player Information**: Name, Position, Age, Team, Trade (true if traded during the season), Past MVP (true if win MVP in the past)
 - **Stats per Game**: Minutes Played, Field Goals, 2P, 3P, Free throws, Offensive Rebounds, Defensive Rebounds, Assists, Steals, Blocks, Turnovers, Points
 - **Aggregated /Calculated Measures Across the season**: Total minutes played, Player Efficiency Rating, True Shooting percentage, Offensive win shares, Value over replacement player
 - **Team Performance Stats**: Wins and losses, Points per Game, Conference Ranking, Leagues Ranking , Playoffs (true if gets into playoff)

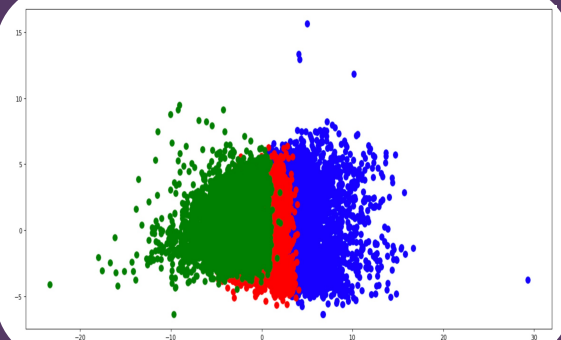
* Stats Glossary: <https://www.nba.com/stats/help/glossary/#fta>

Clustering

Elbow Method



Cluster Visualization



- K-means Clustering
- We used the elbow method for determining the number of clusters in our data set;
- For each value of K, we are calculating Within-Cluster Sum of Square;
- The Elbow Point was located at 3, so we would have 3 clusters;
- Use PCA to reduce dimensions

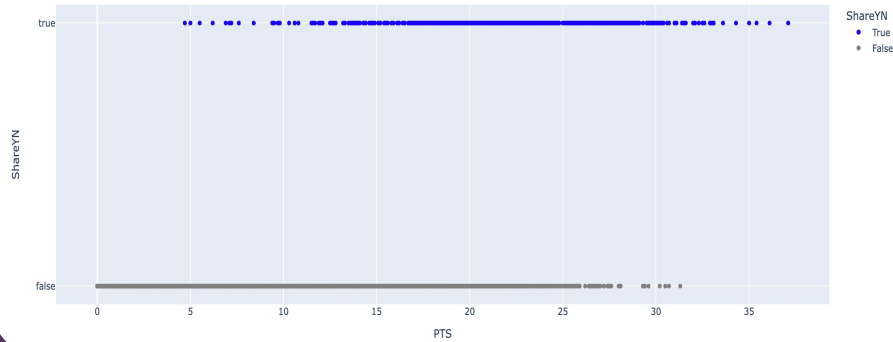
Cluster Means

	ORB	DRB	AST	STL	BLK	3P	OWS	DWS	PTS	MVP	counts
Cluster											
0	1.672283	4.280565	3.399674	1.101804	0.728717	0.679109	3.850261	2.770957	15.263391	0.00913	4600
1	1.047965	2.554253	1.900262	0.696472	0.410533	0.501362	0.848803	0.988699	8.325939	0.00000	5725
2	0.585473	1.303825	0.812817	0.338286	0.219217	0.204689	0.148363	0.551710	3.469842	0.00000	5438

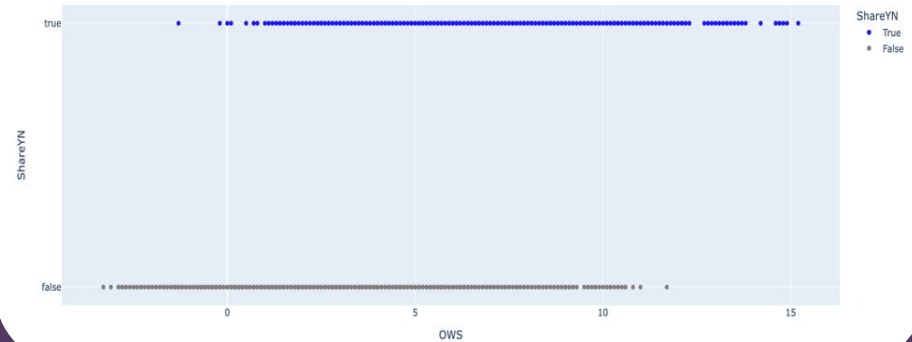
- Comparing the cluster means, we can find generally Cluster 0 players perform better than Cluster 1 than Cluster 2;
- The differences of cluster means show that our features can distinguish players well by their performance and value
- All MVPs are in cluster 0.

Data Exploration and Analysis I

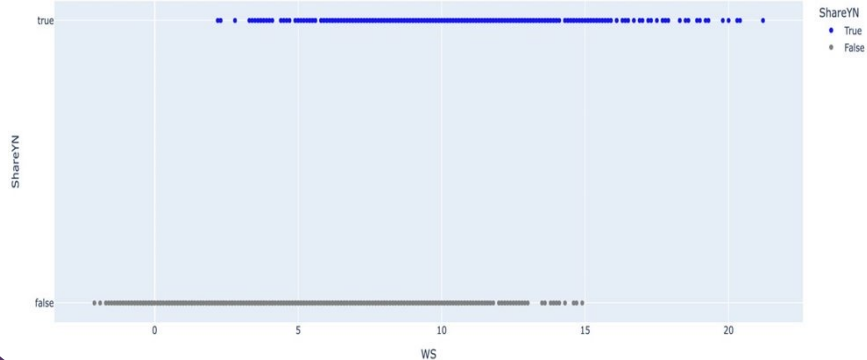
PTS vs. ShareYN



OVS vs. ShareYN



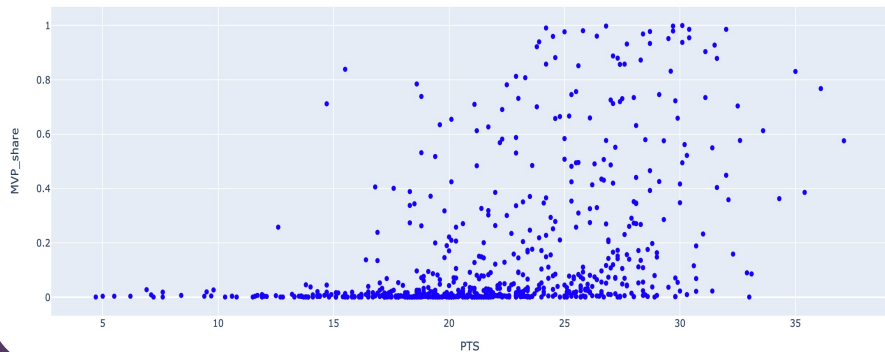
WS vs. ShareYN



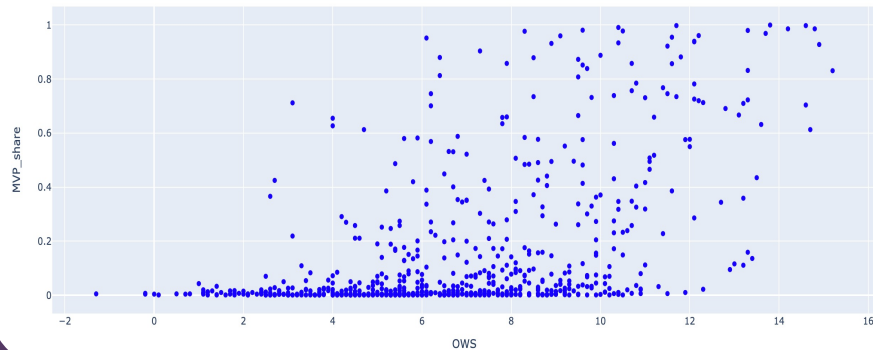
- Based on the correlation matrix and mutual information scores plot, we made scatter plots for the features which had high positive correlations with MVP share, the points represent each players from 1980–2021 with the actual MVP candidates highlighted in blue;
- The y-axis means whether a player could become MVP candidate or not: 'True' means the player was the MVP candidate and 'False' means not;
- The x-axis means the score one player got for this feature;
- Players who was chosen to be candidate generally scored higher than those who didn't.

Data Exploration and Analysis II

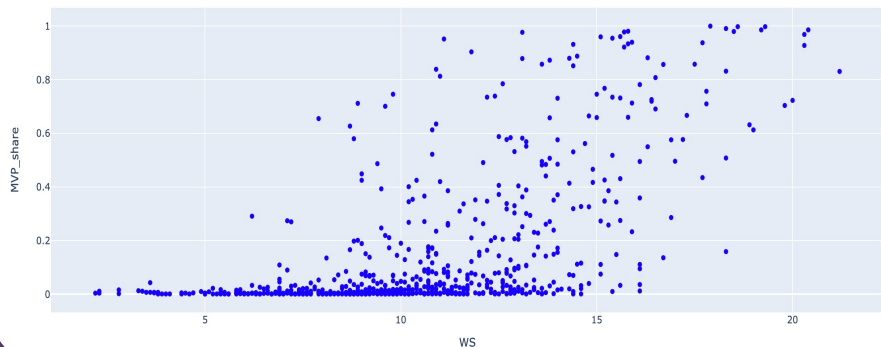
PTS vs. MVP Share



OWS vs. MVP Share

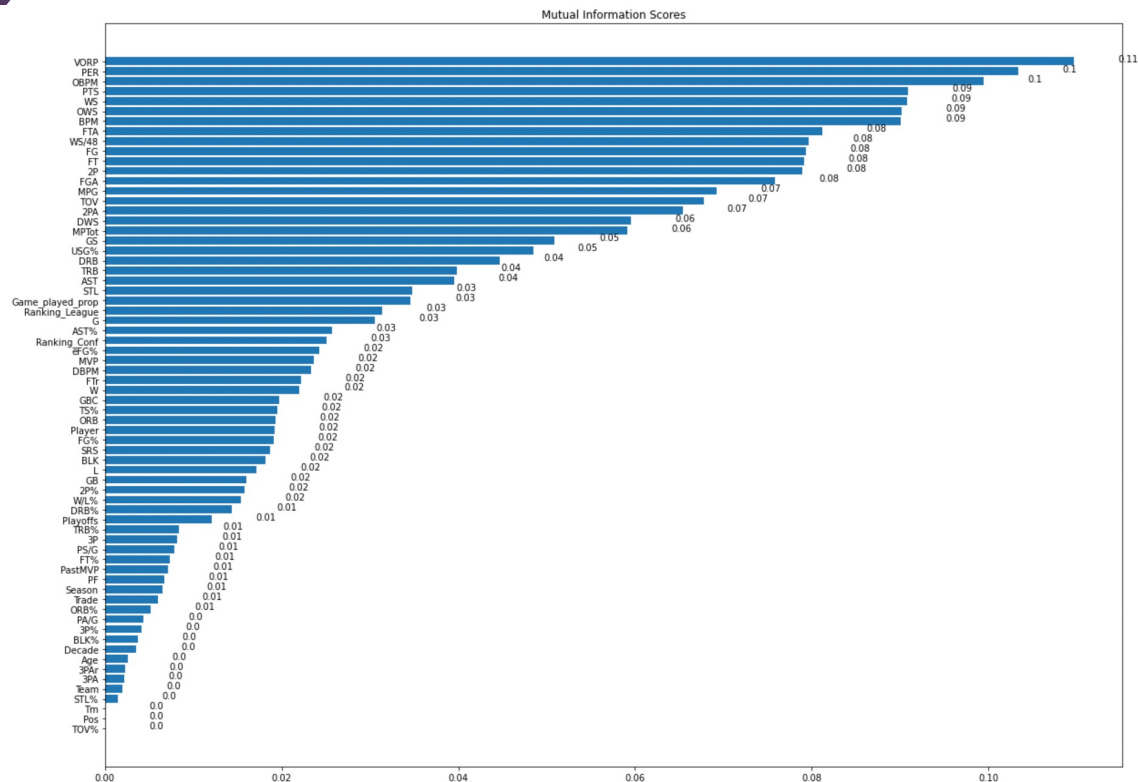


WS vs. MVP Share



- In these scatter plots, all blue points represent for the MVP candidates from 1980-2021 ;
- The y-axis means the MVP share one candidates got, from 0-1;
- The x-axis means the score one candidates got for this feature;
- If a candidate scored higher than others, his MVP share would be higher accordingly

Feature Importance Exploration



Using Mutual information method to measure association between features and MVP_share

As we can see from the plot, features like VORP, WS, PER, PTS, FG, 2P are much more related.

Age, Pos, Team, Trade, ORB are less related.

Findings

- 3-Point is not important in MVP election: From the Feature information chart, we can see 3P related data has about 0 information score, compared to 0.1 for 2P. The mean value for 3-Point Field Goal Percentage has almost no difference between MVPs and normal players. (0.284 vs 0.216)
- Only teams that perform extraordinarily can have an MVP. The mean win-loss ratio for the MVP team is 74.5%, compared to only 50% for other teams. The min win-loss ratio is 71% for the MVP teams, which means only teams winning more than 70% of games can have an MVP.
- Most of the time players younger than 26 cannot win MVP. For Age, the mean value of MVP (27.8) is older than the normal players (22.6), the first quartile of MVP is 26, and the first quartile of normal players is just 23.
- MVPs have a higher free throw rate. The mean value of free-throw Percentage for MVP is 0.8 and for normal players is 0.7. Also from the feature information chart, free-throw has a 0.08 information score.

03

Data Processing

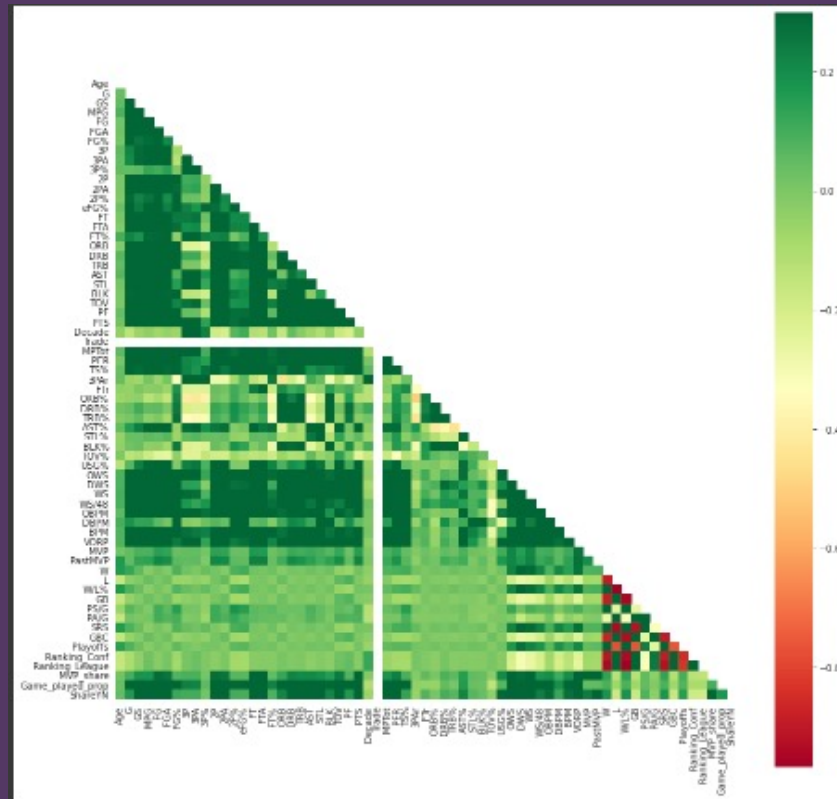
- > Features Selection
- > Train Test Split
- > SMOTE
- > PCA

Features Selection

Reduced to 49 Features

```
[ ] unusefull_columns = [
    'Team',           #same with TM
    'GS',             #Game started, unuseful info, only partilly filled
    'FG',             #Sum of 2P and 3P, depend on 2P and 3P
    'FGA',            #Sum of 2PA and 3PA
    'FG%',            #Depending of 2P% and 3P%
    '2PA',            #Depending of 2P and 2P%
    '3PA',            #Depending of 3P and 3P%
    'FTA',            #Depending of FT and FT%
    'TRB',            #Sum of ORB and DRB
    'Decade',         #Unuseful
    'W',              #covered by W/L ratio
    'L',              #covered by W/L ratio
    'MPTot',          #covered by MP per Game, and MP total depends on nu
    'GB',             #Highly correlated
    'Trade',          #The trade column are all false, which means only p
    'Player',         #The name of player is not useful
    'Tm',             #The name of team is not useful
    'Season'          #Unuseful
]
```

- ☐ Drop Redundant Features
- ☐ Drop Highly Correlated Features
- ☐ Missing Value Imputation
- ☐ Standard Scaler
- ☐ Label Encoder



Train-Test Split

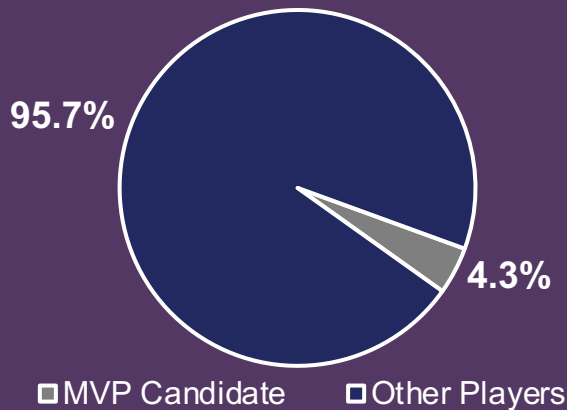
SMOTE

PCA

To Avoid Seasons cut offs

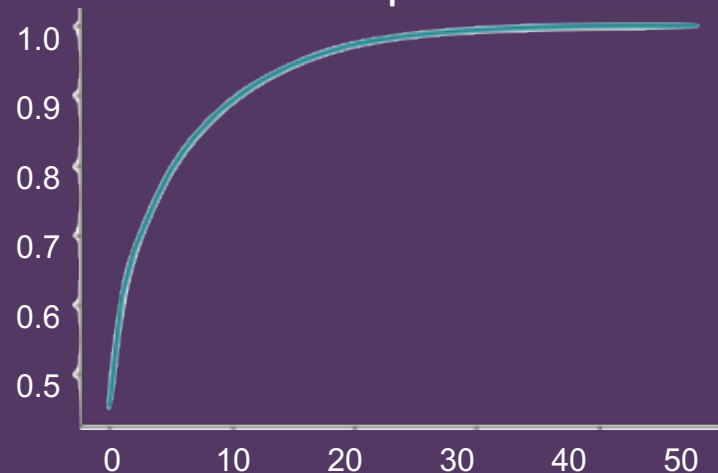
- ☐ Train: Before S2015
- ☐ Test: After S2015

NBA Players



☐ Apply SMOTE to Train Data

Cumulative Explained Variance



☐ Threshold=95%

Reduced to 14 Features

04

Model Selection

- >Potential Candidate: Binary Classification
- >The MVP : Regression
- >Models Performance

Binary Classification

Determine if a player is selected as an MVP candidate



	Logistic Regression (L2)	SVM	Random Forest
Accuracy:	95.42 %	97.25 %	98.38%
Precision:	37.04 %	49.47 %	64.29%
Recall:	100.0 %	94.00 %	90.00%
F-1 Score:	54.05 %	64.83%	75.00%

Model interpretation

Accuracy: Not informative due to imbalanced data

Recall: Very important since we wanted to get every potential candidate

L2: Helps reduce model complexity and multi-collinearity

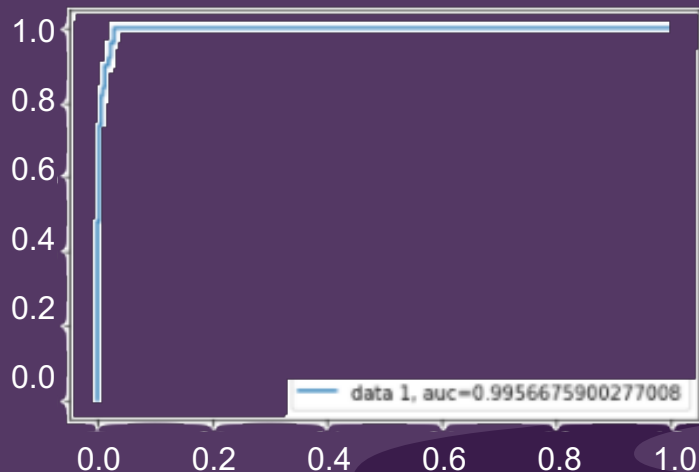
		Predicted Label	
Actual Label	0	1720	85
	1	0	50
		0	1

Binary Classification

Determine if a player is selected as an MVP candidate

Best Model: Logistic Regression with L2 Regularization

ROC



AUC=0.996

Reference: Lecture Slide

10-Fold Cross Validation

Accuracy: 96.27 %	
Precision:	41.32 %
Recall:	100.0 %
F-1 Score:	58.48 %

2021-2022 Predicted MVP Candidates Pool

0	Giannis Antetokounmpo	21	Khris Middleton	40	Kyrie Irving
1	Nikola Jokić	22	Darius Garland	41	Jonas Valančiūnas
2	Joel Embiid	23	Nikola Vučević	42	Tobias Harris
3	Luka Dončić	24	Anthony Davis	43	D'Angelo Russell
4	Ja Morant	25	Zach LaVine	44	Evan Mobley
5	LeBron James	26	Fred VanVleet	45	Brandon Ingram
6	Stephen Curry	27	Kyle Lowry	46	Steven Adams
7	Chris Paul	28	Paul George	47	Mike Conley
8	DeMar DeRozan	29	Julius Randle	48	Al Horford
9	Jimmy Butler	30	Deandre Ayton	49	Shai Gilgeous-Alexander
10	Trae Young	31	Jaylen Brown	50	Andrew Wiggins
11	Karl-Anthony Towns	32	LaMelo Ball	51	Miles Bridges
12	Kevin Durant	33	Jarrett Allen	52	Anthony Edwards
13	Jayson Tatum	34	Russell Westbrook	53	JaVale McGee
14	Devin Booker	35	Draymond Green	54	Jalen Brunson
15	Donovan Mitchell	36	Bradley Beal	55	Mikal Bridges
16	Bam Adebayo	37	Robert Williams	56	Bobby Portis
17	Rudy Gobert	38	Tyler Herro	57	Scottie Barnes
18	Jrue Holiday	39	Jaren Jackson Jr.		
19	Pascal Siakam				
20	Dejounte Murray				

Regression

Predict the MVP from MVP Candidates' Pool

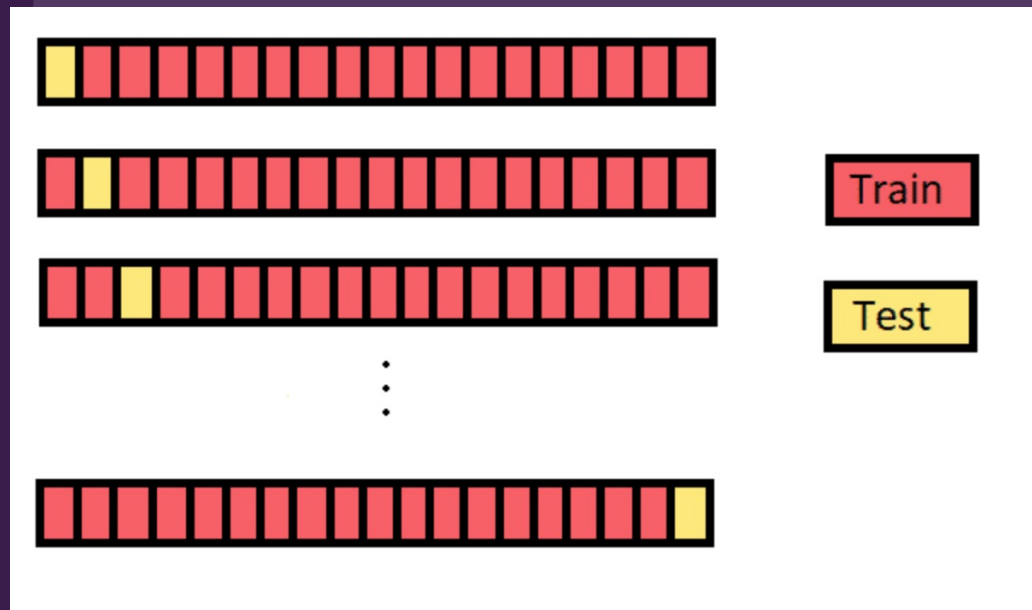
	Linear Regression	Random Forest	XGBoost	KNN
RMSE	3.211	2.546	2.548	3.350
R-Square	0.547	0.715	0.715	0.507
Prediction Accuracy	?	?	?	?

XGBoost:

- Performs well in ranking problem
- Relatively hard to do hyper parameter tuning

- Random Forest and XGBoost have a similar performance
- Need to introduce a new evaluation metric (Prediction Accuracy) to distinguish them

Leave-One-Out Cross Validation (LOOCV)



Result from Random Forest:

	Predicted MVP	Actual MVP	Label
year			
1980	Larry Bird	Kareem Abdul-Jabbar	incorrect
1981	Julius Erving	Julius Erving	correct
1982	Magic Johnson	Moses Malone	incorrect
1983	Larry Bird	Moses Malone	incorrect
1984	Larry Bird	Larry Bird	correct
1985	Larry Bird	Larry Bird	correct
1986	Larry Bird	Larry Bird	correct
...
2015	James Harden	Stephen Curry	incorrect
2016	LeBron James	Stephen Curry	incorrect
2017	Russell Westbrook	Russell Westbrook	correct
2018	James Harden	James Harden	correct
2019	James Harden	Giannis Antetokounmpo	incorrect
2020	James Harden	Giannis Antetokounmpo	incorrect
2021	Nikola Jokić	Nikola Jokić	correct


Regression

Predict the MVP from MVP candidates' pool

Leave-One-Out Cross Validation:

1. Provide an extra evaluation metric: Prediction Accuracy
2. Approximately unbiased result
3. Computational Expensive

LOOCV Results:

	Linear Regression	 Random Forest	XGBoost	KNN
R-Square	0.4455	0.6227	0.615	0.4963
MAE	2.3484	1.6712	1.684	1.9857
Accuracy	0.6381	0.7143	0.666	0.5476

Predicted MVP: Nikola Jokić



Rank	Linear Regression	Random Forest	XGBoost
1	Nikola Jokić	Nikola Jokić	Giannis Antetokounmpo
2	Giannis Antetokounmpo	Giannis Antetokounmpo	Nikola Jokić
3	Joel Embiid	Joel Embiid	Joel Embiid
4	Rudy Gobert	Luka Dončić	Ja Morant
5	Karl-Anthony Towns	Ja Morant	Luka Dončić

2021-22 NBA MVP Award Tracker

The NBA MVP Award Tracker ranks candidates based on a model built using previous voting results. Players must have played in at least 70% of the league-wide average for team games to qualify.



Top Candidates [Compare the top candidates](#) [Share & Export](#) [Glossary](#)

Rk	Player	Team	W	L	W/L%	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	Prob%
1	Nikola Jokić	DEN	41	28	.594	62	62	33.1	9.9	17.3	.572	1.5	4.2	.349	8.4	13.1	.643	.614	4.8	5.9	.804	2.8	11.0	13.8	8.1	1.4	0.8	3.8	2.6	26.0	40.4%
2	Giannis Antetokounmpo	MIL	43	26	.623	58	58	32.8	10.1	18.5	.548	1.1	3.8	.303	9.0	14.7	.611	.579	8.3	11.5	.721	2.0	9.6	11.5	5.9	1.1	1.4	3.2	3.3	29.7	29.0%
3	Joel Embiid	PHI	41	26	.612	55	55	33.3	9.3	19.3	.485	1.4	3.7	.366	8.0	15.5	.514	.521	9.8	11.9	.822	2.1	9.2	11.3	4.3	1.1	1.5	3.0	2.7	29.9	10.3%
4	Chris Paul	PHO	54	14	.794	58	58	33.0	5.6	11.5	.487	1.0	3.1	.330	4.6	8.4	.546	.532	2.7	3.2	.843	0.3	4.2	4.5	10.7	1.9	0.3	2.4	2.0	14.9	5.9%
5	Luka Dončić	DAL	42	26	.618	52	52	35.7	9.8	21.7	.453	2.9	8.5	.339	7.0	13.2	.526	.519	5.5	7.4	.741	0.9	8.4	9.3	8.6	1.2	0.6	4.5	2.3	28.0	3.3%

Vegas INSIDER

MARCH MADNESS NBA NCAAB NHL UFC AUTO GOLF MOR

NBA Home NBA Odds NBA Futures Market NBA Buy Picks NBA Free Picks NBA Scores

2021-22 NBA MVP ODDS

Player (Team)	Odds
Joel Embiid (Philadelphia)	-135
Nikola Jokic (Denver)	+200
Giannis Antetokounmpo (Milwaukee)	+800
Ja Morant (Memphis)	+1200
DeMar DeRozan (Chicago)	+1800

05

Conclusion

Improvements

Design Improvement

- Include features in other areas besides players' performances.
 - Use NLP to study the opinions of the media
 - Subject matter experts

Model Improvement

- Models to explore
 - Hyper-parameter tuning on XGBoost
 - Non-Linear Models

Future Extension

- Other Use Case
 - NBA Trade Result
 - All-STAR prediction

Meet the Team's pets



Crystal Xu



Jiaqi Feng



Zoey Zhou



Diandian Yuan



Jack Chen

THANKS

