Improving Predictions of Which NCAA Players Are Most Likely to Be Drafted into NBA

Table of Contents

1. Introduction

1.1 Problem Statement

The NBA (National Basketball Association) Draft is the primary pathway for NCAA (National Collegiate Athletic Association) players to enter professional basketball league. Despite media exposure and scouting reports, draft decisions still rely heavily on subjective evaluations rather than data-driven insights. Some talented college players go undrafted due to inefficient scouting processes that fail to integrate performance metrics, physical attributes, and historical draft trends effectively. This project aims to build a predictive model that uses machine learning to enhance draft forecasting accuracy and scouting efficiency. By analyzing NCAA player stats, this model will provide a systematic approach to identify draft-worthy talents.

1.2 Background and Context

The NCAA serves as the primary feeder system for the NBA, with hundreds of college players declaring for the draft each year. However, only 60 players are selected in the two-round NBA Draft. Traditional scouting relies on real-time game observation, game film analysis, and team assessments. With the increasing adoption of analytics in professional sports, NBA teams have started integrating statistical models into their draft processes. Currently, many existing models only rely on one certain attribute like size or shooting efficiency rather than generating an overall evaluation of to what extend the player is positioned in this year’s draft. This leaves room for improvement in draft predictions.

1.3 Importance of the Problem

Improving draft predictions benefits multiple stakeholders. NBA teams can optimize their draft picks, reduce selection risks, and improve long-term roster success. Scouts can refine evaluation methods, making talent identification more efficient. NCAA players and coaches gain insights into which skills are valued, allowing for better development. Similar predictive approaches can also enhance fairness and decision-making in recruitment, education, and other talent-focused fields.

1.4 Project Goals

This project aims to enhance NBA draft predictions by using machine learning techniques and integrating diverse data sources, including NCAA performance metrics and players' physical attributes. The goal is to develop a data-driven model that improves scouting accuracy, identifies high-value prospects, and adapts to the evolving strategic demands of NBA teams.

2. Related Work

2.1 Summary of Existing Research and Solutions

Several studies have examined factors influencing NBA draft selections. A Springer study on draft selection (*Kong, X. F., Mu; Zhang, Hui.,n,d*) identified key

predictors, including physical attributes, college statistics, and team preferences. Another study on draft trend (ThinkingBeyondtheScore, n,d) utilized machine learning to categorize players based on their likelihood of success. Traditional models have used performance metrics like points per game, efficiency ratings, and player impact estimates. Although these methods offer valuable insights, they often fail to adapt to the teams' strategic demands for player evaluation.

2.2 Identification of Gaps

Current draft models emphasize traditional stats and physical traits but neglect the overall impact of these two attributes in talent evaluation. They also fail to incorporate the teams' strategic demands, meaning that the talents are evaluated without considering the teams actual needs. The Springer study highlights player attributes but lacks predictive modeling, while the K-mean analysis (*ThinkingBeyondtheScore, n.d*) clusters players without considering too much historical data. Additionally, existing models may fail to predict draft choices due to the evolving team needs. This project will integrate machine learning and available NCAA player datasets to ensure relevance in forecasting draft outcomes based on modern NBA strategies and the teams' needs.

2.3 Novelty of Project

This project introduces a novel approach by combining traditional performance metrics and physical data (e.g., height, height). By applying machine learning

clustering techniques like k-mean clustering, this project will offer precise draft predictions distinct from prior models. This model could enhance talent evaluation processes, providing NBA teams, scouts, and players with detailed, data-driven insights to guide draft decisions.

3. Business Model

3.1 Business/Application Analysis

This project uses data science to improve NBA draft predictions by analyzing player performance and physical attributes. By applying machine learning to NCAA datasets, teams can make informed scouting decisions based on their own needs. The project reduces selection risks, improves accuracy, and increases efficiency in talent selection. Primary stakeholders include NBA teams, scouts, and players. They all benefit from the enhanced draft insights. Introducing analytics into recruitment optimizes draft strategies, while NCAA athletes can understand what attributes improve their chances to be drafted in the evolving trends. To ensure the model's relevance in recruitment, this model also takes the most advantage of the 4 Vs in big data, which are moderate Volume (multi-season datasets), Velocity (real-time updates if possible), Variety (base stats, biometrics), and Veracity (accuracy through validation across history datasets).

3.2 Benefits and Value Creation

This project enhances scouting efficiency by providing data-driven insights on prospects, helping teams and scouts select players aligned with their needs. Players benefit by the identified valuable attributes. College coaches can use these insights to better prepare young players for their future career. By integrating historical data and trained models, this project provides a new recruitment strategy, providing long-term value to all stakeholders.

3.3 Primary Stakeholders

Primary stakeholders include NBA teams, scouts, college coaching staff, and NCAA players. Teams benefit from improved draft accuracy. They are more likely to select players that fit roster needs and their long-term plans. Scouts get access to advanced analytical tools and refine their accuracy of talent evaluations. NCAA players receive valuable insights into what skills and attributes NBA teams prioritize. Based on the insights from the model, the NCAA players can refine their training and career decisions. College coaching staff can also align player development with NBA preferences, ensuring their athletes are prepared for the professional league.

4. Characterizing and Analyzing data

4.1 Data Sources and Characteristics

Source 1: NBA Daft History from the official NBA stats site *(National Basketball Association, n.d.)*, including draft results, player names, draft year, and pick numbers.

Source 2: NCAA Team rosters and player stats, scraped from Sports Reference *(Sports Reference, n.d.)*, including players physical attributes and their performance in NCAA.

Both data sets are publicly available and scraped from the web pages by Python. Data from Source 1 is cleaned into NCAA players only. In Source 2, the full data set is divided by years, meaning that a player who played more than 2 years exists in all the data sets of the year when he played.

In terms of 4V's, both data sources have moderate Volume, which contains multiple years of draft results and player stats. Source 1 shows low but necessary Variety while Source 2 shows high Variety which contains categorical data (college, position), numeric data (physical attributes, base stats) and the year of draft. The Velocity is low to medium because both datasets are updated annually. In the current project, this Velocity is sufficient to build the model since the draft is an annual event. There are other live-updating sources like ESPN live. However, in considering the Veracity, those live sources' data qualities are lower than the selected sources which are corrected and uploaded annually.

To collect the data, Python scripts are used for web scraping and initial processing (removing missing values, correct college names). Then, R is used to create a label indicating whether the player is drafted after comparing the 2 data sources. The cleaned data set could be downloaded from my public GitHub repository (https://github.com/kkgob/FIT5145_assignmnet_3_data.git).

4.2 Data Analysis Techniques and Statistical Methods

This project applies a hybrid approach combining dimensionality reduction, clustering, and supervised learning to identify NCAA players that most likely to be drafted.

First, five core features were selected: Height_cm, Weight_kg, PTS, REB, and AST. All variables were normalized to ensure uniform contribution. Then, Principal Component Analysis (PCA) was applied to extract the top components.

To classify player types, K-Means clustering was performed on the PCA-reduced data. The optimal number of clusters was chosen based on both the Elbow Method (WSS) and Silhouette Scores, balancing cluster cohesion with separation.

Two logistic regression models were fitted. The first used only the cluster label (KMeans_Cluster) to predict draft status. The second, improved model combined cluster labels with original stats (Height_cm, Weight_kg, PTS, REB, AST) to simulate a realistic evaluation pipeline like real-world scouting.

This pipeline integrates unsupervised and supervised techniques to deliver interpretable, data-driven draft predictions aligned with actual scouting outcomes.

4.3 Demonstration

The train data is from the 2022 draft results and 2022 NCAA rosters. The test data is from 2021 draft results and 2021 NCAA rosters.

In the training data, before applying PCA, all numerical features were normalized to ensure equal influence across scales. Principal Component Analysis was then used to reduce the data into two primary components that retained most of the variance.

```
=== PCA Feature Contributions ===
                PC1         PC2         PC3
Height_cm -0.3082625 -0.58780594  0.02608806
Weight_kg -0.3070865 -0.57306514  0.45571629
PTS       -0.5587380  0.28403232 -0.23513014
REB       -0.5984789  0.02189142 -0.47502934
AST       -0.3745757  0.49490093  0.71463488
PC1 is most influenced by: REB (-0.5985)
PC2 is most influenced by: Height_cm (-0.5878)
PC3 is most influenced by: AST (0.7146)
```

Figure 1.1 PCA Feature Contributions

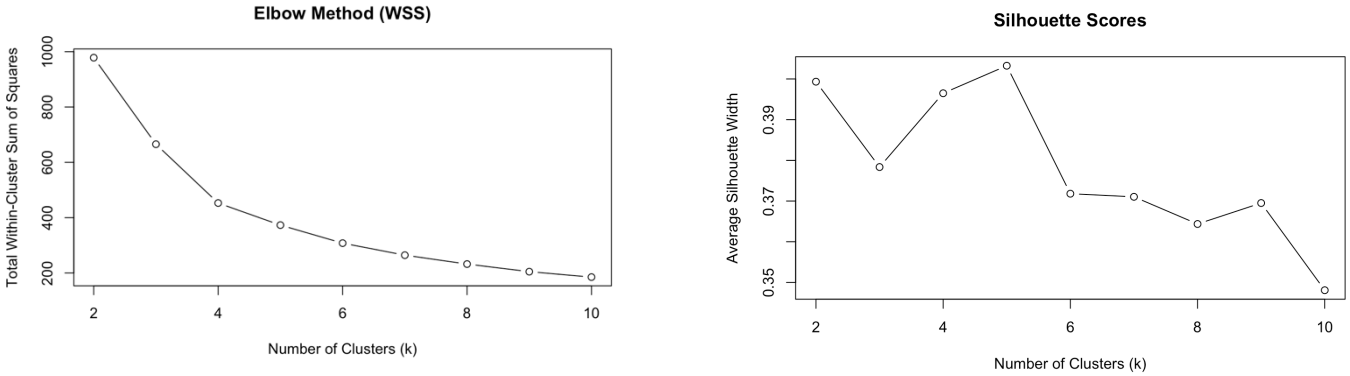

Figure 2.1 Elbow Method and Silhouette Scores

Based on Figure 1.1, PC1 is heavily influenced by rebounding and scoring ability (REB and PTS), reflecting a player's overall statistical impact. PC2 is dominated by

height, weight, and assist, separating players based on physical size and playmaking tendencies. These components capture the main differentiators among NCAA players and serve as the input space for clustering and classification tasks.

K-Means clustering was applied to the PCA-transformed data to group players into distinct types based on their underlying statistical profiles. To determine the optimal number of clusters, both the Elbow Method and Silhouette Scores were evaluated. The Elbow plot showed a clear inflection at k = 5, indicating a balance between model complexity and within-cluster variance. Similarly, the Silhouette analysis peaked around k = 5, suggesting good separation and cohesion among clusters.

From the Review of the clustering results in the R script, k = 5 revealed meaningful player groupings that reflect diverse roles on the court—such as elite scoring bigs, versatile wings, and low-impact bench players. These clusters are used as categorical inputs in the later models. This classification captures practical differences in player types, especially relevant in a positionless basketball context where traditional roles no longer apply to modern basketball.

In Model 1, the logistic regression model used only the KMeans_Cluster label to predict whether a player was drafted.

```r
#logistic regression using cluster membership only
logit_model_cluster <- glm(Drafted ~ KMeans_Cluster, data = pca_data, family = "binomial")
```

Figure 3.1 Model 1 (Cluster_only)

```
Call:
glm(formula = Drafted ~ KMeans_Cluster, family = "binomial",
    data = pca_data)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.8177     0.7148  -5.341 9.26e-08 ***
KMeans_Cluster2   2.1130     0.8116   2.603  0.00923 **
KMeans_Cluster3   1.6205     0.8184   1.980  0.04769 *
KMeans_Cluster4   3.3928     0.7799   4.350 1.36e-05 ***
KMeans_Cluster5 -15.7484   981.7028  -0.016  0.98720
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 228.63  on 377  degrees of freedom
Residual deviance: 167.19  on 373  degrees of freedom
AIC: 177.19

Number of Fisher Scoring iterations: 18

AUC (Cluster Only): 0.8570024
Optimal Threshold (Cluster Only): 0.06075269
         Actual
Predicted   0   1
        0 211   2
        1 133  32
Model Accuracy (Cluster Only, Optimal Threshold): 0.6428571
```

Figure 3.2 Model 1 training performance

While the model achieved a decent AUC (0.857), its overall performance was limited. Although some clusters—specifically Cluster 2 ($p < 0.01$), Cluster 3 ($p < 0.05$), and Cluster 4 ($p < 0.001$)—were statistically significant, the model still produced a very low optimal threshold (~0.06) and a high number of false positives (133). This suggests it was overly confident in predicting players as drafted, even when they were not. With an overall accuracy of only 64.3%, the model lacked precision and calibration. This confirmed that cluster membership alone was not sufficient for reliable draft prediction and needed to be combined with player-level statistics for improved performance.

In Model 2, KMeans_Cluster labels are considered with core player stats (Height_cm, Weight_kg, PTS, REB, AST).

```
#fit logistic regression model using cluster + player stats
logit_model_full <- glm(Drafted ~ Height_cm + Weight_kg + PTS + REB + AST + KMeans_Cluster,
                        data = data_with_cluster, family = binomial(link="logit"))
```

Figure 3.3 Model 2 (Cluster and player stats)

```
AUC (Cluster + Stats): 0.9209986
Optimal Threshold (Cluster + Stats): 0.1250014
         Actual
Predicted   0   1
        0 294   5
        1  50  29
Model Accuracy (Cluster + Stats, Optimal Threshold): 0.8544974
Precision: 0.367
Recall (Sensitivity): 0.853
F1-Score: 0.513

Call:
glm(formula = Drafted ~ Height_cm + Weight_kg + PTS + REB + AST +
    KMeans_Cluster, family = binomial(link = "logit"), data = data_with_cluster)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -16.12961    9.02150  -1.788  0.07379 .
Height_cm          0.09840    0.04747   2.073  0.03816 *
Weight_kg         -0.09061    0.03463  -2.617  0.00888 **
PTS                0.21935    0.07945   2.761  0.00576 **
REB                0.23615    0.17890   1.320  0.18684
AST                0.29577    0.25734   1.149  0.25041
KMeans_Cluster2   -1.33604    1.46341  -0.913  0.36126
KMeans_Cluster3   -0.80525    1.05693  -0.762  0.44614
KMeans_Cluster4   -0.46383    1.21355  -0.382  0.70231
KMeans_Cluster5  -16.39502 1550.55196  -0.011  0.99156
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 228.63  on 377  degrees of freedom
Residual deviance: 138.29  on 368  degrees of freedom
AIC: 158.29

Number of Fisher Scoring iterations: 19
```

Figure 3.4 Model 2 training performance

Based on Figure 3.4, Model 2 shows significantly improved performance. The AUC increased to 0.921, indicating excellent discrimination between drafted and undrafted players. The model achieved an accuracy of 85.4%, with high recall (0.853)—correctly identifying 29 out of 34 drafted players—while maintaining an F1 score of 0.513. Although precision (0.367) was moderate due to some false positives (50), this is acceptable in scouting contexts where missing talent is costlier than over-selecting. The optimal threshold of 0.125 reflects class imbalance. Statistically significant predictors included Height_cm ($p < 0.05$), Weight_kg ($p < 0.01$), and PTS ($p < 0.01$), confirming that physical and scoring attributes were key drivers in draft prediction. While cluster terms provided role-based structure, individual player metrics added most of the predictive power.

To test Model 2, the test data was used under the same selected features and same optimal threshold in classification. With the projected PCAs, the cluster Labels were assigned to the test data set.

```
         Actual
Predicted   0   1
        0 420   5
        1  68  33
Test Set Accuracy (Cluster + Stats, Optimal Threshold): 0.861
Test Set Precision: 0.327
Test Set Recall (Sensitivity): 0.868
Test Set F1-Score: 0.475

Call:
glm(formula = Drafted ~ Height_cm + Weight_kg + PTS + REB + AST +
    KMeans_Cluster, family = binomial(link = "logit"), data = data_with_cluster)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -16.12961    9.02150  -1.788  0.07379 .
Height_cm          0.09840    0.04747   2.073  0.03816 *
Weight_kg         -0.09061    0.03463  -2.617  0.00888 **
PTS                0.21935    0.07945   2.761  0.00576 **
REB                0.23615    0.17890   1.320  0.18684
AST                0.29577    0.25734   1.149  0.25041
KMeans_Cluster2   -1.33604    1.46341  -0.913  0.36126
KMeans_Cluster3   -0.80525    1.05693  -0.762  0.44614
KMeans_Cluster4   -0.46383    1.21355  -0.382  0.70231
KMeans_Cluster5  -16.39502 1550.55196  -0.011  0.99156
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 228.63  on 377  degrees of freedom
Residual deviance: 138.29  on 368  degrees of freedom
AIC: 158.29

Number of Fisher Scoring iterations: 19
```

Figure 4.1 Model 2 test performance in 2021 data set

In the 2021 test set, Model 2 maintained strong generalization performance. The model achieved an accuracy of 86.1%, closely matching its training accuracy, and a high recall (0.868), correctly identifying 33 out of 38 drafted players. Although precision (0.327) declined slightly from training, this is expected in imbalanced datasets and acceptable in scouting, where missing talent is riskier than false positives. The F1 score of 0.475 reflects this trade-off. Key predictors remained

consistent—Height_cm ($p < 0.05$), Weight_kg ($p < 0.01$), and PTS ($p < 0.01$) were statistically significant, confirming their robust impact across seasons. Therefore, the model demonstrated strong out-of-sample reliability, suggesting practical applicability for future draft prediction tasks.

5. Standards for Data Science Process, Data Governance, and Management

5.1 Standards in Data Science Process

This project follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, which provides a structured approach through stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This standard ensures that the project remains iterative, traceable, and reproducible.

For data processing, a documented pipeline was followed, including feature standardization, PCA for dimensionality reduction, and rigorous model evaluation using metrics such as accuracy, AUC, and F1-score. Clustering and supervised models were benchmarked using both training and test sets, supporting generalizability and alignment with the model governance principle of fairness and transparency.

5.2 Data Governance and Management

The project uses two public datasets:

NBA Draft History, containing player names, draft years, and pick numbers.

NCAA Player Stats, scraped from Sports Reference, including physical attributes and NCAA performance metrics.

Although neither dataset contains personally identifiable information, responsible data governance practices were applied. Data use complies with the Australian Privacy Act 1988, which mandates reasonable steps for securing personal data.

Key governance actions included:

Access Control: Cleaned datasets were stored on a public GitHub repository (link) with version control. Only essential features were stored to minimize exposure.

Lifecycle Management: Data was cleaned, analyzed, and transformed within a documented pipeline, ensuring traceability and integrity across stages.

Ethical Use: No commercial use or redistribution of the scraped NCAA data was performed. The analysis remained within the scope of academic purposes.

Anonymity and Security: All player identifiers were retained solely for clustering and were not exposed in any sensitive outputs. No financial data or biometric identifiers were involved.

References

Kong, X. F., Mu; Zhang, Hui. (2024). Factors Affecting NBA Player Draft Selection: An Analysis Based on a Generalized Linear Mixed Model Springer, Singapore. https://link.springer.com/chapter/10.1007/978-981-97-2898-5_14

ThinkingBeyondtheScore. (2024). NBA Draft Patterns Through Clustering Analysis. https://medium.com/%40thinkingbeyondthescore/nba-draft-patterns-through-clustering-analysis-541e89fcc02b

Sports Reference. (n.d.). *College basketball schools index*. Sports-Reference.com. https://www.sports-reference.com/cbb/schools/#NCAAM_schools

National Basketball Association. (n.d.). *NBA draft history*. NBA.com. https://www.nba.com/stats/draft/history