Scouting for Future NBA Stars: Implementation of Multi-classification Machine Learning Methods in Sports Analytics

Liangze Ke

Abstract— **This paper presents an innovative approach to scouting future NBA stars using multi-classification machine learning methods in sports analytics. Focusing on the crucial balance between team budgets and player salaries within the NBA's wage cap constraints, the study proposes a several ML method involving k-means, logistic regression, and Random Forest to predict and classify players' potential and performance levels. The research analyzes data from players drafted between 2007 and 2015, achieving an 82% accuracy rate in predicting career outcomes. This approach provides a strategic and systematic framework for teams to assess player value accurately, offering significant implications for player scouting and team management in professional basketball.**

I. Introduction

In the high-stakes world of the National Basketball Association (NBA), the balance between team budgets and player salaries is a critical factor shaping the league dynamics. Despite the NBA's substantial profits and the generous salaries awarded to players, each team operates under a stringent constraint known as the "wage cap." This cap limits the total amount a team can spend on player salaries, irrespective of the team's financial capacity. Given these financial constraints, it becomes imperative for teams to assess the value of players accurately, particularly in the context of the scouting process.

The scouting process is a crucial element in predicting and evaluating the potential of young, emerging talent within the league. Players entering the NBA are designated as "rookies," typically bound by a standard four-year contract. This contract includes two guaranteed years, followed by two optional years. Post these initial years, a player attains the status of a free agent, opening the doors for public assessment and potential contracts from interested teams.

In this landscape, a critical question arises: What is the true worth of a player based on their performance during the rookie years, especially considering the limitations imposed by the wage cap? This essay aims to address this question by proposing a novel, strategic, and systematic approach based on Random Forrest, k-means, SVM, logistic regression, and so on. Our method seeks to predict a player's future performance level and classify their potential tier in the professional hierarchy.

As we delve into this analysis, several key terms and concepts will be frequently referenced, which are vital for understanding our methodology and conclusions:

1) **rookie**: An NBA player who has been drafted in recent years, typically two years. In this article, it will be referred to as the first four years instead of two.
2) **draft** (also known as a **class**): A selection conference in which each qualified team can choose their designated player from the potential rookie pool, which typically includes high-ranking college basketball players and overseas players. The year a player is chosen to become an NBA rookie is called "draft year".
3) **rookie contract**: A contract signed by a newly drafted NBA rookie, which contains the first two years of guaranteed salary and a third and fourth years of optional salary, four years total.
4) **free agent**: A player that has abided every period of the previous contract, and currently seeking a new contract by any team in the league. In this article, it is referred to as the first time a rookie player becomes a free agent, which is the end of their rookie contract.
5) **season**: An NBA season refers to a series of matches that will result in a single champion among teams in the league. Typically, seasons will be named based on calendar years, like 16-17 season, which means a season that holds across the 2016 and 2017 calendar years. In this article, to get rid of ambiguity, the season is named based on the starting calendar year. The 16-17 season will be referred to as the 2016 season.
6) **wage cap**: A price ceiling set by the NBA official that no teams could spend more than that price to sign or pay their players.

As it could all possibly be searched, this article is a pioneer in the scouting field. There has been little analogous literature that could be found. Nevertheless, some similar (but not analogous) articles are worth mentioning and comparing.

The endeavor of this paper can be divided into three parts, data preprocessing (including scraping, cleaning, and clustering), classification model training (based on NBA players who were drafted from 2007 to 2015 stats), and predictive scouting trial (apply the trained model to 2018 drafted rookies.) To the best effort, and 82% accuracy rate has been achieved based on the training dataset, and the 2018 scouting trial has resonated with current commentators' and senior correspondents' mainstream opinions.

II. Literature Review

While most sports analytics are done in scouting agencies, and whose works are concealed, there are several literal works published about classifying or predicting NBA players. A k-means classification trained by Chi and Chi[4] shows a high statistic performance. They classify a panel of NBA players based on the variables they meticulously choose. However, the studies lack practical inferences. Their classes are "supporting", "key", and "bench". This classification, though achieves high statistical scores, lacks evidence bases. Questions like "What would be given to a "supporting" achieve?", or "What does a supporting player have in common in NBA?". The links between their classes and an expectation of classes from a regular NBA supporter are relatively opaque. Moreover, their studies lack commercial value. Their statistical significance would not be marketable if no monetary context was spoken. A K Nearest Neighbors classifier was trained by Wu [9] and his colleagues and achieved remarkable statistics scores. They successfully put their findings into a monetary-oriented field. However, their methods are too simple and monotone, as the job of selecting the best method amongst them is hardly visible. Moreover, their dependent variable, NBA player salary, is also based so much on off-court behaviors (like engaging in commercials, politics, and social media). Such a confounding effect is so conspicuous that the result could hardly be convincing. Ergual [5] did a logistic regression to classify NBA teams based on their season performances. They have done a meticulous cause-and-effect analysis on their logistic regression, and their independent variables have achieved satisfying robustness. However, team classification has little practical meaning because teams are volatile due to players' transactions, managerial change, and unpredictable fan base. Other works, including [8], and [7], are using similar methods attributed to sports analytics.

III. Data

Two datasets are scraped using BeautifulSoup from the NBA official site[1], ESPN Sports[2], Sky Sports[3], NBA2konlinec[4], one containing player statistics (team_stats.csv) and the other an award list(player_stats.csv). Two datasets are mingled into a panel dataset, which observes 900 unique players spanning 14 years of NBA drafts, with detailed numerical records of player performance for each season. 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

TABLE I
NBA Player Performance Metrics

A. Training dataset: `Training.csv`

The endeavor is to integrate and compose a fruitful and informational training dataset. The units of observation are players who were drafted from 2007 to 2015. They are players who have already retired or past their primes (around 7 seasons after he was drafted). Therefore, it is sophisticated enough to conclude their career outcome. However, acknowledging, that there are a few players whose prime periods would have come late, they are considered outliers and the amount of this scenario is too scarce to be worth noticing. After pruning all of the units of observations from the raw dataset, calculations of the new variables based on Table I variables have been meticulously done.

- Season standardization. A regular season always contains 82 games between 30 teams in the NBA. However, there were only averaging 66 games played in 2011 due to a "lockout". Therefore, 2011 season statistics will be linearly scaled into 82 games.
- Stats averaging. The stats are manipulated using the following equations.
- Single season labeling.
- Career outcome labeling (6 categories). The explanation provided is about a system for classifying NBA players based on the highest level of success they achieve in each season of their career. Here's a breakdown of the concept:
  Season-Level Outcome: A player's outcome for a given season is determined by the highest level

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

VARIABLES AND EQUATIONS FOR NBA PLAYER PERFORMANCE ANALYSIS

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

NBA PLAYER CLASSIFICATION CRITERIA (6-CATEGORY)

of success they achieve in that season. The levels are categorized as "Elite," "All-Star," "Starter," "Rotation," "Roster," and "Out of the League."
Example: If Shai Gilgeous-Alexander was both an All-NBA 1st team member and an All-Star in the 2022 season, his outcome for that season would be "Elite," as that is the highest level of success he achieved. Career Outcome: A player's career outcome is determined by considering their highest level of success across all seasons, with a focus on their achievements after the first four seasons.
Example 1: A player with season outcomes of Roster, Rotation, Rotation, Roster, Roster, and then Out of the League, would be classified as "Out of the League" for his career outcome. This is because, after his first four seasons, his highest level of success was only a single "Roster" year, which doesn't qualify him for a higher category.
Example 2: A player with season outcomes starting from Roster and progressing through Rotation, Starter, and reaching All-Star and Elite, but then dropping back to Starter, would have a career outcome of "All-Star." This is because, after his first four seasons, his highest sustained level of success was at the All-Star level.
Example 3: A player who reaches the "Starter" level for several seasons but does not consistently achieve "All-Star" or "Elite" status, will be classified as a "Starter" in their career outcome.
Hypothetical Case of Shai Gilgeous-Alexander: It's noted that Shai Gilgeous-Alexander, in this hypothetical scenario, would have a career outcome of "All-Star" if, after being "Elite" in one season, he later achieves only one more All-Star level but no more Elite status.

- Simple clustering (3-categories): The above shows how to classify into six categories. However, the k-means method shows that three clusters is the best fit for our dataset (explained in a later section). This makes sense because when commentators and fans usually discuss players, they tend not to classify in high detail. The difference between detailed (six categories) are less distinguishable without knowledge of detailed data. Therefore, we also produce a 3-category labeling version based on:

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)

B. Descriptive Analysis of `Training.csv`

Here is our descriptive analysis based on the training dataset.

A. The first figure contains four histograms, namely the distribution of Draft picks, the distribution of Rookie Points per Game, the distribution of points per game over a player's first four year, and finally the total minutes played distribution.

- Draft Pick Distribution: Players are heavily concentrated at the higher draft numbers, especially at the 60th pick, possibly representing a large group of undrafted players assigned the final number for record-keeping purposes.
- Rookie PPG Distribution: Most rookies score below 5 points per game, with a declining frequency of players as average points increase, suggesting high-scoring rookies are relatively uncommon.
- Average 4-Year PPG Distribution: The distribution of average points per game over a player's first four years skews right, indicating that while scoring generally increases with experience, exceptionally high averages are rare.
- Total Minutes Played Distribution: A notable peak at the lower end signifies many players play

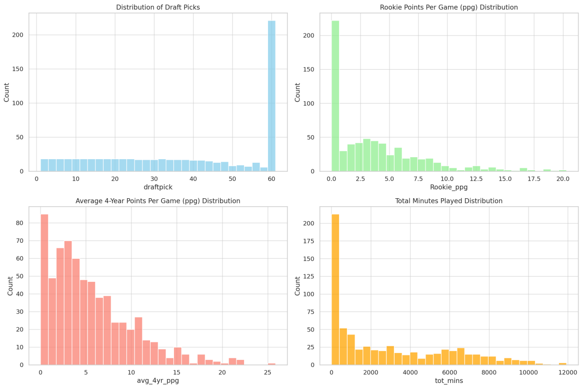

Fig. 1. distribution plots

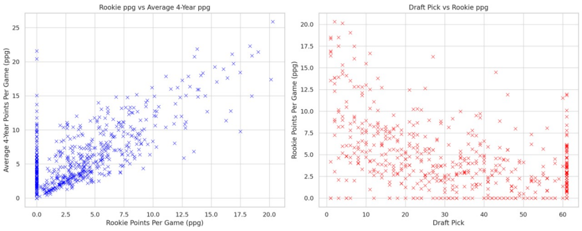

Fig. 2. scatter plots

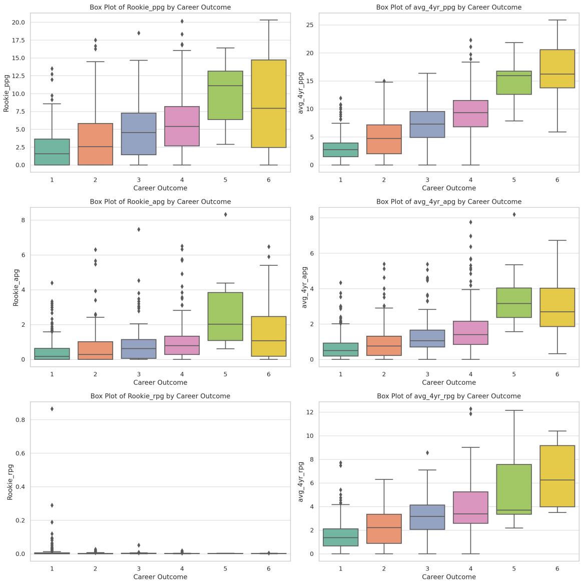

Fig. 3. Enter Caption

only briefly, with fewer players logging extensive minutes on the court, reflecting the rarity of high endurance or consistently selected players.

These histograms collectively offer a visual synopsis of draft selection tendencies and the typical performance arc of NBA players in their initial years.

B. Here is two scatter plots relatively showing the relationship between Rookie ppg and Average 4-Year ppg, and the relationship between Draft Pick and Rookie ppg.

- In the first graph, there's a positive correlation, indicating that players who score more in their rookie year tend to maintain or improve their scoring in the subsequent years.
- The second plot shows a wide spread with no clear trend, suggesting that draft pick position does not strongly predict rookie scoring performance.

C. The following series of box plots illustrating the distribution of NBA player performance metrics by career outcomes, categorized from 1 (least successful) to 6 (most successful).

- Rookie PPG: Median points per game during the rookie year increase with higher career outcomes, with considerable variability within each category.
- Avg. 4-Year PPG: Average points over the first four years rise with improved career outcomes, suggesting a correlation between early-career scoring and long-term success.
- Rookie APG: Initial assists per game show a trend of increasing median values in line with better career outcomes, with a wide spread indicating diverse player performances.
- Avg. 4-Year APG: Consistent with rookie assists, the four-year average assists per game also elevate with higher career outcomes.
- Rookie RPG: Rebounds per game in the rookie year exhibit higher median values for players with more successful careers.
- Avg. 4-Year RPG: The trend of rebounds per game over the first four years mirrors that of points and assists, with higher averages linked to more successful outcomes.

Outliers mark significant deviations from other players in the same career category, implying that individual performance can greatly vary. While, generally, the data suggests a strong association between initial performance metrics and overall career success.

D. In order to understand the relationship between each columns in the training dataset and their career outcome, it is helpful to draw a correlation heatmap. The correlation matrix reveals the following about the relationship:

- `fourth_year_achieved`, `tot_mins` (total minutes played), and `tot_games_start` (total games started) have strong positive correlations with `career_outcome`, suggesting that players with more minutes, starts, and achievements in their fourth year are likely to have better career outcomes.
- `avg_4yr_ppg` (average points per game over 4 years), `tot_games`, and `third_year_achieved` also show strong positive relationships with career outcomes, indicating that consistent performance and achievements in the third year are predictive of a successful career.

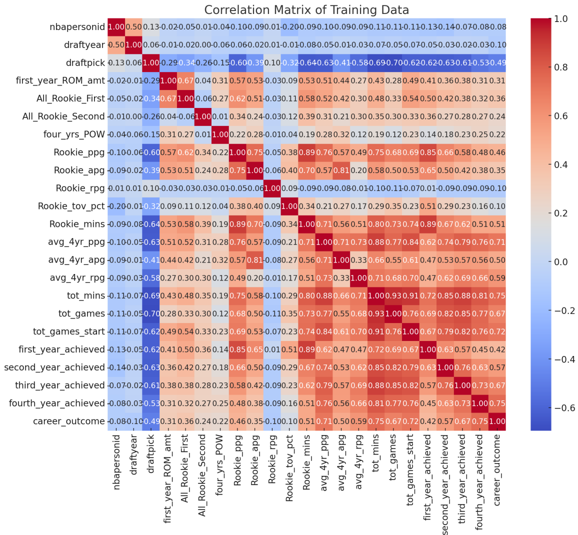

Fig. 4. Enter Caption

- Rookie year statistics like Rookie_mins, Rookie_ppg (rookie points per game), and Rookie_apg (rookie assists per game) have moderate positive correlations with career outcomes, implying that early performance can be indicative of future success.
- Interestingly, draftpick has a significant negative correlation with career_outcome, which might reflect that players selected earlier in the draft tend to have better career outcomes compared to those picked later.
- Metrics like Rookie_rpg (rookie rebounds per game) and Rookie_tov_pct (rookie turnover percentage) have weaker correlations with career outcomes.
- The negative correlation with nbapersonid and draftyear is not informative as these are likely categorical identifiers rather than quantitative variables.

Overall, the data suggests that early career achievements and consistent performance, especially in terms of minutes played and points scored, are good indicators of a player's long-term career success in the NBA. Draft pick position also appears to be a significant predictor, with higher picks generally having better career outcomes.

IV. Hypotheses and Goals

The motivating hypothesis is that there is a causality association between the first four years of performance and awards achieved and the final career outcome of an NBA player. Accordingly, our objective is to predict players' career outcomes based on their first four seasons of performance after being drafted, using player stats and rewards as explanatory variables. This has been theoretically aligns with insights from ESPN reporter Stephen A. Smith[**?**], who emphasized the predictive value of a player's initial four years for future success. We'll use data from players drafted between 2007 and 2015, splitting it 80-20 for training and testing. Our analysis will focus on accuracy and feature evaluation, with the possibility of adjustments. Players drafted in 2016 will serve as a prediction set to estimate their future value, compared with ESPN experts' opinions. Ultimately, we aim to create a Machine Learning system for rookie-year analysis with commercial potential and high data accuracy.

Our data analysis will use three key machine learning algorithms: *K-means++*, *Logistic Regression* and *Random Forest*, each offering unique advantages. We'll perform cross-validation and feature weighting to identify the best predictive features and refine our models using accuracy, precision, and recall metrics. We expect to benchmark these algorithms, aiming for a categorizing accuracy rate above 80%. Critically, a firm causality is expected: the average scoring, assisting, rebounding, stealing, minutes played and started, and awards achieved in the rookie years (independent variables) to be conditional to the career outcome category. There exists not only statistical significance but also theoretical evidence to support the causal association between the conditions and the results.

V. Methods

A. *K-means Clustering*

To start exploring the distance between each unit of observation, Elbow testing with the K-means clustering method is applied. The Elbow Method involves plotting the explained variation as a function of the number of clusters, and the "elbow" of the plot typically represents where the rate of decrease sharply changes, which can be considered a good trade-off between the number of clusters and the within-cluster sum of squares (WCSS). From this graph, we can observe the WCSS decreases as the number of clusters increases. The rate of decrease is sharp up to 3 clusters, and beyond this point, the rate of decrease starts to slow down, indicating that adding more clusters does not significantly improve the fitting of the model. Therefore, the optimal number of clusters is 3 in this case with a silhouette score of 0.45.

Before implementing the clustering method, relevant columns were selected, standardized, and PCA reduced. The results are displayed as following:

B. *Logistic Regression*

In this study, we first employed logistic regression, a widely used statistical method in predictive analytics, particularly for classification problems. Our objective was to predict the career outcomes of basketball players based on various features indicative of their performance and potential. Logistic regression was chosen due to its effectiveness in handling binary and categorical outcome variables, making it suitable for classifying players into distinct career outcome categories.

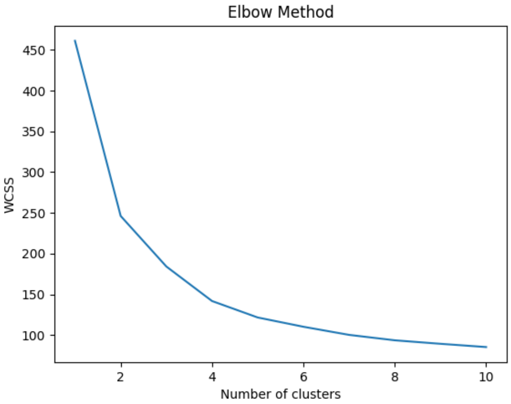

Fig. 5. Elbow Test

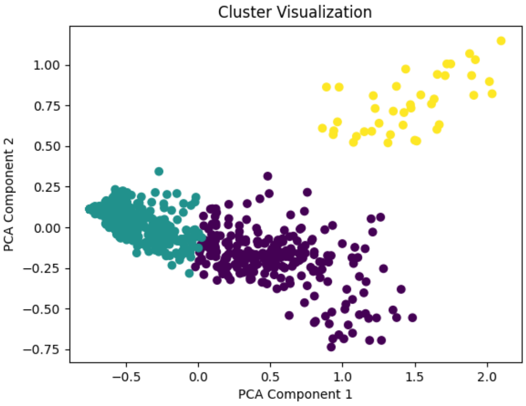

Fig. 6. PCA and K-means Clustering

Logistic regression was chosen for its simplicity, interpretability, and efficiency in handling categorical data. Its ability to provide probabilities for each class makes it an excellent tool for understanding the likelihood of different career outcomes. Furthermore, logistic regression's performance in scenarios with a moderate number of features and data points, as in our case, is generally robust and reliable.

The dataset comprised various features, including `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`, and `tot_games_start`. These features were selected for their significant relevance and potential impact on a player's career trajectory. Initially, career outcomes were categorized into six labels: `Elite`, `All-Star`, `Starter`, `Rotation`, `Roster`, and `Out of the League`. To achieve a balanced dataset and simplify the model's interpretability, feature engineering was employed to consolidate these six labels into three main categories. The `career_outcome` variable was thus transformed into a simpler categorical variable, `simple_label`. The new labeling system grouped `Elite` and `All-Star` into one category, combined `Starter`, `Rotation`, and `Roster` into another, and retained `Out of the League` as a separate category. This reduction in the number of labels was a critical step to decrease complexity and enhance the predictive model's accuracy and interpretability.

The data was split into training and testing sets, with 70% used for training and 30% for testing, ensuring a balanced approach to model training and evaluation. We trained two logistic regression models: one predicting the original 'career_outcome' and the other using the simplified 'simple_label'. The models were iterated for 1000 cycles to ensure convergence.

C. Random Forest

The Random Forest classifier, a robust and versatile machine learning algorithm, was utilized for its proficiency in handling complex classification tasks. This method is particularly effective in managing large datasets with multiple features, making it an excellent choice for our study.

Our dataset contained a mix of numerical and categorical features, including `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`, and `tot_games_start`. Feature engineering involved the application of one-hot encoding to categorical features, thus converting them into a format suitable for modeling. This process enhanced the model's ability to interpret categorical data accurately.

Additionally, a unique 'personal_score' was created by summing standardized numerical features, providing a composite metric indicative of a player's overall performance and potential. This engineered feature was then combined with the original numeric features to form a comprehensive feature set. Recognizing the potential issue of class imbalance in our dataset, we employed Synthetic Minority Over-sampling Technique (SMOTE). This technique artificially generates new instances of the minority class, thus balancing the dataset and improving the classifier's ability to learn from underrepresented classes.

Each model was evaluated on a test set, with the primary metric being accuracy. Additionally, confusion matrices and classification reports were generated for each model, providing detailed insights into their performance across different classes. These reports included precision, recall, and F1-scores, which are critical for assessing the models' ability to handle class imbalances.

VI. Result

A. Logistic Regression

The results of your logistic regression model with six labels (ranging from 'Out of the League' to 'Elite') provide several insights into its performance and areas for improvement. The model correctly predicts the career outcomes 58% of the time. While this is better than random guessing, there is room for improvement. This accuracy might be limited by the complexity of predicting six distinct classes, especially with a dataset that might have imbalances among these classes. The model struggles significantly with label 5 (*All-Star*) and label 6 (*Elite*), indicated by low precision and recall. This could be due to fewer instances of these classes in the dataset or the features not capturing the nuances required to identify top-performing players.

TABLE V
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

In this case, we simplified the target variable `career_outcome` from six categories to three. This simplification involves combining *Roster*, *Rotation*, and *Starter* into one category (labeled 2), *All-Star* and *Elite* into another (labeled 3), and keeping *Out of the League* as is (labeled 1). The new classification report is shown in Table VI.

TABLE VI
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

The simplification of categories in Report (2) has contributed to the improved accuracy. The model might have found distinguishing between broader groups easier by combining similar categories (such as All-Star and Elite into a single class). Despite improvements, both reports indicate ongoing challenges with the highest tier (All-Star/Elite), which might require further data collection, feature engineering, or alternative modeling approaches to enhance predictive accuracy.

Having analyzed the logistic regression model's performance, particularly its limitations in accurately predicting higher-tier career outcomes, we shifted our focus toward a more complex model: the Random Forest classifier. This transition was motivated by the need to address the logistic regression model's shortcomings and explore the potential of a more robust, ensemble-based approach in handling the complexities of our dataset.

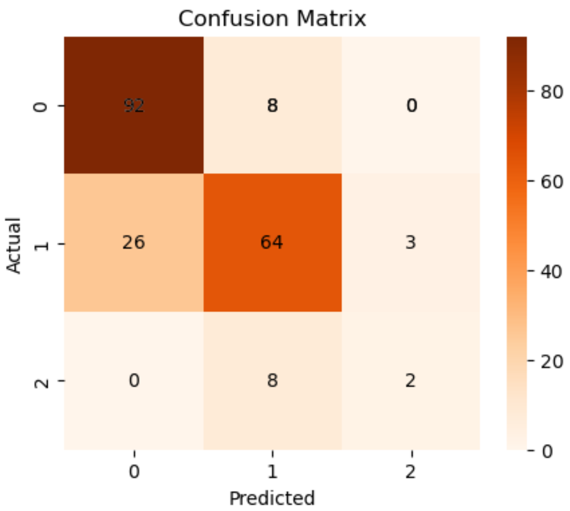

Fig. 7. Enter Caption

B. Random Forest

The implementation of the Random Forest classifier, further enhanced with SMOTE [2] for addressing class imbalances, marked a significant turning point in our analysis. Achieving an overall accuracy of 83%, this model demonstrated a substantial improvement compared to the logistic regression models. This underscores the effectiveness of combining advanced feature engineering, class balancing techniques, and the robustness of the Random Forest algorithm in our predictive modeling endeavor [1].

The model exhibited high precision (88%) and good recall (82%) for this class, reflected in an F1-score of 0.85. This indicates a strong capability in accurately identifying players who are 'Out of the League'. For Class 2, the model also showed impressive results with a precision of 82% and a recall of 85%, leading to an F1-score of 0.83. This signifies effective prediction capabilities for this combined category. However, Class 3 presented more challenges, showing moderate precision (38%) and higher recall (50%), resulting in an F1-score of 0.43. Although the model identified these players more frequently than chance, it also incurred a higher number of false positives.

The confusion matrix provided further insights: most predictions for Class 1 were accurate, with minimal instances being misclassified into Class 2. Class 2 had a high rate of correct predictions, though some misclassifications into Classes 1 and 3 were observed. Despite being the smallest group, Class 3 displayed significant misclassifications. Nonetheless, the model's

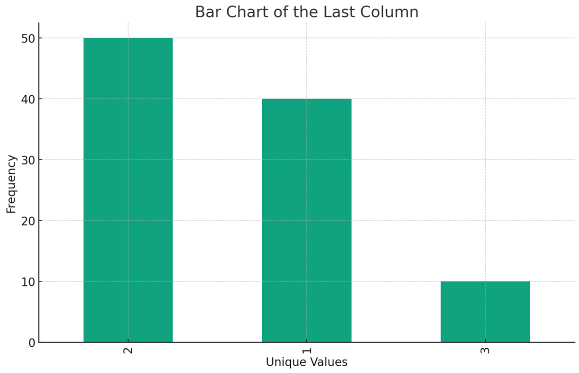

Fig. 9. bar chart of success level

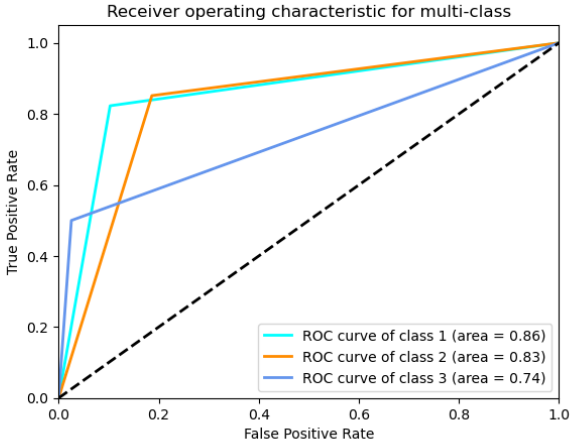

Fig. 8. Enter Caption

ability to correctly identify half of these instances is noteworthy. The ROC line plot, as depicted in Fig. 6, offers a graphical representation that complements our earlier analysis. The commendable AUC values for Classes 1 and 2 (0.86 and 0.83) signify the model's strong discriminatory power in these categories. This high level of performance is visually represented in the ROC plot, where the curves for these classes rise steeply towards the upper left corner, indicating a favorable balance between true positive and false positive rates. Conversely, the curve for Class 3, with an AUC of 0.74, rises less steeply, reflecting the model's relative struggles in accurately distinguishing the 'Combined All-Star, Elite' category from others. This graphical representation aligns with the challenges identified in the confusion matrix, where Class 3 showed a considerable number of false positives, underlining the complexities inherent in predicting this less frequent and more nuanced category. The ROC plot, thus, not only visually validates the numerical metrics we've analyzed but also provides an intuitive understanding of the model's varying performance across different classes.

C. Prediction

The model with the highest performance was applied to the 2018 class players, who have just finished their four rookie-year performances. Their classification results were captured in Fig 7. With a meticulous comparison between our results and renowned commentators' opinion made, there has been a few controversies.

Young superstars (those who have been voted as All-Stars), like Luca Doncic and Trae Young, were all successfully classified as top-tier by our model. Marvin Bagley and Kevin Huerter have been criticized by contemporary reporters like Steven A. Smith, Doris Burke, and Danny Smith [3],[10]. They believe that Bagley and Huerter are not a "superstar" despite their great performance in their rookie years. Jalen Brunson, who has been greatly appraised by the media[6], has been classified as middle tier by our model. Jalen Brunson shows an average score sheet among his peers in his rookie years, but his effort and humbleness have brought him great success just recently.

To summarize, the prediction of 2018, a "mock scouting" process, shares similar views with contemporary experts and commentators. This illustrates the marketable property of our studies.

VII. Conclusion

This study embarked on an ambitious journey to predict the career trajectories of NBA players, leveraging a data-driven approach that combines various machine-learning techniques. Our analysis, pivoting on the rich dataset of player performance and awards, has yielded noteworthy insights into the predictive power of early career statistics in forecasting long-term player outcomes.

Our initial approach with logistic regression revealed the complexities inherent in predicting a player's career trajectory. While moderately successful, the model struggled with the finer nuances of the six-category system. Simplifying these categories into three broader groups improved accuracy, underscoring the importance of data interpretability and model simplicity. The transition to a Random Forest classifier, enhanced with SMOTE to address class imbalances, significantly improved predictive accuracy. This model, with an 83% accuracy rate, demonstrated the effectiveness of ensemble methods in handling complex datasets and the importance of addressing class imbalance in predictive.

The in-depth analysis of model performance across different classes revealed distinct challenges. While the model excelled in predicting outcomes for more common player categories, it faced difficulties in accurately classifying players in the rare 'All-Star' or 'Elite' categories. This finding echoes a common challenge in sports analytics and machine learning at large: effectively predicting rare but significant events.

The findings from this study have significant implications for NBA teams in their scouting and player evaluation processes. By leveraging such predictive models, teams can make more informed decisions, especially under the constraints of wage caps.

While our models have demonstrated high accuracy, there is room for further refinement. Future studies could explore the inclusion of additional variables, such as player injuries or team dynamics, which might enhance predictive accuracy. Moreover, extending the analysis to other sports or leagues could further validate the robustness of these methods.

A critical area for future research lies in improving predictions for rare categories. This might involve exploring alternative machine learning algorithms, more sophisticated feature engineering, or even integrating unstructured data like player interviews or expert opinions.

In conclusion, this study contributes significantly to the field of sports analytics by demonstrating the potential of machine learning in predicting the career outcomes of professional athletes. While challenges remain, particularly in predicting top-tier outcomes, the progress made lays a promising foundation for future research and practical applications in the realm of sports analytics and beyond.

REFERENCES

[1] Using random forest to learn imbalanced data. Technical Report 666, Department of Statistics, Berkeley, July 2004.

[2] J. Brownlee. Smote for imbalanced classification with python, Mar. 2021.

[3] A. Chaudhary. 'huerter put them on a flight like an ig model,' mark jones from espn made a wild statement during the live game, Jan. 2023.

[4] Y. N. Chi. A mixed model for performance-based classification of nba players: Performance-based classification of nba players, 2021. Accessed: December 12, 2023.

[5] B. Ergül. Classification of nba league teams using discriminant and logistic regression analyses, 2014. Accessed: December 12, 2023.

[6] A. Kini. Knicks: Jalen brunson's epic 4-word response to 3-point celebration question, Apr. 2023.

[7] G. R. Levine. *All-NBA Team Voting Patterns: Using Classification Models to Identify How and Why Players Are Nominated.* PhD thesis, [The University Name], 2019. Accessed: [Your Access Date Here].

[8] N. H. Nguyen, D. T. N. Nguyen, B. Ma, and J. Hu. The application of machine learning and deep learning in sport: Predicting nba players' performance and popularity. *Journal of Information and Telecommunication*, 6(2):217–235, Sept. 2021.

[9] Wwu. Github - wwu2020/knn-nba-fa: Source code for 'classification of nba salaries through player statistics' - sagb, 2020.

[10] A. Yerocostas. Season review: Marvin bagley iii, June 2021.