Scouting for Future NBA Stars:

Implementation of Multi-classification Machine Learning Methods in Sports Analytics

Liangze Ke

Introduction: Background

1. What happens when a player gets into NBA?
   - Drafted, Sign a Rookie Contract, And become a free agent.
2. What are we doing here?
   - Evaluate how well a player has done in his rookie years (first four years after he was drafted).
   - In language of data science, predictive analysis: X (variables involved in the rookie years stats), Y (career outcome)
3. It is extremely marketable research because agencies and scouts make huge cashbacks after a player signs a good deal of contract. And signing players not just a simple procedure, but an investment.

Data

Two datasets we used are team_stats.csv and player_stats.csv, which relatively contains players' statistics and the award list.

They are mingled into a panel dataset, which has 900 unique players spanning 14 years of NBA drafts.

Each season (yearly basis), each player, their performance statistics and award statistics are recorded.

Throughout initial inspections, the following variables are relevant.

Abbreviation/Acronym	Term Meanings
season/nbapersonid	Season and Player ID
All NBA Defensive Team	Elected to All NBA Defensive Team
All NBA Team	Elected to All NBA Team
All Rookie Team	Elected to All Rookie Team
Finals MVP	Won the Finals MVP award
Player Of the Month	Player of the Month awards won
Player Of the Week	Player of the Week awards won
Rookie Of the Month	Rookie of the Month awards won
all_star_game	Selected to an all-star team
rookie_all_star_game	Selected to a rookie all-star team
allstar_rk	Player's rank in all-star voting
DPOY_rk	Player's rank in DPOY voting
MIP_rk	Player's rank in MIP voting
MVP_rk	Player's rank in MVP voting
ROTY_rk	Player's rank in ROTY voting
draftyear	Year the player was drafted
draftpick	Draft pick number
team	Team abbreviation
games	Number of games played
games_start	Number of games started
mins	Total minutes played
off_reb	Offensive rebounds
def_reb	Defensive rebounds
tot_reb	Total rebounds
ast	Total Assists
steal	Total Steals

X and y, how we feature engineered?

Variable	Equation
First_year_ROM_amt	$\sum$ Rookie of the Month awards in draft year
All_Rookie_First	1 if (rookie_all_star_game == 1), 0 otherwise
All_Rookie_Second	1 if (rookie_all_star_game == 2), 0 otherwise
four_yrs_POW	$\sum_{\text{year=draftyear}}^{\text{year=draftyear+3}}$ Player of the Week awards
Rookie_ppg	$\left(\sum_{i=1}^{82} \text{points\_draftyear}/82\right)/4$
Rookie_apg	$\left(\sum_{i=1}^{82} \text{assists\_draftyear}/82\right)/4$
Rookie_rpg	$\left(\sum_{i=1}^{82} \text{tot\_rebound\_draftyear}/82\right)/4$
Rookie_tov_pct	$\sum_{\text{year=draftyear}}^{\text{year=draftyear+3}} \text{tov\_pct}/4$
Rookie_mins	mins_draftyear
avg_4yr_ppg	$\sum_{\text{year=draftyear}}^{\text{year=draftyear+3}} \left(\sum_{i=1}^{82} \text{points\_i}/82\right)/4$
avg_4yr_apg	$\sum_{\text{year=draftyear}}^{\text{year=draftyear+3}} \left(\sum_{i=1}^{82} \text{assists\_i}/82\right)/4$
avg_4yr_rpg	$\sum_{\text{year=draftyear}}^{\text{year=draftyear+3}} \left(\sum_{i=1}^{82} \text{tot\_rebound\_i}/82\right)/4$
tot_mins	$\sum_{\text{year=draftyear}}^{\text{year=draftyear+3}}$ mins
tot_games	$\sum_{\text{year=draftyear}}^{\text{year=draftyear+3}}$ games
tot_games_start	$\sum_{\text{year=draftyear}}^{\text{year=draftyear+3}}$ games_start

TABLE II

Player Classification	Criteria
Elite	Won any All NBA award (1st, 2nd, 3rd team), MVP, or DPOY
All-Star	Selected as an All-Star in the season
Starter	Started in at least 41 games OR played at least 2000 minutes in the season
Rotation	Played at least 1000 minutes in the season
Roster	Played at least 1 minute in the season but did not meet any of the above criteria
Out of the League	Not in the NBA in that season

TABLE III

Classification	Career Outcome Criteria
Top Class	Career outcome is either Elite or All-Star
Middle Class	Career outcome is Starter, Rotation, or Roster
Incompatible	Career outcome is Out of the League

TABLE IV

PLAYER CAREER OUTCOME CLASSIFICATIONS (3-CATEGORY)

Train? Test? Predict?

Train-test: 2007-2015 drafted players (they have already got their career outcome)

Develop a model: Logistics, Random Forest

Predict: 2018 drafted player (new emerging stars have just finished their four rookie years)

Literature review: We are pioneers

Only few literature works were published about data analytics of NBA players. Most works were concealed by agencies and specialists.

A Mixed Model for Performance-Based Classification of NBA Players

Yeong Nain Chi[1], and Jennifer Chi[2]
[1]University of Maryland Eastern Shore, [2]University of Texas at Dallas

1. Lack marketable endeavor
2. Classify into three categories “key”, “bench”, and “supporting”, which are not mainstream categories (awards-oriented)

Goals and Hypotheses

Goals: Predicting 2018 class players. What categories will 2018 class players fall into?

Hypothesis: Their career outcome is based on the rookie years performances

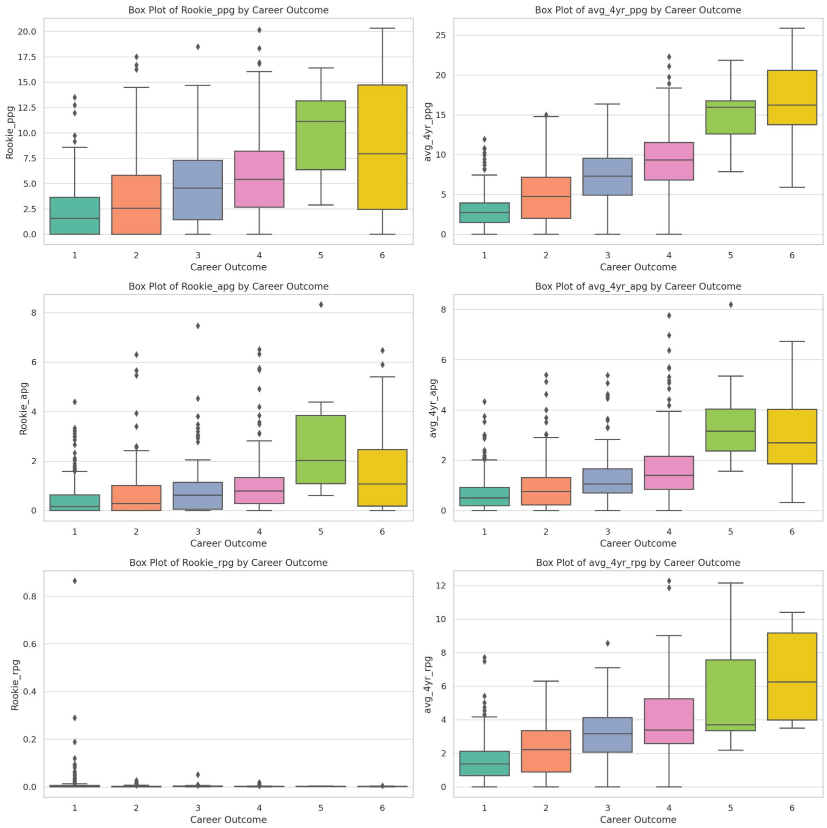
Reasons:

1. Rookie years determine how much a coach value a player.
2. Recent basketball leagues are abundant, so if a player did not perform well in his rookie year, he might as well go overseas.
3. Rookie years determine the “initial impression” of supporters, and their votes are crucial for selective accolades, which determine their career outcome.

Descriptive analysis

This figure shows a series of box plots displaying the distribution of several numerical column across different career outcomes.

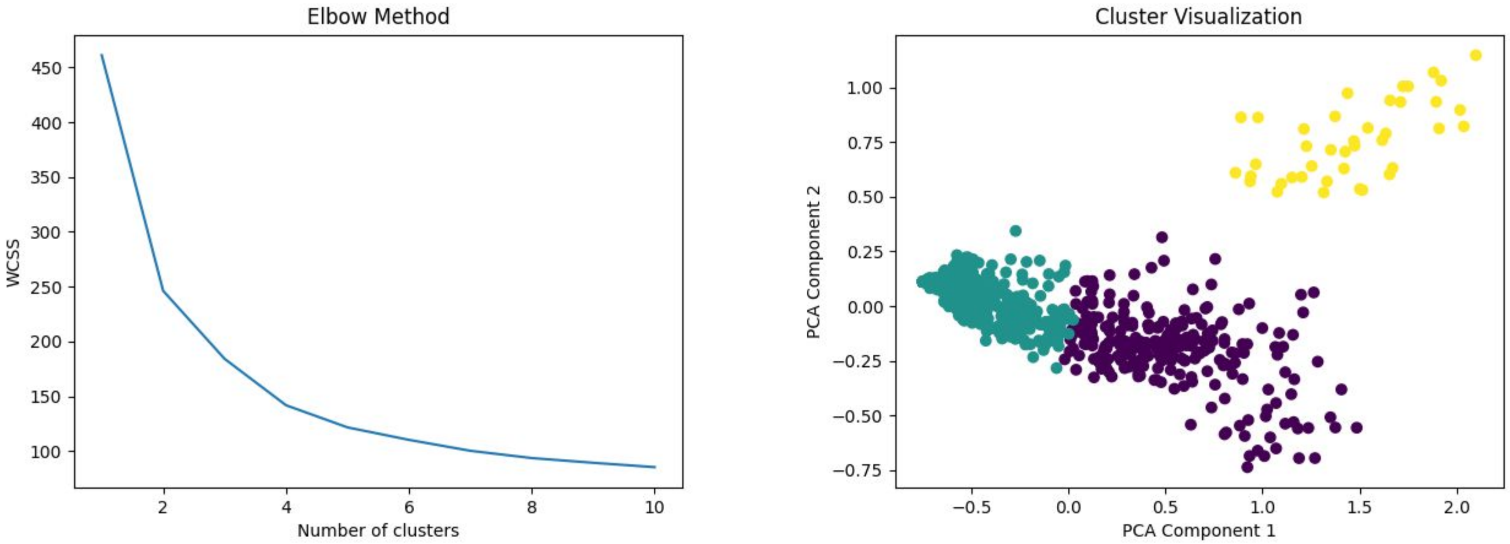
Generally, We can found out that, as rookie players' career outcome become more successful, their first year gained points per game, average 4 year points per game, average 4 year assist per game, and average 4 year rebound per game increases as well.



Method-- Introduction

1. K-Means Clustering
2. Logistic Regression
3. Random Forest Classification

Method-- K-means clustering



Method-- Logistic Regression (1)

1. Model: llog_reg = LogisticRegression(max_iter=1000)

2. Feature Engineering
    a. X = 'draftpick', 'first_year_ROM_amt', 'All_Rookie_First', 'All_Rookie_Second', 'four_yrs_POW', 'Rookie_ppg', 'Rookie_apg', 'Rookie_rpg', 'Rookie_tov_pct', 'Rookie_mins', 'avg_4yr_ppg', 'avg_4yr_apg', 'avg_4yr_rpg', 'tot_mins', 'tot_games', 'tot_games_start'

    b. Y = mapping = { 'Elite': 6, 'All-Star': 5, 'Starter': 4, 'Rotation': 3, 'Roster': 2,'Out of the League': 1}

Result-- Logistic Regression (1)

LOGISTIC REGRESSION MODEL CLASSIFICATION REPORT (1)

Label	Precision	Recall	F1-score	Support
1 (Out of the League)	0.72	0.90	0.80	103
2 (Roster)	0.25	0.06	0.09	35
3 (Rotation)	0.27	0.23	0.25	26
4 (Starter)	0.48	0.48	0.48	31
5 (All-Star)	0.00	0.00	0.00	3
6 (Elite)	0.18	0.40	0.25	5
Accuracy			0.58	203
Macro Average	0.32	0.35	0.31	203
Weighted Average	0.52	0.58	0.53	203

Method-- Logistic Regression (2)

1. Model: log_reg = LogisticRegression(max_iter=1000)

2. Feature Engineering
   a. Simplify player outcome labels

```python
# Mapping the 'career_outcome' to 'simple_label'
def map_to_simple_label(career_outcome):
    if career_outcome in [2, 3, 4]:  # Combining 'Roster', 'Rotation', 'Starter'
        return 2
    elif career_outcome in [5, 6]:  # Combining 'All-Star', 'Elite'
        return 3
    else:  # 'Out of the League'
        return 1
```

Result-- Logistic Regression (2)

LOGISTIC REGRESSION MODEL CLASSIFICATION REPORT (2)

Label	Precision	Recall	F1-score	Support
1 (Out of the League)	0.78	0.92	0.84	100
2 (Roster)	0.80	0.69	0.74	93
3 (All-Star/Elite)	0.40	0.20	0.27	10
Accuracy			0.78	203
Macro Average	0.66	0.60	0.62	203
Weighted Average	0.77	0.78	0.77	203

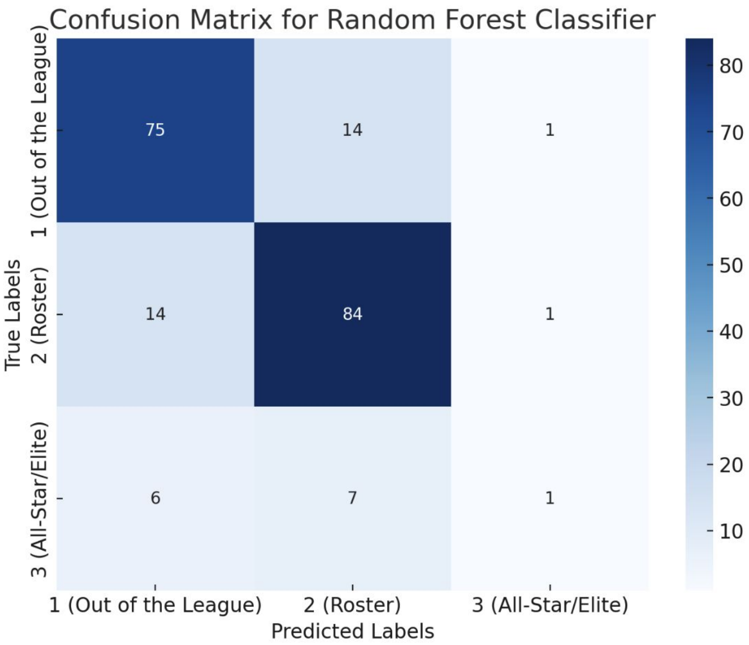
Method-- Random Forest Classifier (1)

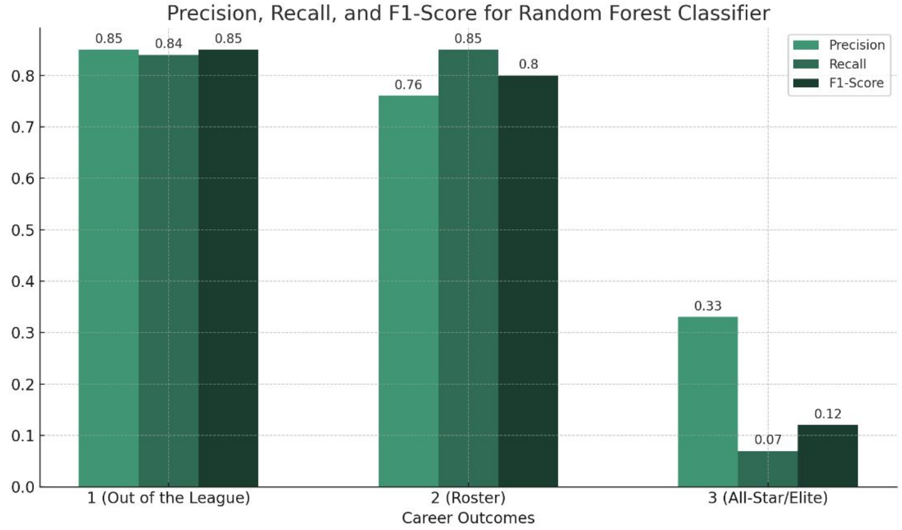
1. Model: rf_classifier = RandomForestClassifier()

2. Feature Engineering
    a. Simplify player outcome labels

```
# Mapping the 'career_outcome' to 'simple_label'
def map_to_simple_label(career_outcome):
    if career_outcome in [2, 3, 4]:  # Combining 'Roster', 'Rotation', 'Starter'
        return 2
    elif career_outcome in [5, 6]:  # Combining 'All-Star', 'Elite'
        return 3
    else:  # 'Out of the League'
        return 1
```

Result-- Random Forest Classifier (1)





Method-- Random Forest Classifier (2)

1. Model: rf_classifier = RandomForestClassifier()

2. Feature Engineering
    a. Simplify player outcome labels

```
# Mapping the 'career_outcome' to 'simple_label'
def map_to_simple_label(career_outcome):
    if career_outcome in [2, 3, 4]:  # Combining 'Roster', 'Rotation', 'Starter'
        return 2
    elif career_outcome in [5, 6]:  # Combining 'All-Star', 'Elite'
        return 3
    else:  # 'Out of the League'
        return 1
```

Method-- Random Forest Classifier (2)

1. Feature Engineering
   a. Simplify player outcome labels
   b. Separating numerical and categorical features

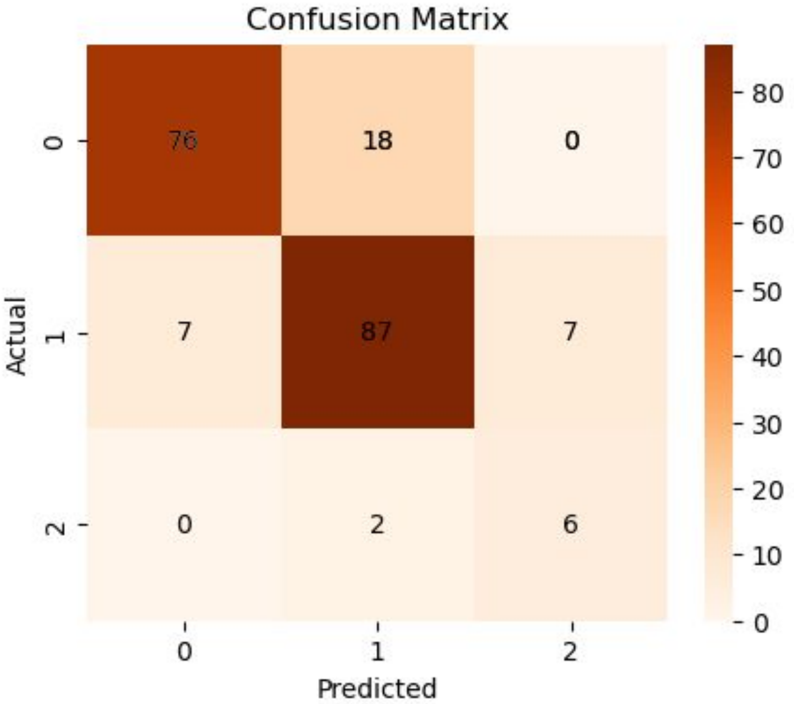
```
numerical_features = X[['first_year_ROM_amt', 'Rookie_ppg', 'Rookie_apg',
                        'Rookie_rpg','Rookie_tov_pct', 'Rookie_mins',
                        'avg_4yr_ppg', 'avg_4yr_apg', 'avg_4yr_rpg',
                        'tot_mins', 'tot_games', 'tot_games_start']]
categorical_features = X[['draftpick', 'All_Rookie_First', 'All_Rookie_Second',
                        'four_yrs_POW']]
```

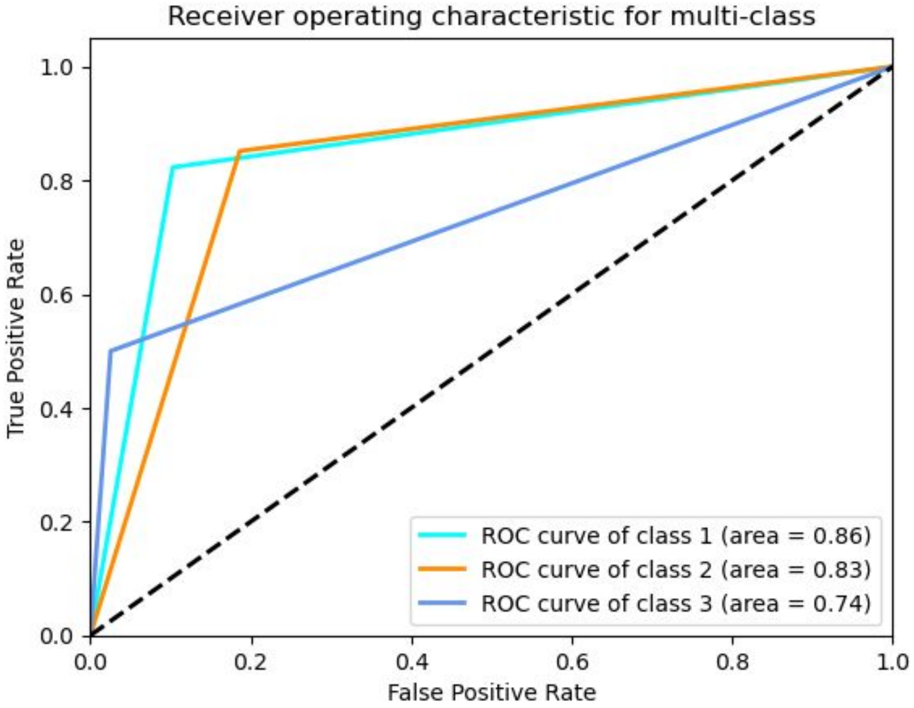
Method-- Random Forest Classifier (2)

1. Feature Engineering
    a. Simplify player outcome labels
    b. Separating numerical and categorical features
    c. Apply SMOTE on training data

```
smote = SMOTE()
X_train_smote, y_train_smote = smote.fit_resample(X_train_combined, y_train_combined)
```

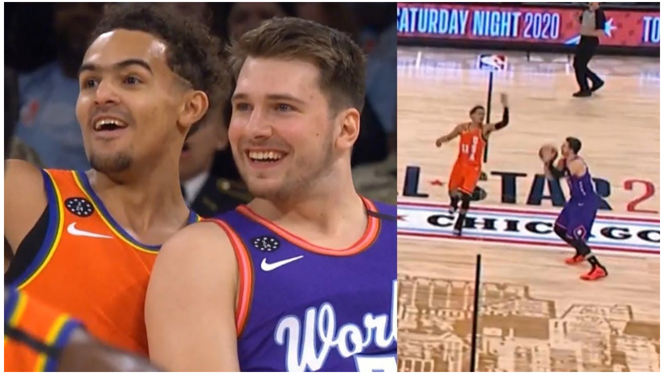
Result-- Random Forest Classifier (2)





Try our model with the 2018 class prediction

player	model_results
Marvin Bagley	3
Shai Gilgeous-Alexander	3
Collin Sexton	3
Trae Young	3
DeAndre Ayton	3
Luka Dončić	3
Jaren Jackson	3
Kevin Huerter	3
Jemerrio Jones	2
Haywood Highsmith	2
Rawle Alkins	2
Grayson Allen	2
Mohamed Bamba	2
Keita Bates-Diop	2
Mikal Bridges	2
Miles Bridges	2
Bruce Brown	2
Troy Brown	2
Jalen Brunson	2
Jevon Carter	2
Wendell Carter	2
Hamidou Diallo	2
Donte DiVincenzo	2
Kevin Hervey	2
Aaron Holiday	2
Chandler Hutchison	2
De'Anthony Melton	2
Shake Milton	2
Svi Mykhailiuk	2
Josh Okogie	2



Conclusion

- Sports analytics is a good area to implement newly developed models.
- It can be evidently concluded that rookie years are crucial for a NBA player’s career.
- Rookie years stats are extremely important to the contract signed as a free agent.
- However, there are some players lies above our prediction, which tells us the effort will finally make impact even though a player did not start off well.

Difficulties

1. Lack of big data of players
2. Lack of past literature, we are on our own
3. Subjective-based theory and hypothesis