IMPROVING PREDICTIONS OF WHICH NCAA PLAYERS ARE MOST LIKELY TO BE DRAFTED INTO NBA

Data Science in Sports

Sucheng Li

Introduction

Background:

- The NBA (National Basketball Association) Draft is the primary pathway for NCAA (National Collegiate Athletic Association) players to enter professional basketball league.
- Hundreds of college players declaring for the draft each year. However, only 60 players are selected in the two-round NBA Draft.
- Despite media exposure and scouting reports, draft decisions still rely heavily on subjective evaluations rather than data-driven insights.
- This project aims to build a predictive model that uses machine learning to enhance draft forecasting accuracy and scouting efficiency.

Value of this project and Business Model:

- This project uses data science to improve NBA draft predictions by analyzing player performance, physical attributes, and historical draft trends. The project reduces selection risks, improves accuracy, and increases efficiency in talent selection.
- Similar predictive approaches can also enhance fairness and decision-making in recruitment, education, and other talent-focused fields.

Data Sources and Characteristics

- Data 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.
- https://www.nba.com/stats/draft/history
- Moderate Volume, Low but Necessary Variety, Low Velocity (updated annually), High Veracity

- Data 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.
- https://www.sports-reference.com/cbb/schools/#NCAAM_schools
- Moderate Volume, High Variety, Low Velocity (updated annually), High Veracity

Data Analysis Techniques and Statistical Methods

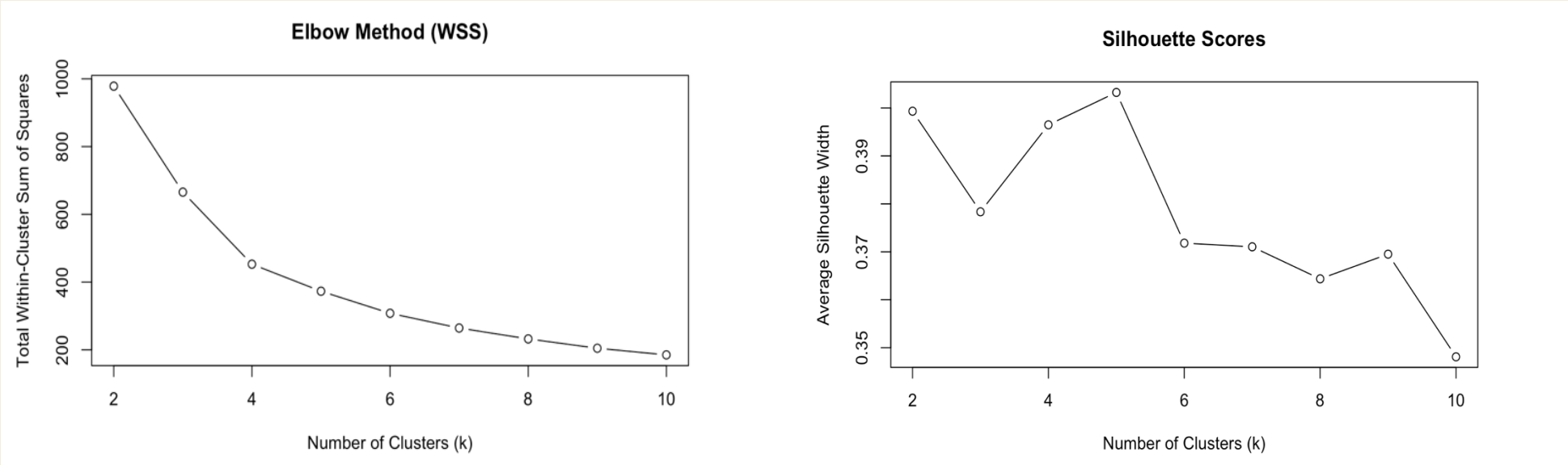
- Feature Normalization (Height_cm, Weight_kg, PTS, REB, and AST).
- Principal Component Analysis (PCA) was applied to extract the top 2 components.
- K-Means clustering was performed on the PCA-reduced data.
- Logistic regression models were trained based on 2022 draft.
- Logistic regression models were tested based on 2021 draft.

Demonstration - PCA

- 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.

```
=== 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)
```

Demonstration – K-Mean Clustering



- 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.

Demonstration - K=5

- This classification is based on the players' actual contribution on basketball court and their physical attributes.
- This is a Novel classification method that captures the current trend of positionless basketball.

Cluster 1: Low-Impact and Bench Players
- Drafted Rate: 0.0%
- Player Count: 116
- Profile: Players with limited minutes, production, or visibility. Often bench players or non-rotational contributors, with minimal statistical impact.
- Drafted Examples: None

Cluster 2: Playmaking Guards and Slashing Scorers
- Drafted Rate: ~6.9%
- Player Count: 87
- Profile: Young, energetic guards with upside. These players often filled combo guard or secondary ball-handler roles and showed flashes of scoring or playmaking.
- Drafted Examples:
- Jaden Ivey (Purdue): 17.3 PTS, 4.9 REB, 3.1 AST
- Andrew Nembhard (Gonzaga): 11.8 PTS, 5.8 AST
- Malaki Branham (Ohio State): 13.7 PTS, 3.6 REB
- Blake Wesley (Notre Dame): 14.4 PTS, 3.7 REB
- Ryan Rollins (Toledo): 18.9 PTS, 6.0 REB

Cluster 3: Elite Scoring Bigs and Offensive Anchors
- Drafted Rate: ~36.2%
- Player Count: 58
- Profile: High-usage forwards and centers who dominated offensively. They contributed heavily in PTS and REB, often leading their teams.
- Drafted Examples:
- Paolo Banchero (Duke): 17.2 PTS, 7.8 REB
- Chet Holmgren (Gonzaga): 14.1 PTS, 9.9 REB, 3.7 BLK
- Walker Kessler (Auburn): 11.4 PTS, 8.1 REB, 4.6 BLK
- Keegan Murray (Iowa): 23.5 PTS, 8.7 REB
- Mark Williams (Duke): 11.2 PTS, 7.4 REB

Cluster 4: Defensive and Role-Specific Players
- Drafted Rate: Low (~1 player drafted)
- Player Count: 93
- Profile: Primarily defense-first players – centers and forwards with limited offensive contribution but strong rebounding or rim protection. Often situational role players.
- Drafted Examples:
- Jaylin Williams (Arkansas): 10.9 PTS, 9.8 REB, 1.3 STL

Cluster 5: Versatile Wings and Two-Way Forwards
- Drafted Rate: ~14.3%
- Player Count: 77
- Profile: Wings with balanced stat lines across scoring, rebounding, and assists. Known for adaptability and versatility on both ends of the court.
- Drafted Examples:
- AJ Griffin (Duke): 10.4 PTS, 3.9 REB
- Kendall Brown (Baylor): 9.7 PTS, 4.9 REB
- Dalen Terry (Arizona): 8.0 PTS, 4.8 REB, 3.9 AST
- Max Christie (Michigan State): 9.3 PTS, 3.5 REB
- Caleb Houstan (Michigan): 10.1 PTS, 4.0 REB

Demonstration – Logistic Regression Model (Training Performance)

```
#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"))
```

```
Optimal Threshold (Cluster + Stats): 0.1250014
          Actual
Predicted   0   1
        0 294   5
        1  50  29
Model Accuracy (Cluster + Stats, Optimal Threshold): 0.8544974
```

- The optimal threshold of 0.125 reflects class imbalance since only few players were selected.
- 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 (False Negatives) is riskier than over-selecting (False Positives).

Demonstration – Logistic Regression Model (Test Performance)

```
=== True Positives (Correctly Predicted Drafted Players) ===
 [1] "Cade Cunningham"        "Evan Mobley"            "Scottie Barnes"         "Corey Kispert"          "Jalen Suggs"
 [6] "Franz Wagner"           "Isaiah Livers"          "Jared Butler"           "Ziaire Williams"        "James Bouknight"
[11] "Chris Duarte"           "Moses Moody"            "Tre Mann"               "Kai Jones"              "Jalen Johnson"
[16] "Keon Johnson"           "Isaiah Jackson"         "Josh Christopher"       "Quentin Grimes"         "Santi Aldama"
[21] "Jeremiah Robinson-Earl" "Jason Preston"          "Miles McBride"          "JT Thor"                "Ayo Dosunmu"
[26] "Neemias Queta"          "Luka Garza"             "Joe Wieskamp"           "Kessler Edwards"        "Dalano Banton"
[31] "Charles Bassey"         "Sandro Mamukelashvili"  "Aaron Wiggins"

=== False Negatives (Missed Drafted Players) ===
[1] "Davion Mitchell" "Joshua Primo"    "Jericho Sims"    "Jaden Springer"  "Day'Ron Sharpe"
```

```
          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
```

- 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 (False Negatives) is riskier than over-selecting (False Positives).
- Therefore, the model demonstrated strong out-of-sample reliability, suggesting practical applicability for future draft prediction tasks.

Who's Most Likely to Be Drafted?

- Cluster 3: consistently exhibits the highest draft rates, indicating that PTS + REB-heavy bigs and scoring forwards are valued highly in the NBA draft.
- Cluster 2: playmaking guards shows moderate draft rates, suggesting that guards with offensive output and playmaking stats are noticed, but only when combined with efficiency or athletic ceiling.
- Cluster 5: a hybrid zone — drafted players here tend to be well-rounded or NBA-ready wings.

Data Governance and Management

- Data 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.
- Data 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.
- Cleaned Data Set: Cleaned 2021-2022 draft results, Cleaned 2021-2022 NCAA rosters, uploaded to public GitHub repository (https://github.com/kkgob/FIT5145_assignmnet_3_data.git).
- Access Control: Cleaned datasets were stored on a public GitHub repository 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 data. 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.