Predictive Analysis of NBA Draft Outcomes Using Machine Learning Techniques

A Comprehensive Study Using NCAA Basketball Player Data from 2009-2021

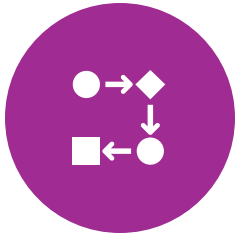
By: Stephen Marsella, Shih-Fan Liu, Edna Maldonado, Arya Bibhukalyan, Megha Goyal

Agenda

1. OVERVIEW OF OUR DATASET

2. DATA PREPROCESSING

3. MODEL EVALUATION AND SELECTION



4. MODEL TUNING, OPERATION, AND RESULTS

5. CONCLUSION AND AREAS FOR IMPROVEMENT

Introduction

- Objective: Harness machine learning to predict NBA draft picks.
- Scope: Analysis of a decade's NCAA player data for draft outcome patterns.
- Impact: Aid stakeholders with actionable sports analytics insights.

Overview of our dataset

- Data Source: Aggregated NCAA records (2009-2021).
- Volume: 61,061 player-season records.
- Key Attributes:
  - Player Demographics (e.g., age, height)
  - Game Statistics (e.g., points scored, assists)
  - Performance Metrics (e.g., offensive/defensive ratings)
  - Efficiency Scores (e.g., field goal percentage)
  - Team and Conference Information
- Exclusion Criteria: Omitted 2022 data to maintain dataset consistency.

Data Preprocessing – Addressing missing data

- Missing Value Treatment:
  - Columns with over 4,000 missing entries removed.
  - Missing 'blk' and 'pts' values replaced with column means.
- Impact:
  - Reduced feature set for enhanced model efficiency.
  - Preserved statistical integrity for key performance indicators.

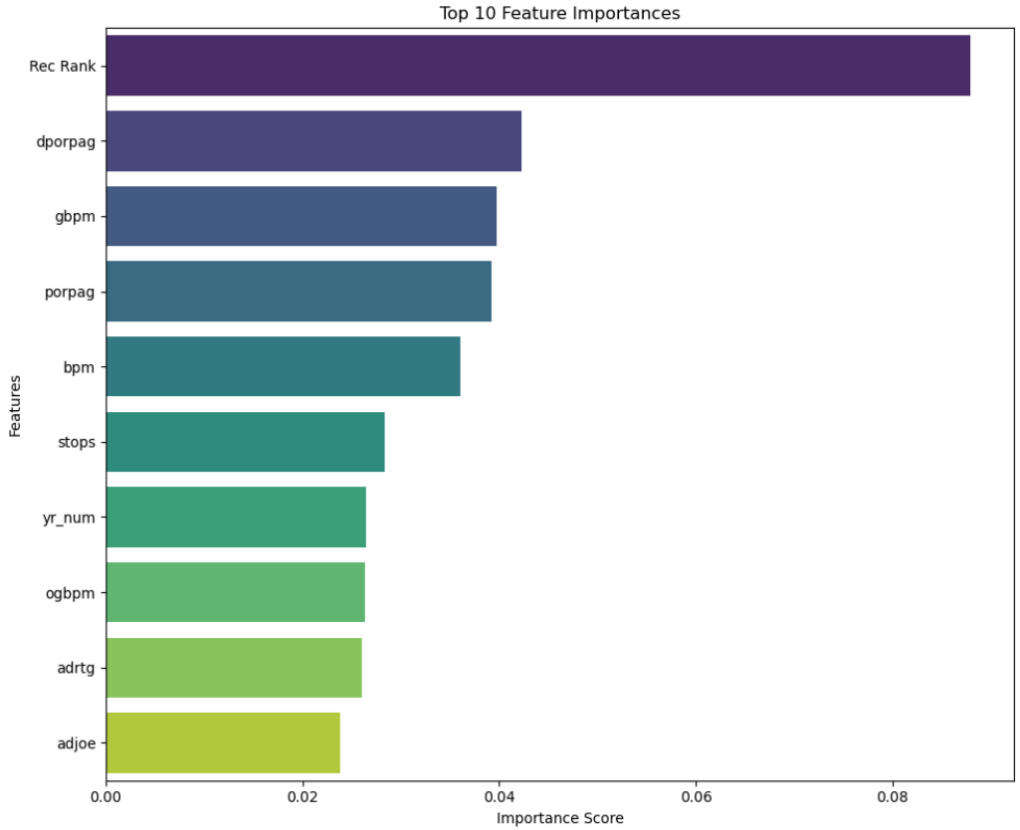
Rec Rank	69.751560
ast/tov	7.741439
rimmade	10.365045
rimmade+rimmiss	10.365045
midmade	10.365045
midmade+midmiss	10.365045
rimmade/(rimmade+rimmiss)	16.557213
midmade/(midmade+midmiss)	17.007583
dunksmade	10.365045
dunksmiss+dunksmade	10.365045
dunksmade/(dunksmade+dunksmiss)	54.879547
pick	97.649891
drtg	0.073697
adrtg	0.073697
dporpag	0.073697
stops	0.073697
bpm	0.073697
obpm	0.073697
dbpm	0.073697
gbpm	0.073697
mp	0.062233
ogbpm	0.073697
dgbpm	0.073697
oreb	0.062233
dreb	0.062233
treb	0.062233
ast	0.062233
stl	0.062233
blk	0.062233
pts	0.062233
Unnamed: 64	7.671018
Unnamed: 65	0.073697

Data Preprocessing - Correlation Analysis

- Correlation Focus: Four performance metrics correlated with 'Drafted' status.
  - 'ortg' (Offensive Rating): Correlation coefficient 'X'
  - 'usg' (Usage Rate): Correlation coefficient 'Y'
  - 'min_per' (Minutes per Game): Correlation coefficient 'Z'
  - 'GP' (Games Played): Correlation coefficient 'A'
- Significance: 'ortg' and 'usg' emerge as top predictors for drafting.

	GP	Min_per	Ortg	usg
count	61061.000000	61061.00000	61061.000000	61061.000000
mean	22.797760	37.12839	75.821091	18.126341
std	10.166805	28.05805	21.210093	6.253742
min	1.000000	0.00000	0.000000	0.000000
25%	15.000000	9.30000	75.821091	14.500000
50%	27.000000	35.60000	75.821091	18.100000
75%	31.000000	62.00000	88.900000	21.800000
max	41.000000	98.00000	100.000000	50.000000

Data Preprocessing - Feature Importance Insights



- Feature Significance via Machine Learning:
  - Random Forest and XGBoost utilized to evaluate feature significance across an initial broader set of features.
- 22 Key Features:
  - Selected using a weighted sum method, focusing on the most impactful predictors.
  - Feature Categories Include:
    - Performance Metrics: Offensive Rating (ortg), Usage Rate (usg).
    - Statistical Measures: Minutes per Game (min_per), Games Played (GP).
    - Efficiency Scores: Effective FG% (eFG), True Shooting % (TS_per).
    - Physical Attributes: Height, weight.
    - Team Dynamics: Player roles, conference strength
- Outcome:
  - Refined feature set enhances model's predictive accuracy and focuses on crucial draft determinants.

Data Preprocessing - Feature Engineering and Transformations

Binary Variable Creation:

- Engineered 'Drafted' variable from 'pick' data to indicate draft status:
  - Drafted = 1 (selected)
  - Not Drafted = 0 (not selected)

Categorical Transformation:

- One-hot encoding applied to:
  - 'Yr' (Year of play)
  - 'conf' (Conference)
- Transforms categorical data into binary vectors for machine learning analysis.

Feature Standardization:

- Implemented z-score normalization:
  - Standardizes data by adjusting for mean and scaling to unit variance.
  - Ensures each feature contributes equally to model predictions.

Model Selection Criteria

- Objective: Identify the best machine learning models for predicting NBA draft outcomes.
- Criteria for Model Evaluation:
  - Accuracy: Measures the overall correctness of the model.
  - Recall: Focuses on the model's ability to correctly identify all drafted players.
  - F1-Score: Harmonic mean of precision and recall, balancing both metrics.
- Models Considered:
  - Random Forest
  - XGBoost

Random Forest Model Evaluation

- Implementation: Utilized an ensemble of decision trees to manage data complexity and variance.
- Feature Handling: Automatic feature selection through ensemble averaging.
- Performance Metrics (Initial Test):
  - Accuracy: 93.98%
  - Recall for drafted players: 32%
  - F1-Score for drafted players: 44%

XGBoost Model Evaluation

- Implementation: Applied gradient boosting techniques for scalable and efficient predictions.
- Feature Handling: Handles missing values and supports various custom optimization objectives.
- Performance Metrics (Initial Test):
  - Accuracy: 94.92%
  - Recall for drafted players: 48%
  - F1-Score for drafted players: 59%

Model Evaluation & Selection – Key Findings/Decisions

- Comparative Insights:
  - XGBoost slightly outperforms Random Forest in accuracy, recall, and F1-score.
  - Both models show high overall accuracy but struggle with recall for drafted players.
- Selection Justification:
  - XGBoost's ability to handle class imbalance and provide higher recall makes it preferable for predicting drafted players.
  - Future tuning and adaptations will focus on improving recall and precision, particularly for the drafted player category.

Overview of Model Tuning, Operation, and Results

- Purpose: Enhance model performance by optimizing key parameters.
- Tuning Techniques Used:
  - RandomizedSearchCV for systematic exploration of parameter space.
  - Focus on hyperparameters such as tree depth, learning rate, and the number of trees.
- Goal: Achieve better precision and recall, especially for drafted players.

Model Tuning – Addressing a data imbalance

Challenge:

- Significant class imbalance with drafted players making up only 2.3% of the dataset, influencing model predictions.

Impact:

- Models biased towards predicting the majority class (not drafted), reducing effectiveness in identifying potential NBA players.

Strategies for Balance:

- Use stratified sampling
- Randomly select 15% of the non-drafted records, combined with all the drafted records

Results:

- Keep the drafted vs non-drafted ratio in the data set at approximately 1:2

Tuning the Random Forest Model

Parameter Adjustments:

- Number of Trees: Increased to improve model stability and accuracy.
- Maximum Depth: Limited to prevent overfitting.

Results After Tuning:

- Improved Precision for drafted players from 72% to 75%.
- Recall for drafted players enhanced slightly, indicating better identification of true positives.

Tuning the XGBoost Model

Parameter Adjustments:

- Eta: 0.1771313
- Max Depth: 4
- Min Child Weight: 5.9497414
- Subsample: 0.6436925

Results After Tuning:

- The combination of these factors resulted in an ideal score of 0.9119, demonstrating the efficiency of our tuning method.

Comparative Results and Final Model Selection

Post-Tuning Performance Comparison:

- XGBoost shows superior performance in both precision and recall post-tuning.
- Random Forest remains robust but slightly less effective in identifying drafted players.

Decision:

- XGBoost selected as the primary model due to its enhanced ability to predict drafted players accurately.
- Plan to implement additional strategies to further improve recall and reduce false negatives.

Conclusion – Summary of Findings

Key Outcomes:

The XGBoost model demonstrated superior accuracy and recall, particularly for drafted players.

Identified significant predictors of draft outcomes: 'ortg' (offensive rating) and 'usg' (usage rate).

Model Achievements:

XGBoost achieved an accuracy of 94.92% and a recall rate of 55% for drafted players post-tuning.

Effectively handled class imbalance and optimized for predictive precision.

Conclusion - Areas of Improvement

Feature

Further refine feature selection to enhance model accuracy and relevance.

Explore additional data sources to enrich the dataset and provide new insights.

Class

Continue to address class imbalance through advanced sampling techniques or cost-sensitive learning.

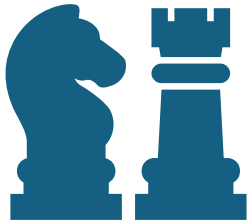
Implement ensemble methods that better predict minority class outcomes.

Conclusion: Future Research & Development

Next Steps:

Integrate more granular data, such as player injury history and psychometric assessments, to improve model robustness.

Apply deep learning techniques to explore non-linear relationships and complex patterns.



Long-Term Goals:

Develop a real-time predictive model that can be used by scouts and team managers during live games and drafts.

Collaborate with sports analysts and data scientists to continuously update and validate the model against new draft seasons.

References

- AIML.com. (2023, October 3). What are the advantages and disadvantages of Decision Tree model? Retrieved from https://aiml.com/what-are-the-advantages-and-disadvantages-of-using-a-decision-tree/
- AIML.com. (2023, October 3). Advantages and disadvantages of Random Forest. Retrieved from https://aiml.com/what-are-the-advantages-and-disadvantages-of-random-forest/
- Berkeley Statistics. (n.d.). Using Random Forest to Learn Imbalanced Data. Retrieved from https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf
- GeeksforGeeks. (2023, February 6). XGBoost. Retrieved from https://www.geeksforgeeks.org/xgboost/
- Hindawi. (2021). Hyperparameter Optimization & Tuning for Machine Learning (ML). Retrieved from https://www.hindawi.com/journals/complexity/2021/6676297/
- IEEE Xplore. (n.d.). Tuning the hyper-parameters of an estimator — scikit-learn 1.4.2 documentation. Retrieved from https://ieeexplore.ieee.org/abstract/document/7460114
- IEEE Xplore. (n.d.). Hyperparameter tuning for Machine learning models. Retrieved from https://ieeexplore.ieee.org/abstract/document/8943156
- Peargin, G. (2019). Random Forest for Prediction with Unbalanced Data. Retrieved from https://opus4.kobv.de/opus4-hm/frontdoor/deliver/index/docId/305/file/Peargin_2019_MA.pdf
- SpringerLink. (2024). Hyperparameter Optimization & Tuning for Machine Learning Models. Retrieved from https://link.springer.com/article/10.1007/s10115-024-02092-9
- Towards Data Science. (n.d.). Hyperparameter tuning for Machine learning models | by Jaswanth Badvelu. Retrieved from Towards Data Science website.
- Verma, A. (2023, February 4). Decision Trees: Advantages, Disadvantages, and Applications. DEV Community. Retrieved from https://dev.to/anurag629/decision-trees-advantages-disadvantages-and-applications-25b2

Thank you!

We welcome any questions or further discussion on the research, methodologies, or findings