EIN6905 Data Analytics for Social Good

Evaluation of All NBA Team Selection Mechanism

Chien-Wei, Chen

April 2022

Contents

List of Figures

Chapter 1

Project Overview

1.1 Motivation, Data, and Methodology Introduction

All NBA Team selection is the highest honor that one player can achieve, only top 15 players in this league would be crowned with this title. Most of the time, the title is strongly related to player's income, as many organizations reward their players with a great amount of money if selected as one of the All NBA Team players, using it as an incentive to motivate their players to thrive for greatness. Therefore, how fair the selection mechanism is becomes a worth-discussing problem. Current selection is solely based on the media voting, and that leads to an issue - Subjectivity, which I'd like to know the magnitude of it. Furthermore, a machine learning model would be built to predict whether a player would be selected as an All NBA Team player by inputting one's statistics.

Two data sets would be used in this project, one is the 2021 advanced statistics and the other one is advanced statistics from 2010 to 2021. The reason why traditional statistics, one that contains points, rebounds, and assists per game, is not considered in this project is because traditional statistics could be very misleading and misrepresenting. Since one player could hold impeccable personal statistics, yet his team is one of the worst teams in the league. Russell Westbrook was a salient example who could corroborate my reasoning. He averaged nearly triple double (Double-digit in points, rebounds and assists) throughout the season, but his team did not even make it to the playoffs (Houston Rockets 2020).

For the first data set, gurobipy optimization would be used to determine whether the 2021 All NBA Team selection is fair or not. And the second data set, containing roughly 6200 records, would be used for the machine learning classification model, determining whether a player would be selected as All NBA Team player.

1.2 Data Source

https://www.basketball-reference.com/leagues/NBA_2021_advanced.html

Chapter 2

Optimization

2.1 Data Preprocessing

2.1.1 Data Overview

Rk	Player	Pos	Age	Tm	G	GS	MP	PER	TS%	3PAr	FTr	ORB%	DRB%	TRB%	AST%	STL%	BLK%	TOV%	USG%	OWS	DWS	WS	WS/48	OBPM	DBPM	BPM	VORP	ALL NBA
1	Precious Achiuwa	PF	21	MIA	61	4	737	14.2	0.550	0.004	0.482	11.5	20.6	16.1	6.1	1.3	4.0	13.5	19.5	0.3	1.0	1.3	0.085	-3.6	-0.5	-4.1	-0.4	0
2	Jaylen Adams	PG	24	MIL	7	0	18	-6.5	0.125	0.250	0.000	0.0	16.9	8.8	12.7	0.0	0.0	0.0	18.6	-0.1	0.0	-0.1	-0.252	-15.1	-4.6	-19.8	-0.1	0
3	Steven Adams	C	27	NOP	58	58	1605	15.1	0.596	0.010	0.438	14.4	20.4	17.4	9.1	1.6	2.2	17.5	11.7	2.3	1.7	4.0	0.119	-0.4	0.1	-0.3	0.7	0
4	Bam Adebayo	C	23	MIA	64	64	2143	22.7	0.626	0.010	0.443	7.7	22.6	15.3	26.9	1.7	3.2	15.0	23.7	5.6	3.2	8.8	0.197	2.9	2.0	4.9	3.7	0
5	LaMarcus Aldridge	C	35	TOT	26	23	674	15.7	0.556	0.270	0.159	3.0	15.8	9.4	11.0	0.8	3.7	7.9	22.2	0.5	0.6	1.1	0.080	-0.2	-0.2	-0.3	0.3	0
5	LaMarcus Aldridge	C	35	SAS	21	18	544	15.1	0.545	0.302	0.149	3.3	15.4	9.2	10.2	0.7	2.8	7.0	22.7	0.3	0.5	0.8	0.067	-0.2	-0.7	-0.9	0.2	0
5	LaMarcus Aldridge	C	35	BRK	5	5	130	18.2	0.611	0.104	0.208	1.8	17.8	10.2	14.3	1.1	7.4	11.8	19.9	0.2	0.2	0.4	0.135	0.1	2.1	2.2	0.1	0
6	Ty-Shon Alexander	SG	22	PHO	15	0	47	4.2	0.349	0.750	0.167	4.9	19.0	12.1	15.4	0.0	1.9	18.9	15.0	-0.1	0.0	0.0	-0.048	-4.8	-1.7	-6.5	-0.1	0
7	Nickeil Alexander-Walker	SG	22	NOP	46	13	1007	12.5	0.522	0.478	0.144	1.4	14.1	7.8	14.7	2.2	2.1	12.4	23.2	-0.3	1.0	0.7	0.035	-1.4	0.1	-1.3	0.2	0
8	Grayson Allen	SG	25	MEM	50	38	1259	12.8	0.586	0.662	0.220	1.6	12.0	6.7	11.5	1.7	0.6	9.6	16.8	1.5	1.2	2.7	0.101	-0.2	0.1	-0.2	0.6	0

Figure 2.1: Data Overview 2021

Important Features	Explanation
G	Game played
GS	Game played as starting lineup
MP	Minutes played
PER_1	Player Efficiency Rating
BPM_2	Box Plus/Minus

PER: The Player Efficiency Rating sums up all player's positive accomplishments, subtracts the negative accomplishments, and returns a per-minute rating of a player's performance.
BPM: Box Plus/Minus is a basketball box score-based metric that estimates a basketball player's contribution to the team when that player is on the court.

2.1.2 Data Cleaning

As shown in Figure 2.1, there are duplicates in the data set, it is a normal phenomenon as player could be traded during the season. Duplicates and null values need to be cleaned.

Based on All NBA Team selection criteria, players who are not qualified should also be factored out. For instance, a player had incredible advanced statistics, but he only played 2 games, this situation would greatly impact the optimization model and give us misrepresenting result. Therefore, only players who played over 45 games as starting lineup and played over 900 minutes would be considered.

2.2 Optimization

After applying gurobipy optimization, the players that are selected by the model is 73% similar to the players that were selected by the media, as shown below (See code for details).

```
# Check how accurate the model is by comparing it with the actuall All NBA team players
# Filtering
Selected = list(df[df['ALL NBA'] == 1]['Player'])
# Comparing
correct = set(Selected_from_Model).intersection(set(Selected))
print('Accuracy:',len(correct)/len(Selected))
# print('Players who are selected by the model and the media', correct)
print('Players who are selected by the model, but not selected by the media', set(Selected_from_Model).difference(set(Selected)))
print('Players who are selected by the media, but not selected by the model', set(Selected).difference(set(Selected_from_Model)))
```

```
Accuracy: 0.7333333333333333
Players who are selected by the model, but not selected by the media {'Jonas Valančiūnas', 'Karl-Anthony Towns', 'Bam Adebayo', 'Zion Williamson'}
Players who are selected by the media, but not selected by the model {'Julius Randle', 'Bradley Beal', 'Chris Paul', 'Paul George'}
```

Figure 2.2: Result_1

Why is there such a big discrepancy? Is there any constraints that has not been taken care of?

The answer is yes. After investigation, even though it was not clearly stated in the criteria, position is one of the important factor that should be considered. Normally All NBA Team is consisted of 3 centers, 6 forwards and 6 guards. Based on the data of past 10 years and the current NBA trend, the demarcation between forwards and guards is less significant, since there are many versatile players who can play both forward and guard. Therefore, the constraint of selecting 6 forwards and 6 guards is lifted in this model. However, the number of center being selected has been relatively consistent throughout the years. Hence, one more constraint is imposed to the model, that being the number of center should be 3.

```
# Comparing
new_selected_from_model = Selected_centers + Selected_f_g

correct = set(new_selected_from_model).intersection(set(Selected))
print('Accuracy:',len(correct)/len(new_selected_from_model))
# print('Players who are selected by the model and the media', correct)
print('Players who are selected by the model, but not selected by the media', set(new_selected_from_model).difference(set(Selected)))
print('Players who are selected by the media, but not selected by the model', set(Selected).difference(set(new_selected_from_model)))
```

```
Accuracy: 0.8666666666666667
Players who are selected by the model, but not selected by the media {'Trae Young', 'Zion Williamson'}
Players who are selected by the media, but not selected by the model {'Julius Randle', 'Paul George'}
```

```
df_differnece_2 = df_all_nba[(df_all_nba['Player'] == 'Trae Young')|(df_all_nba['Player'] == 'Zion Williamson')|
                             (df_all_nba['Player'] == 'Julius Randle')|(df_all_nba['Player'] == 'Paul George')]
df_differnece_2.groupby('ALL NBA').mean()
```

ALL NBA	Age	G	GS	MP	PER	TS%	3PAr	FTr	ORB%	DRB%	TRB%	AST%	STL%	BLK%	TOV%	USG%	OWS	DWS	WS	WS/48	OBPM	DBPM	BPM	VORP
0	21.0	62.0	62.0	2075.5	25.05	0.6190	0.1950	0.5005	5.45	12.6	9.05	32.6	1.25	1.15	13.90	31.40	6.5	1.45	7.95	0.1840	5.7	-1.0	4.75	3.5
1	28.0	62.5	62.5	2244.0	20.10	0.5825	0.3655	0.2820	3.20	22.3	12.90	25.9	1.45	0.90	14.15	29.65	3.2	3.30	6.55	0.1395	3.4	0.6	4.00	3.4

Figure 2.3: Result_2

After incorporating position as one of the constraints, the similarity is enhanced up to 87%. However, based on further investigation, Zion Williamson and Trae Young might have lost their All NBA Team selection due to subjectivity, since their overall advanced stats were better than the ones who were selected by the media as shown in Figure 2.3.

Chapter 3

Machine Learning Model

3.1 Data Preprocessing

3.1.1 Data Overview

(6204, 28)

	Rk	Player	Pos	Age	Tm	G	MP	PER	TS%	3PAr	FTr	ORB%	DRB%	TRB%	AST%	STL%	BLK%	TOV%	USG%	OWS	DWS	WS	WS/48	OBPM	DBPM	BPM	VORP	ALL NBA
0	1	Arron Afflalo	SG	24	DEN	82	2221	10.9	0.576	0.426	0.168	3.1	9.9	6.5	9.3	1.0	1.0	10.5	14.0	2.8	1.4	4.3	0.092	-0.2	-0.2	-0.4	0.9	0
1	2	Alexis Ajinça	C	21	CHA	6	30	6.3	0.479	0.000	0.100	4.1	11.8	8.0	0.0	1.8	2.7	16.1	19.3	-0.1	0.0	0.0	-0.013	-6.3	1.0	-5.3	0.0	0
2	3	LaMarcus Aldridge	PF	24	POR	78	2922	18.2	0.535	0.014	0.260	8.1	18.6	13.3	9.9	1.3	1.3	7.4	22.9	5.5	3.3	8.8	0.145	1.4	-0.2	1.2	2.3	0
3	4	Joe Alexander	SF	23	CHI	8	29	2.8	0.273	0.167	0.500	7.8	11.3	9.6	9.3	1.8	2.6	0.0	11.3	0.0	0.0	0.0	0.030	-9.1	0.9	-8.3	0.0	0
4	5	Malik Allen	PF	31	DEN	51	456	5.9	0.431	0.052	0.112	9.2	11.5	10.4	5.1	1.2	0.8	15.3	14.0	-0.3	0.3	0.1	0.009	-4.7	-1.0	-5.7	-0.4	0

Figure 3.1: Data Overview_2010 to 2021

3.1.2 Data Cleaning

Before diving into the machine learning model, the data set needs to be cleaned. Unlike what has been done in the previous task, the duplicates would not be dropped this time, since the data set encompasses the data from 2010 to 2021, the player name would occur repetitively as they played many years in this league. What needs to be cleaned this time are null values, and columns that are not helpful in training the model, such as player name, team name and Rk.

3.2 Machine Learning Model

3.2.1 Feature Selection

```
# Feature selection using "mutual_info_classif" for classification model
bestfeatures = SelectKBest(score_func=mutual_info_classif, k= df2.shape[1]-1)
fit = bestfeatures.fit(X_train,y_train)
scores = pd.DataFrame(fit.scores_)
columns = pd.DataFrame(X.columns)

featureScores = pd.concat([columns,scores],axis=1)
featureScores.columns = ['Features','Score']

data=featureScores.nlargest(df2.shape[1]-1,'Score')
data
```

Figure 3.2: Feature Selection

Feature selection should be carried out as it increases the chance of using significant features rather than the redundant ones. In this case, Select K-best is used. One crucial parameter that needs to be specified is the score function, which is mutual_info_classif in this case, since this is a classification model. Consequently, Age and Position are the features that have to be removed. This result makes sense, since a player should be able to win All NBA selection regardless of his age. Also, position should not be affecting the model as all positions are equally eligible for being selected. Therefore, Age and Position would be dropped before moving onto the next step.

3.2.2 Feature Scaling

The range of the numerical data varies in the data set. Normalization has to be done to deal with the problem that each features has different scale.

The distribution of each features has to be investigated before applying scaling method, as different normalization method should be applied to different kinds of distribution respectively (Data Visualization is used here, see python code for reference). Min-Max normalization would be used for non-Gaussian distribution, while Standardization would be used for Gaussian distribution. The accuracy could be improved from 92% to nearly 95% if feature scaling is properly carried out.

3.2.3 Model and Hyper-parameter Tuning

As this is a classification, several classifier would be used, including Decision Tree Classifier, Logistic Regression and Support Vector Machine as shown below.

```
def models(X_train, y_train, X_test, y_test):

  dtclf = DecisionTreeClassifier(max_depth = 5, criterion='entropy')
  dtclf.fit(X_train, y_train)
  y_pred_train = dtclf.predict(X_train)
  y_pred_test = dtclf.predict(X_test)
  print('Decision Tree Testing set accuracy: ', accuracy_score(y_test, y_pred_test))
```

Figure 3.3: Machine Learning Model

After several rounds of trial and error process in tuning the hyper-parameter and different algorithms, the decision tree classifier with max_depth of 5 and using entropy as criterion actually renders the highest prediction accuracy, which is up to 95%. With only 5 layers and 14 terminal nodes would generalize well for future unknown data.

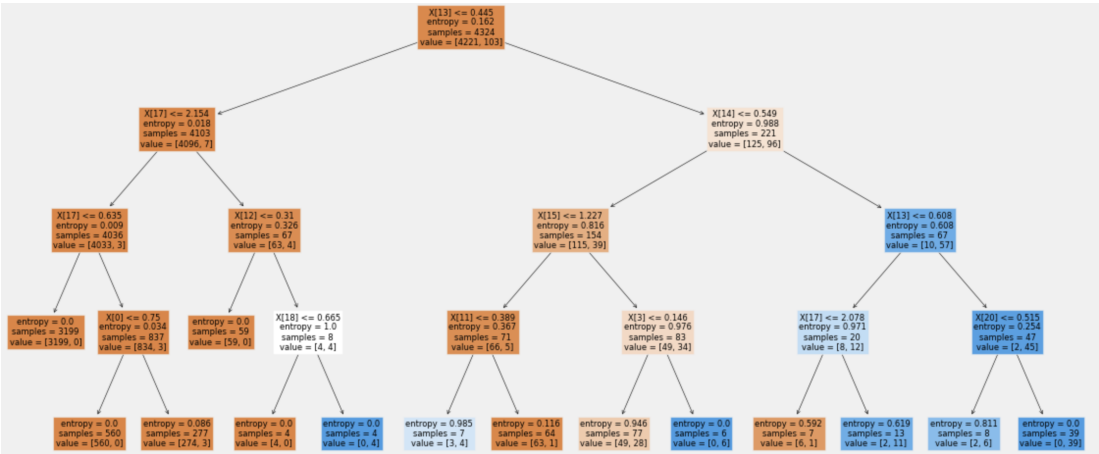

Figure 3.4: Tree plot

Chapter 4

Conclusion

Out of curiosity, I've applied the same gurobi optimization model on statistics from 2010 to 2021 and measured the subjectivity by understanding "How different are the players selected by the media from players selected by the model". I noticed that the overall subjectivity in All NBA Team selection has been mitigated through out the years as science and sports analytic have become more popular than ever in analyzing players' performance.

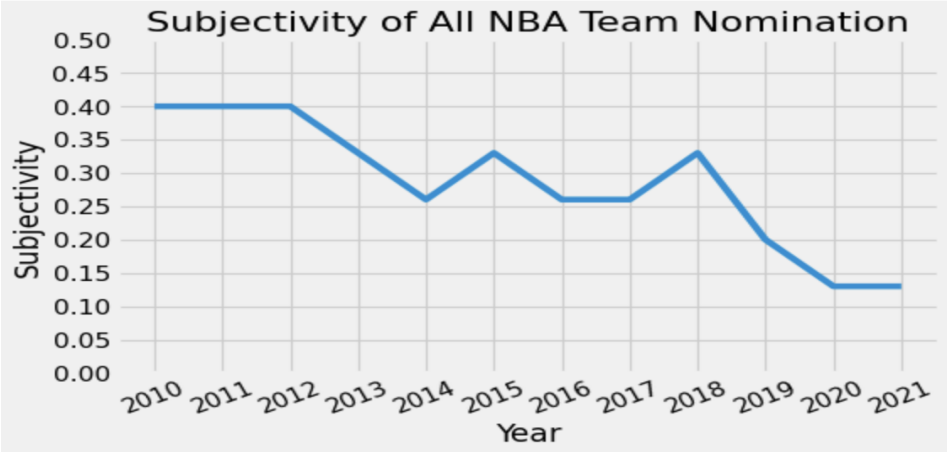

Figure 4.1: Magnitude of Subjectivity

Even though subjectivity is mitigated, it still exists. The selection is still somewhat affected by the impression. For example, how much does a player improve could potentially influence the outcome, player like Julius Randle is a perfect example for that, despite the fact that his statistics was actually not as good as Zion Williamson's, he still won the All NBA Selection. Many sports writer emphasized that he made a huge progress in 2021 compared with that of 2020, this improvement might have earned him some extra points in competing this title.

There are factors that were not considered in this project, and one of which is the business aspect of the NBA. Business is always the top priority of any kinds of professional sports, and it could also be the unwritten rule that influences the outcome. For instance, is it be possible that players who play for the large market base organizations, such as New York or Los Angeles, would have higher chance of winning personal awards? Since they would for sure bring much more revenue than that of smaller organizations like New Orleans and Phoenix. But I can only figure this question out if I enter sports industry in the future.

Chapter 5

Future Research

While using PER and BPM seem to be an objective way of evaluating players' performance, it still not able to encompass the whole picture, especially when it comes to the "clutch moment" or the so called "big time". However, how clutch a player is oftentimes defines a player's career. Take Michael Jordan and Karl Malone as an example, no one would deny the fact that Karl Malone is one of the greatest power forwards of all time, but not being able to make big time play was also the truth that dimmed his legacy. Michael Jordan, on the other hand, is the synonym of G.O.A.T, given that he was able to make clutch shot consistently with over 50% shooting percentage.

A future research would be carried out to address this issue, by using the concept of "weight" taught in this class. An initial idea would be assigning weights that indicate the importance of the time period within the game. For instance, the points that are scored during the last 10 seconds of the game would be weighted heavier than that of at the beginning of the game. Same concept goes to the game itself, a game that determines whether a team clinches to the playoffs or not would be weighted heavier than a normal regular season game. By understanding this, an all-around evaluation mechanism could be built, and be further utilized to evaluate players' performance.