Using Machine Learning to Predict NBA Draft Outcomes

Explore how to apply machine learning techniques to forecast the future performance of NBA draft prospects.

NBA Draft Prediction Workflow

1 Data Extraction

📥 Connect to SQLite database.

📥 Extract data from relevant tables: player attributes, team salaries, player salaries, draft, draft combine, and game data.

📥 Merge draft and draft combine data tables.

📥 Delete unnecessary columns containing 'set' and 'location'.

2 Data Cleaning

🧼 Use mode (categorical) or mean (numeric) to fill missing values.

🧼 Remove irrelevant or redundant rows and columns.

🧼 Verify data integrity after cleaning.

3 Feature Engineering

🛠️ Select relevant features based on domain knowledge.

🛠️ Create new columns to capture position data.

🛠️ Calculate additional metrics (e.g., BMI).

4 Exploratory Data Analysis (EDA)

🔍 Show summary statistics and distributions of key features.

🔍 Analyze the distribution of the target variable (whether drafted or not).

🔍 Visualize relationships and correlations between features.

5 Data Transformation

🔄 Normalize or scale numeric features.

🔄 Encode categorical features using one-hot encoding.

🔄 Split data into training, validation, and test sets.

6 Model Selection and Training

🧠 Select multiple models (e.g., logistic regression, decision trees, random forests, SVM, KNN, gradient boosting, XGBoost).

🧠 Train models using the training data.

🧠 Evaluate models using metrics like accuracy, precision, recall, F1-score, ROC-AUC, and specificity.

7 Model Evaluation and Interpretation

📊 Compare model performance based on evaluation metrics.

📊 Select the best model based on the required metric (e.g., recall).

📊 Perform feature importance analysis to identify key predictive factors.

8 Model Deployment

🚀 Save the best-performing model.

🚀 Load and use the model for real-time predictions.

9 SHAP Analysis for Model Interpretation

🔍 Generate SHAP values to explain model predictions.

10 Real-World Application

🏀 Collect actual data for new players (e.g., Kevin Durant, Bronny James).

🏀 Perform feature engineering and data standardization.

🏀 Use the saved model and SHAP analysis to predict draft position and explain the results.

Introduction

Project Overview

This project explores how to use machine learning to predict NBA draft outcomes and understand the resulting insights.

Importance of Prediction

Accurately predicting draft results is crucial for teams to make wise decisions and build successful rosters.

Objective

The goal is to use data and machine learning to predict which players will be drafted and how they will perform in the NBA.

Motivation

Building Successful NBA Rosters

Accurately predicting draft picks can help teams assess rookie potential and build long-term competitiveness.

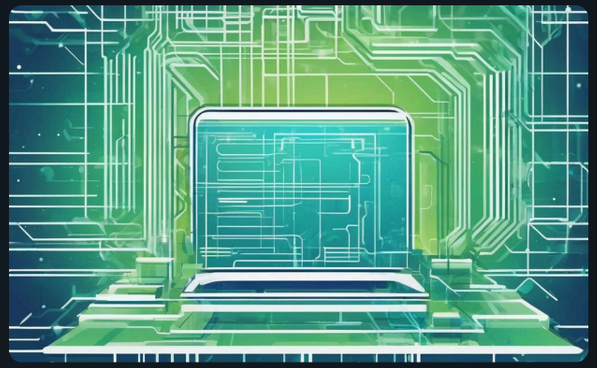
Improving Recruitment Strategies

Data analysis can help more effectively identify promising rookies and improve draft selection success rates.

Personal Interest

As a data analysis enthusiast, I hope to deeply explore the patterns in the NBA draft.

Exploring the Data Journey of the NBA Draft

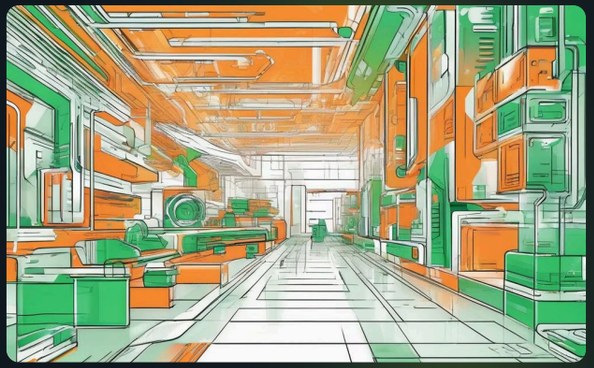


Connect to SQLite Database

Extract player attributes, team salaries, player salaries, draft information, draft combine, and game data from the relevant tables.

Integrate Draft Information

Merge the draft and draft combine tables to consolidate the analysis-required information.



Clean and Preprocess

Clean and preprocess the extracted data to ensure it is in a format suitable for machine learning models.

```python
import sqlite3
import pandas as pd

# 連接到 SQLite 資料庫
conn = sqlite3.connect('/Users/jerrylaivivemachi/Desktop/basketball.sqlite')

# 提取表格數據
player_attributes_df = pd.read_sql('SELECT * FROM Player_Attributes', conn)
team_salary_df = pd.read_sql('SELECT * FROM Team_Salary', conn)
player_salary_df = pd.read_sql('SELECT * FROM Player_Salary', conn)
draft_df = pd.read_sql('SELECT * FROM Draft', conn)
draft_combine_df = pd.read_sql('SELECT * FROM Draft_Combine', conn)
games_df = pd.read_sql('SELECT * FROM Game', conn)

# 合併 Draft 和 Draft_Combine 表格
merged_df = pd.merge(draft_df, draft_combine_df, on='idPlayer', how='right')

# 刪除包含 'set' 和 'location' 字樣的列
columns_to_drop = [col for col in merged_df.columns if 'set' in col or 'location' in col]
merged_df.drop(columns=columns_to_drop, inplace=True)

cols_to_drop = ['heightWShoesInches', 'heightWShoes']
merged_df.drop(cols_to_drop, axis=1, inplace=True)

# 打印合併且整理後的表格
print(merged_df.head())

merged_df.info()
```

```python
import pandas as pd

# 設置顯示選項以顯示所有列和所有行
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)

merged_df.loc[:, ('yearCombine')] = merged_df['yearCombine'].apply(lambda x: x - 1)

merged_df[~(merged_df.isna())][['yearDraft', 'yearCombine', 'namePlayer_y']]

# create a target variable, drafted
merged_df.loc[:, 'drafted'] = (merged_df.yearDraft > 0)

merged_df.drafted.value_counts()

# check that yearDraft
merged_df[merged_df.yearDraft < 2000]

# Drop the mistaken row for Reggie Williams
merged_df = merged_df.drop(index=693, axis=0)

merged_df.info()

# 關閉資料庫連接
conn.close()
```

```python
# Step 2: Data Cleaning

# 處理缺失值的函數，針對數值型和分類型數據分別處理
Codeium: Refactor | Explain | Generate Docstring | ×
def handle_missing_values(df):
    for column in df.columns:
        if df[column].dtype == 'object':
            # 用眾數填充分類型缺失值
            df[column].fillna(df[column].mode()[0], inplace=True)
        else:
            # 用平均值填充數值型缺失值
            df[column].fillna(df[column].mean(), inplace=True)
    return df

# 處理 merged_df 中的缺失值
merged_df = handle_missing_values(merged_df)

# 處理其他數據表中的缺失值
player_attributes_df = handle_missing_values(player_attributes_df)
team_salary_df = handle_missing_values(team_salary_df)
player_salary_df = handle_missing_values(player_salary_df)
games_df = handle_missing_values(games_df)

# 打印處理後的 merged_df 來確認結果
print(merged_df.head())
```

Feature Engineering

Relevant Features

Select key attributes based on domain knowledge

Position Data

Create new columns to capture player position

Additional Indicators

Calculate metrics like BMI to enrich the dataset

As an NBA expert (a basketball fan, that is), I believe these are useful features. The remaining columns are mostly organization names, which have high cardinality and are difficult to use as features.

feature_list = ['idPlayer', 'namePlayer_y', 'yearDraft', 'yearCombine', 'numberPickOverall', 'slugPosition', 'heightWOShoesInches','weightLBS', 'wingspanInches', 'reachStandingInches', 'verticalLeapStandingInches', 'verticalLeapMaxInches', 'repsBenchPress135','timeLaneAgility', 'timeThreeQuarterCourtSprint', 'timeModifiedLaneAgility', 'lengthHandInches', 'widthHandInches', 'pctBodyFat', 'drafted']

Key Observations (Statistical Descriptions)

Feature	Description
Number of Players	1,366 unique players
Player Positions	14 unique positions, with Power Forward (PF) being the most common
Players Drafted	808 players drafted, 586 players undrafted
Draft Years	2000-2020
Combine Years	2000-2020
Height (Without Shoes)	Average 197.1 cm
Weight	Average 98 kg
Wingspan	Average 209.3 cm
Standing Reach	Average 262.7 cm
Standing Vertical Leap	Average 74.2 cm
Max Vertical Leap	Average 87.4 cm
Bench Press (135 lbs)	Average 10.3 reps
Lane Agility	Average 11.4 seconds
3/4 Court Sprint	Average 3.29 seconds
Hand Length	Average 22.1 cm
Hand Width	Average 23.9 cm
Body Fat Percentage	Average 7.6%

Exploring NBA Player Data

1 Missing Values

Only 496 records have data for secondary position, indicating significant missing data

2 Most Common Categories

- Most common player name: Josh Hart (2 times)
- Most common primary position: Power Forward (PF), 376 times
- Most common secondary position: Small Forward (SF), 143 times

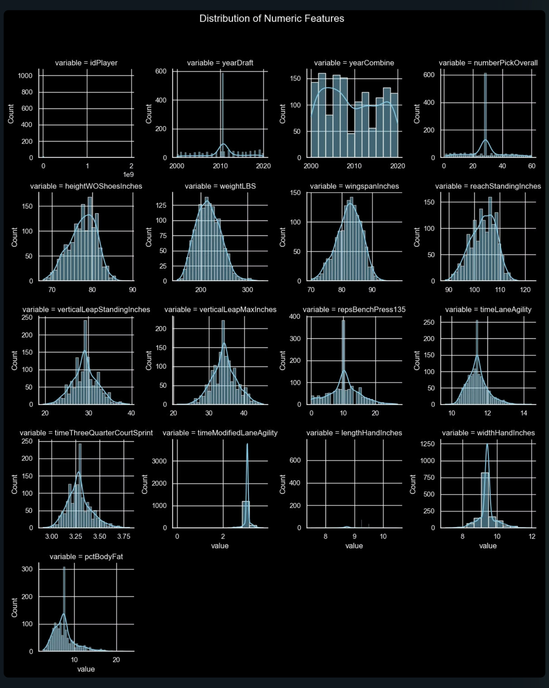
3 Summary

The dataset provides a comprehensive understanding of NBA players' physical attributes and performance, which can aid in scouting, training, and draft decision-making. Addressing the missing data for secondary positions will further improve the utility and reliability of the analysis.

Distribution of Numerical Features

Overview

Histogram analysis of the numerical features in the NBA draft dataset reveals the distribution and trend patterns of various physical and performance indicators for the players. This analysis helps to understand the basic patterns in the data and identify any anomalies or trends.



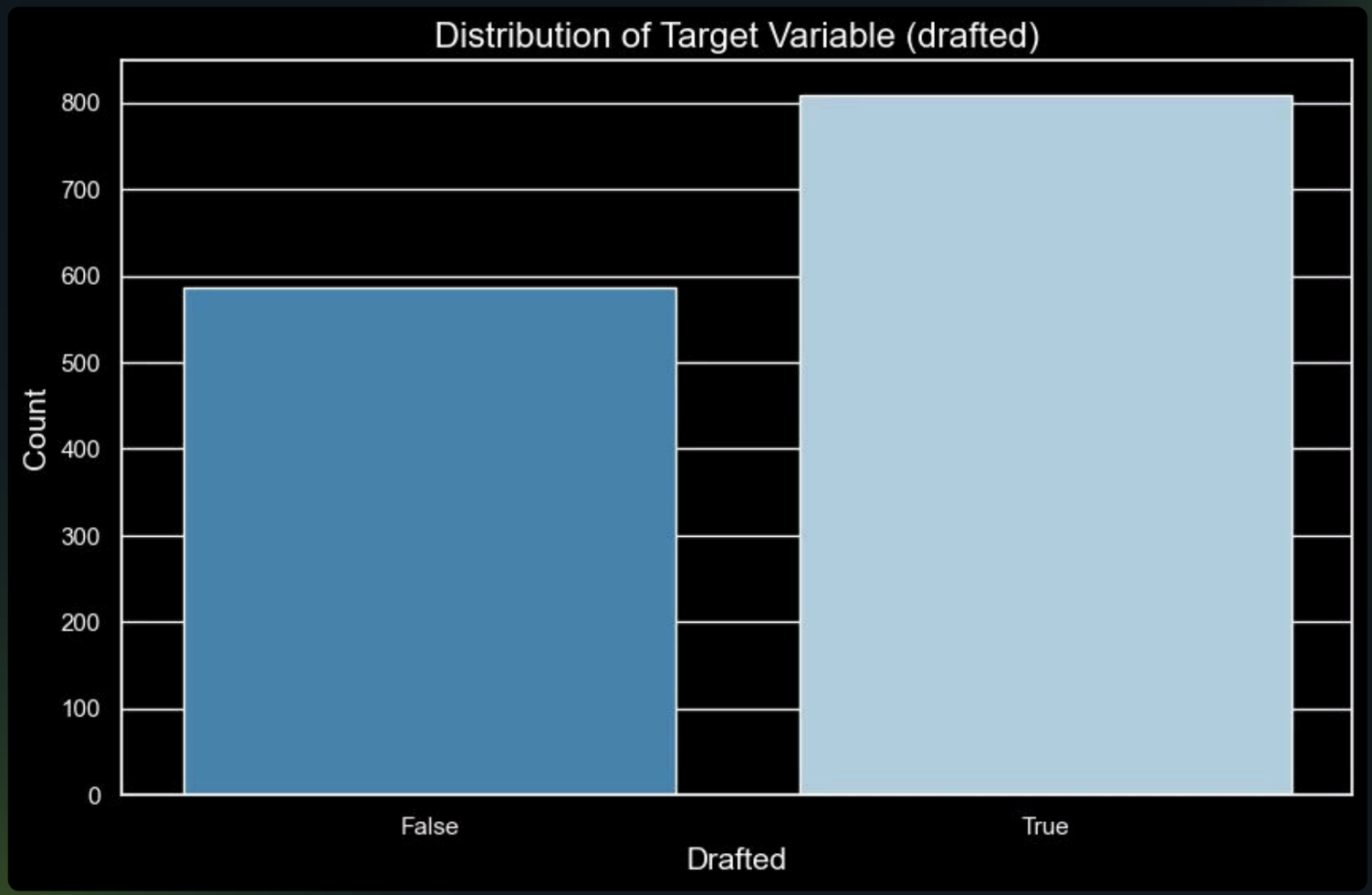
Key Observations

Feature	Distribution	Implications
idPlayer	Highly skewed, with most values concentrated in a specific range	The idPlayer is a unique identifier and does not provide useful information for the analysis
yearDraft	Data shows a large number of players were selected around 2010 and 2015	This indicates the dataset includes players from different draft years, with some years having higher peaks
yearCombine	Similar to yearDraft, with noticeable peaks around 2004 and 2015	The patterns in the combine data match the draft data, reflecting the years when players participated in the combine
numberPickOverall	Higher frequency of lower pick numbers, indicating more players were selected early in the draft	This suggests the dataset includes more high-performing players, which may bias the analysis towards high-performing players
heightWOShoesInches	Roughly normally distributed, centered around 77.6 inches (197.1 cm)	Indicates the typical height of NBA players, with most players within a few inches of the average
weightLBS	Roughly normally distributed, centered around 216.1 lbs (98 kg)	Reflects the average weight of NBA players, providing insights into their typical body types
wingspanInches	Normally distributed, with an average of 82.4 inches (209.3 cm)	Indicates the standard wingspan range of NBA players, which is crucial for evaluating player reach
reachStandingInches	Normally distributed, with an average of 103.4 inches (262.7 cm)	Suggests the typical standing reach of players, which is essential for assessing defensive and offensive capabilities
verticalLeapStandingInches	Roughly normally distributed, centered around 29.2 inches (74.2 cm)	Provides insights into player explosiveness, which is crucial for rebounding and shot blocking abilities
verticalLeapMaxInches	Normally distributed, with an average of 34.4 inches (87.4 cm)	Reflects the maximum vertical leap height, highlighting the players' athleticism
repsBenchPress135	Skewed distribution, with most values concentrated around 10.3 reps	Indicates the upper body strength of players, which is crucial for physical battles in NBA games
timeLaneAgility	Roughly normally distributed, centered around 11.4 seconds	Reflects the agility of players, which is essential for quick movements on the court
timeThreeQuarterCourtSprint	Normally distributed, with an average of 3.29 seconds	Highlights the sprint speed of players, which is crucial for fast breaks and defensive rotations
timeModifiedLaneAgility	Some outliers, with most values concentrated at a specific point	May indicate issues with data collection or specific situations affecting the recorded times
lengthHandInches	Normally distributed, centered around 8.7 inches (22.1 cm)	Reflects the typical hand length, which is related to ball handling and shooting
widthHandInches	Normally distributed, centered around 9.4 inches (23.9 cm)	Suggests the hand width, which affects grip and ball control
pctBodyFat	Skewed distribution, with most values concentrated around 7.6%	Displays the body fat percentage of players, which is important for evaluating physical fitness and condition

Summary

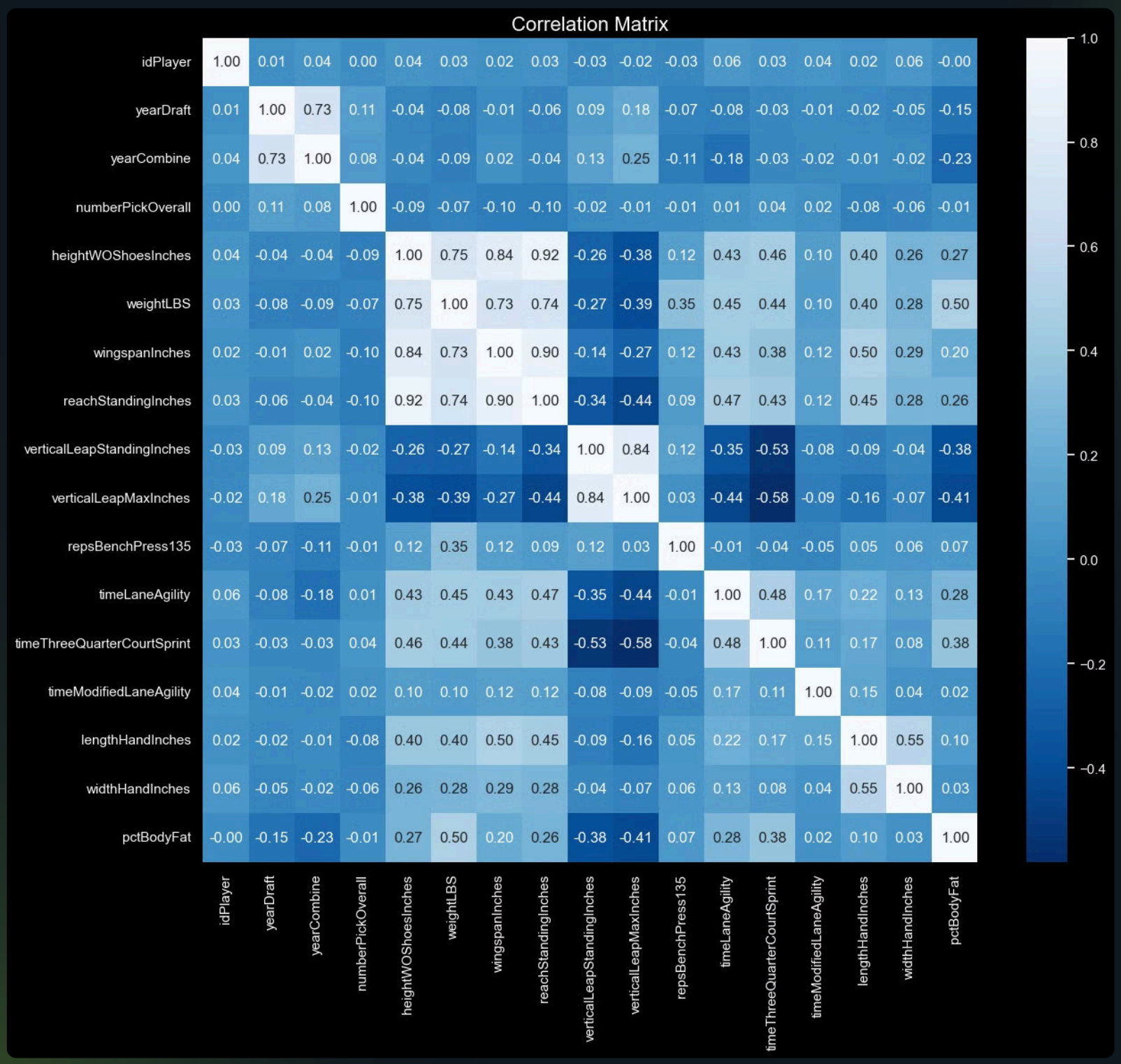
The histograms show that most physical and performance indicators follow a roughly normal distribution, with a few features exhibiting skewness or anomalies. These insights help to understand the typical player profiles in the NBA and identify any outliers or trends in the data. This information is crucial for scouting, training, and predictive modeling in basketball analysis.

Distribution of the Target Variable



The distribution of the target variable `drafted` shows that players who were drafted are slightly more than those who were not. This is helpful for training machine learning models, as it allows the model to learn better and make more accurate predictions. However, to ensure the model performs fairly, we should still try to balance the distribution of the target variable as much as possible.

Correlation Matrix Heatmap

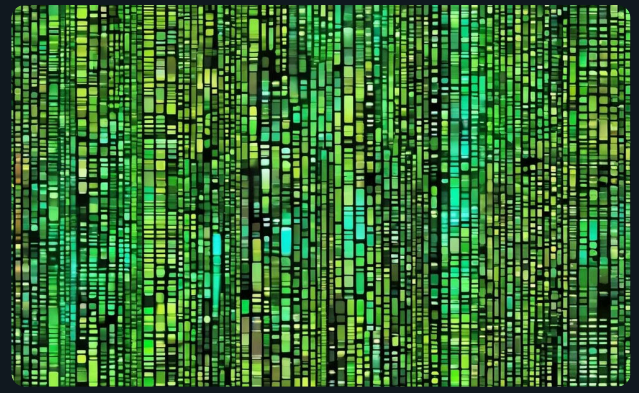


Correlation analysis reveals the important relationships between numerical features in the NBA draft dataset. Strong positive correlations show that height is closely related to standing reach (0.920), wingspan (0.836), and vertical jump (0.839), highlighting the significance of these features for defensive and offensive capabilities. Moderate positive correlations, such as weight with height (0.755) and agility with sprint time (0.485), emphasize the critical nature of body size and athletic ability. Weak negative correlations, like maximum vertical jump with sprint time (-0.580) and body fat with vertical jump (-0.415), indicate the impact of explosiveness and body composition on performance. These insights are crucial for scouting, training, predictive modeling, and player positioning, improving overall team performance and player abilities.

Data Transformation

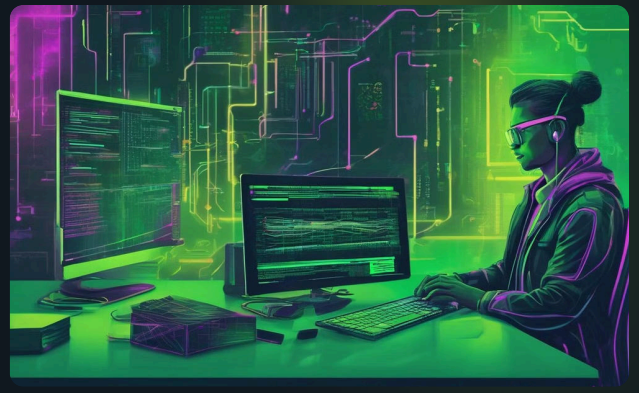
Numerical Feature Normalization

Normalize or scale numerical features to ensure they are all on a similar magnitude.



Categorical Feature Encoding

Use one-hot encoding to transform categorical features into a format that machine learning models can handle.



Data Split

Split the dataset into training, validation, and test sets to properly evaluate model performance.

Model Selection and Training

1

Model Selection

We selected multiple models (such as logistic regression, decision trees, random forests, support vector machines, K-nearest neighbors, gradient boosting, and XGBoost)

2

Model Training

We trained the models using the training data

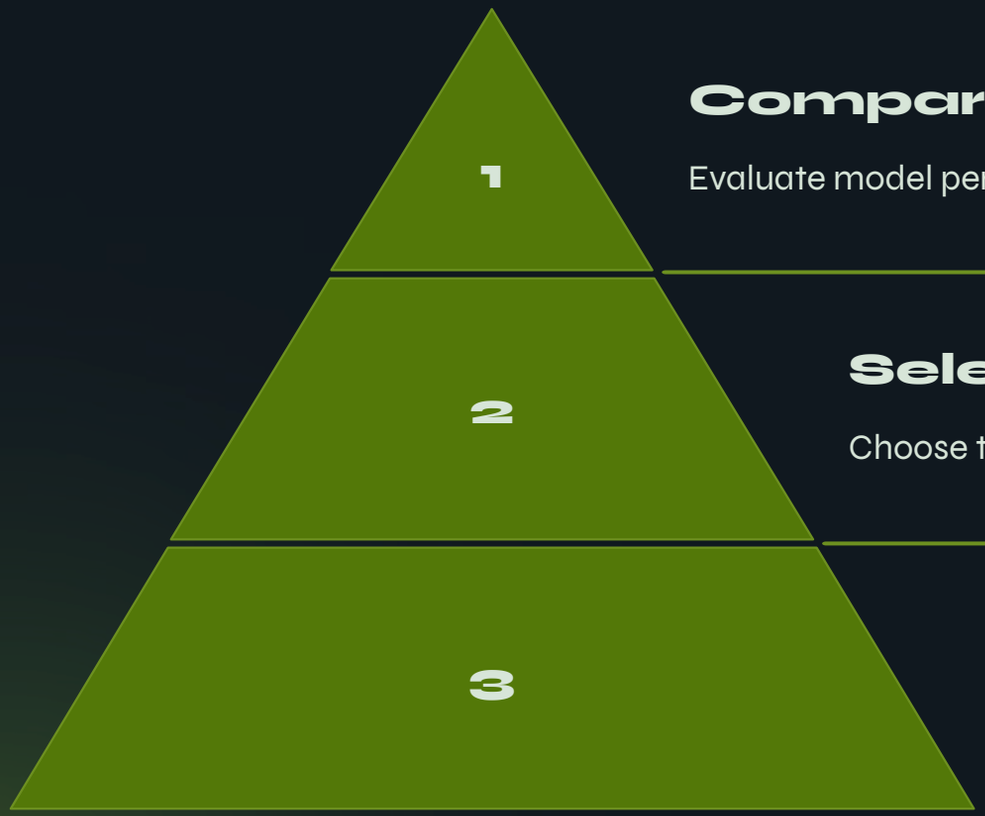
3

Model Evaluation

We evaluated the models using metrics like accuracy, precision, recall, F1-score, ROC-AUC, and specificity

	Accuracy	Precision	Recall	F1 Score	ROC-AUC	Specificity
Logistic Regression	0.637232	0.635841	0.637232	0.636406	0.656439	0.561111
Decision Tree	0.568019	0.560783	0.568019	0.562054	0.550023	0.422222
Random Forest	0.632458	0.626726	0.632458	0.622360	0.661820	0.450000
Support Vector Machine	0.637232	0.634110	0.637232	0.618519	0.663680	0.394444
K-Nearest Neighbors	0.551313	0.537663	0.551313	0.537271	0.589319	0.344444
Gradient Boosting	0.649165	0.644636	0.649165	0.641673	0.676034	0.488889
XGBoost	0.627685	0.621597	0.627685	0.618134	0.657020	0.450000

Model Evaluation and Interpretation



Compare Models

Evaluate model performance using key metrics

Select Best Model

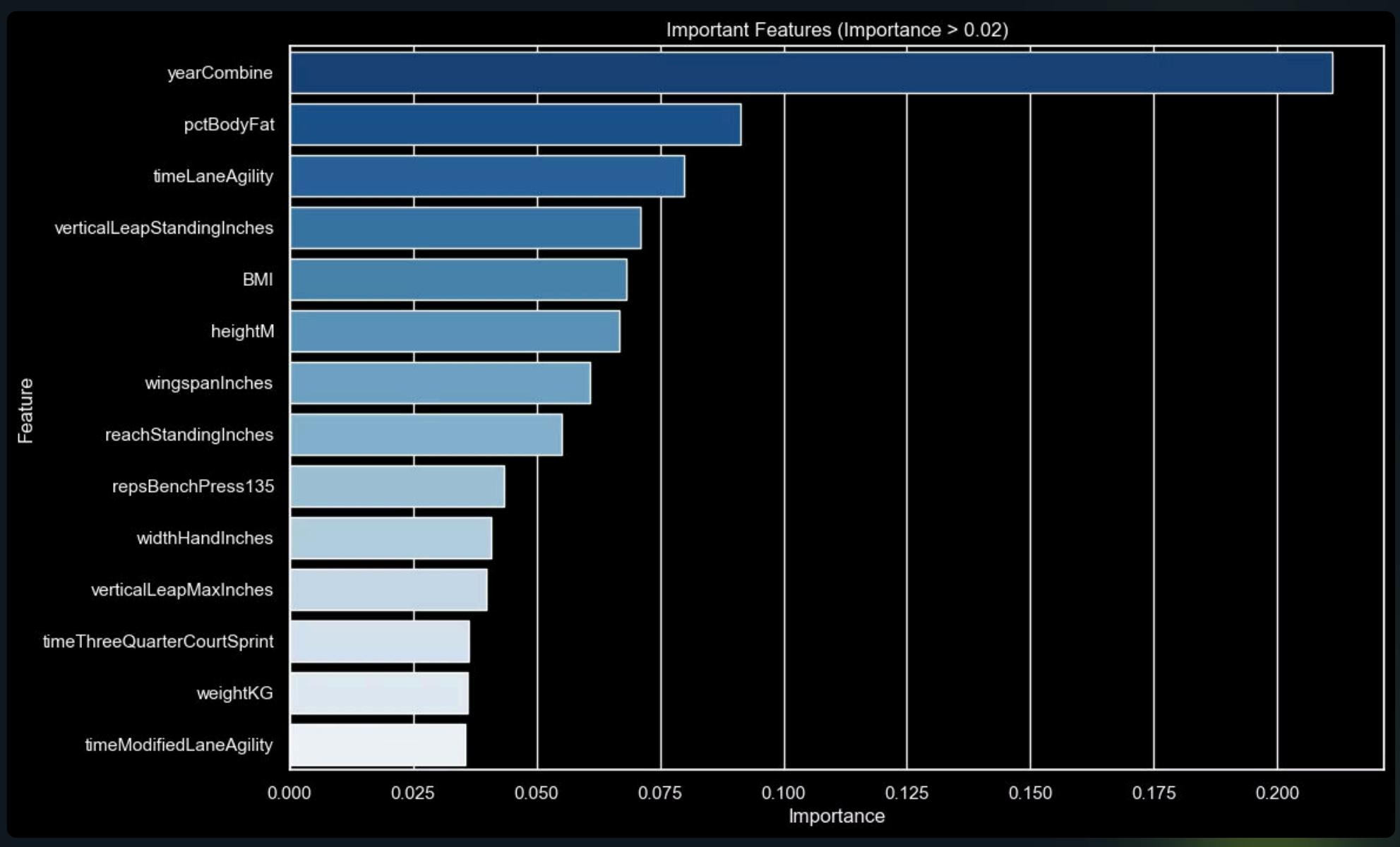
Choose the model with the highest recall (sensitivity)

Analyze Feature Importance

Conduct feature importance analysis

The model evaluation process includes comparing the performance of various models based on key metrics such as accuracy, precision, recall, and F1 score. The model with the highest recall was selected as the best model(Gradient Boosting). Additionally, feature importance analysis was performed to understand which attributes are most important for the model's predictions.

Analyzing Feature Importance



Feature Importance Analysis

Key Attributes Impacting NBA Draft Predictions

- 🏀 **Year of NBA Combine Participation: 0.2126**
  - **Insight**: Recent performance data is highly valuable.
- 🏋️ **Body Fat Percentage: 0.0905**
  - **Insight**: Indicates the player's physical condition.
- 🏃 **Lateral Quickness: 0.0815**
  - **Insight**: Reflects the player's ability to move quickly.
- 🦘 **Vertical Jump: 0.0691**
  - **Insight**: Indicates the player's explosiveness.
- 📊 **BMI (Body Mass Index): 0.0690**
  - **Insight**: Reflects the player's body composition.
- 🧍 **Height: 0.0680**
  - **Insight**: Crucial for both offense and defense.
- 👐 **Wingspan: 0.0593**
  - **Insight**: Important for reach in both offense and defense.
- 📏 **Standing Reach: 0.0535**
  - **Insight**: Key for defensive and offensive actions.
- 💪 **Bench Press Reps: 0.0434**
  - **Insight**: Shows the player's upper body strength.
- ✋ **Hand Width: 0.0412**
  - **Insight**: Indicates ball-handling skills.
- 🏃 **Maximum Vertical Jump: 0.0390**
  - **Insight**: Reflects the player's peak explosiveness.
- 🏃 **3/4 Court Sprint: 0.0363**
  - **Insight**: Measures speed over a short distance.
- ⚖️ **Weight: 0.0360**
  - **Insight**: An indicator of the player's physical build.
- 🏃 **Modified Lane Agility: 0.0355**
  - **Insight**: Indicates the player's agility and endurance.

Application in NBA Teams

These attributes collectively provide NBA teams with a comprehensive evaluation of a player's readiness and potential impact on the court.

Model Optimization

{'Accuracy': 0.6539379474940334, 'Precision': 0.6497070343443696, 'Recall': 0.6539379474940334, 'F1 Score': 0.646548411226029, 'ROC-AUC': 0.6652022315202233, 'Specificity': 0.49444444444444446}

This process focuses on improving the model by emphasizing the key features for predicting NBA draft outcomes. Initially, a gradient boosting model was used to identify feature importance, then features with importance scores greater than 0.02 were selected to retrain the model, in order to improve the recall.

The main recall metric increased from 0.6420 to 0.6516, enhancing the model's ability to correctly identify all selected players while minimizing false positives. This ensures comprehensive identification of potential draftees, which is critical for effective scouting and decision-making.

Key Challenges

Missing Data

Addressing missing data in the dataset is a critical challenge, requiring careful imputation techniques to ensure model reliability.

Model Generalization

Ensuring the model can generalize well to unseen data is crucial for its practical viability.

Balancing Metrics

Balancing different performance metrics, such as accuracy and recall, is necessary to optimize the model for project goals.

Implemented Solutions

Techniques for Handling Missing Values

Using imputation techniques to address missing values in the dataset, ensuring the machine learning model is trained on complete and reliable data.

Cross-Validation

Applying cross-validation to avoid overfitting and ensure the model can generalize well to unseen data.

Focusing on Recall

Prioritizing recall as a key performance metric to minimize false negatives and accurately identify selected players.

SHAP Analysis

1

Explain Predictions

Use SHAP values to understand how each feature impacts the model's predictions.

2

Insights and Interpretability

SHAP analysis provides valuable insights and improves the interpretability of the model's decision process.

Practical Application

1

Data Collection

Collect real data on new players

2

Feature Engineering

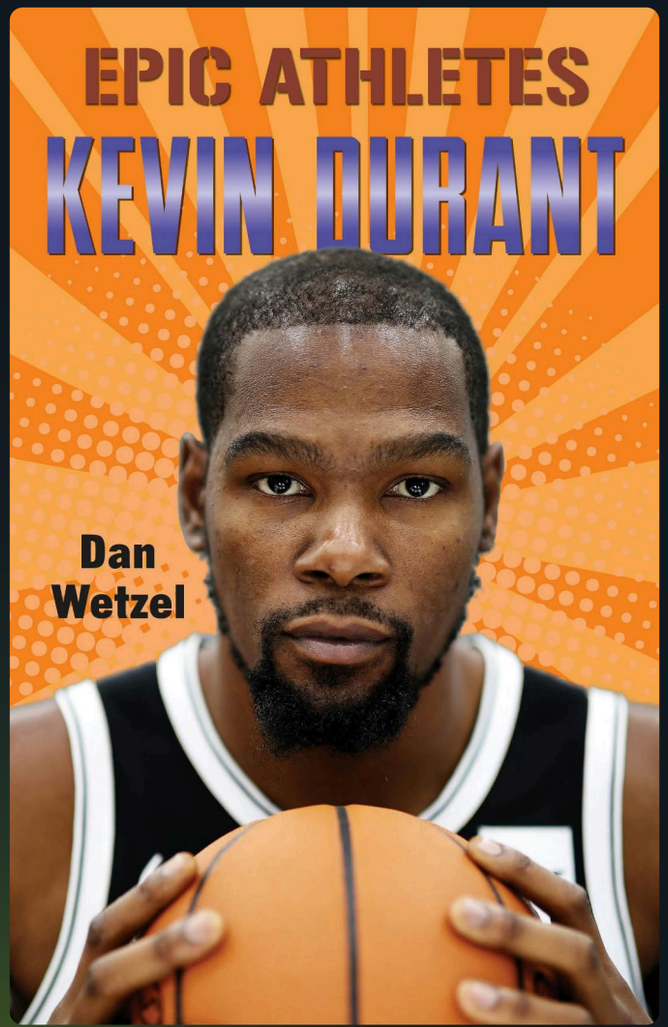
Apply feature engineering and standardization

3

Predict Draft Position

Use the saved model to predict draft position

Kevin Durant Apply



- Height (without shoes): 81.75 inches (6'9.75")
- Weight: 240 lbs
- Wingspan: 87.5 inches
- Standing Reach: 96.0 inches
- Vertical (No-Step): 33.5 inches
- Vertical (Max): 39.0 inches
- Bench Press Reps: 3
- Lane Agility Time: 11.18 seconds
- 3/4 Court Sprint: 3.45 seconds
- Body Fat %: 6.9%
- Combine Participation: 2007

The analysis for **Kevin Durant** shows that he has impressive athletic qualities, and the model predicts he will be selected with a 0.90 probability.

Kevin Durant was selected in the 2007 NBA Draft. He was chosen 2nd overall by the Seattle SuperSonics, who later became the Oklahoma City Thunder.

Bronny James Analysis

Height (without shoes)	74.75 inches (6'2.75")
Weight	180 pounds
Wingspan	80.0 inches
Standing Reach	95.0 inches
Vertical (Standing)	31.5 inches
Vertical (Max)	35.5 inches
Bench Press Reps	10
Lane Agility Time	11.00 seconds
3/4 Court Sprint	3.25 seconds
Body Fat %	7.2%
Participated in Combine	2024

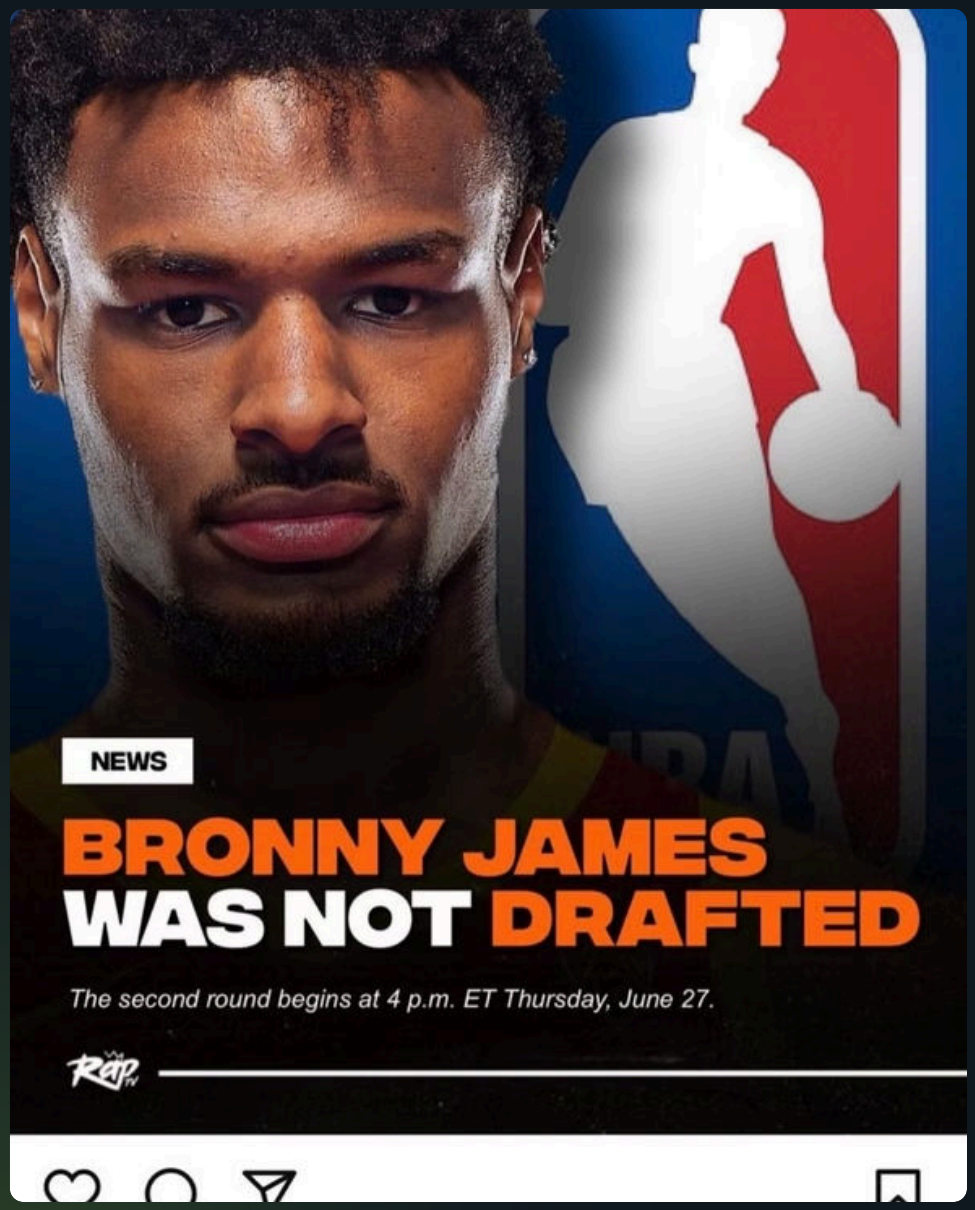
My Bronny James' draft analysis: Drafted (Probability: 0.68), there is a lot of disagreement about whether Bronny James will be drafted in the 2024 NBA Draft.

Many mock drafts, including Sports Illustrated and The Ringer, project him to be selected in the second round, often by the Los Angeles Lakers, partly due to the pairing with his father LeBron James.

Other teams like the Utah Jazz and Miami Heat have also expressed interest, with the Jazz holding the 32nd pick and the Heat having the 15th and 43rd picks. Despite these interests, concerns remain about Bronny's performance at USC and the incident of cardiac arrest, and he was not included in ESPN's latest mock draft.

Betting odds show the Lakers as the most likely to select him, but there is uncertainty around his draft position, reflecting the differing opinions of analysts. Overall, Bronny's draft status is a much-discussed topic.

Bronny James' NBA DRAFT OUTCOME



NBA Draft 1st Round Outcome

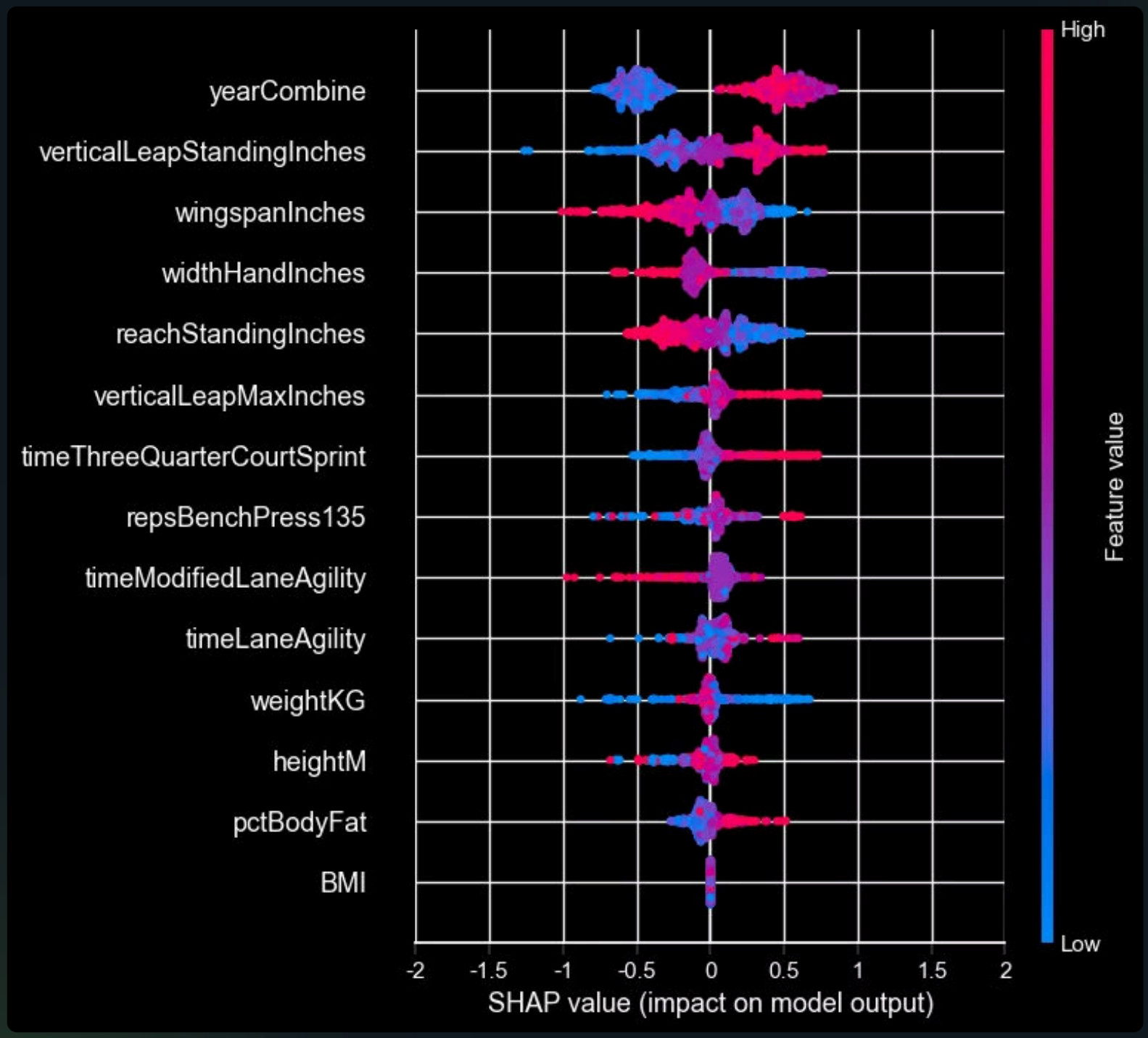
Just as predicted, Bronny James wasn't selected in the first round of the 2024 NBA Draft.

NBA Draft 2nd Round Outcome

But what really blew my mind was that my analysis was spot-on: Bronny James got chosen by the Los Angeles Lakers in the second round

This shows my analysis had value... YAAAAAAAAS! PS: Wrapped up this project on June 13, 2024.

SHAP Analysis: Key Insights



Top Features

yearCombine, verticalLeapStandingInches, and wingspanInches have the widest spread in SHAP values, indicating significant impact on the model's predictions.

Positive Impacts

yearCombine, verticalLeapStandingInches, and wingspanInches positively influence the predictions, suggesting more recent combines and higher physical attributes lead to better performance metrics.

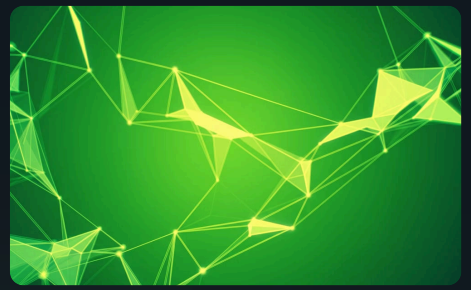
Negative Impacts

timeThreeQuarterCourtSprint and timeModifiedLaneAgility negatively impact the predictions, indicating that faster sprint times and better agility are crucial for higher performance.

Mixed Impact

weightKG shows both positive and negative impacts depending on other factors, suggesting weight is a complex factor in predicting performance.

Reflection on the Model's Improvement



Interaction Features

Create interaction terms between the most impactful features, like `verticalLeapStandingInches * wingspanInches`, to capture combined athleticism metrics.

Feature Engineering

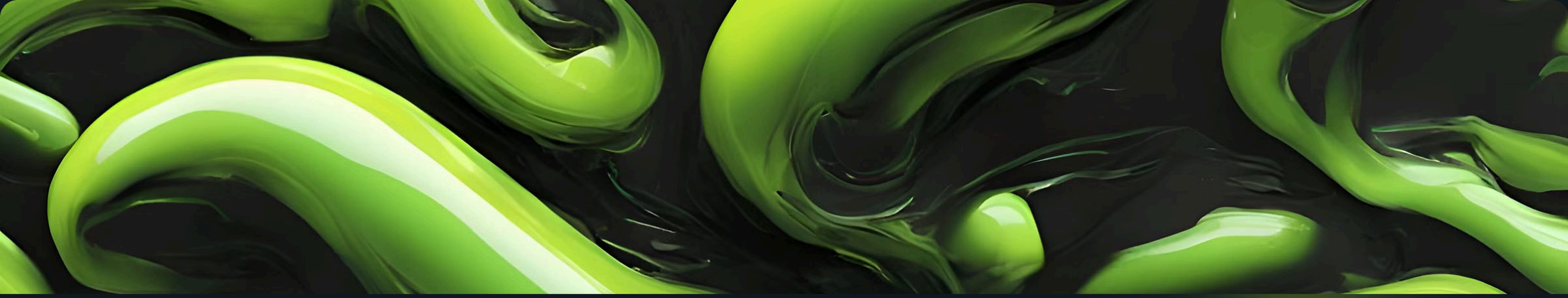
Implementing more relevant features like Points Per Game, Assists Per Game, and Shooting Percentages can significantly improve the model's prediction accuracy.

Leveraging Recent Data

Focusing the model on data from the last five years can better capture the evolving playing styles in the NBA and further improve the model's accuracy.

Excluding Superstars

Filtering out exceptionally talented players (outliers) and focusing the model on predicting the performance of more average players can provide greater insight into the true value of the model.



Final Thoughts

This project demonstrates the powerful potential of data analysis and mahicne learning in the sports domain, and its ability to transform the NBA draft process.

This work emphasizes the importance of continuous learning and adaptation to remain at the forefront of rapidly evolving fields.

Thank you!

I appreciate your time reading my work.