A Machine Learning Approach to Predicting NBA Draft Outcomes

Kevin McAlister| AI.Data Lab 2025

Xiaofei Wang, xiaofei.wang@emory.edu
Brian Li, brian.li@emory.edu
Selena Su, ysu224@emory.edu
Marco Li, marco.li@emory.edu
Zach Schneider, zach.schneider@emory.edu
Keven Liao, kaitao.liao@emory.edu

EMORY UNIVERSITY Center for AI Learning

INTRODUCTION

The Challenge of Talent Evaluation

Every year, NBA teams invest millions of dollars in draft picks, yet many late-round selections and undrafted players become stars while high picks fail to deliver. This represents a significant market inefficiency in talent evaluation.

Key Questions:

- ➢ Can machine learning predict NBA draft outcomes from college statistics?
- ➢ What factors distinguish drafted from undrafted players?
- ➢ Can we identify "diamond" players—late-round picks or undrafted players who become NBA stars—before they're selected?
- ➢ True diamond players are **extremely hard to detect** using college stats alone.
- ➢ Most late-round stars had **ordinary college profiles.**
- ➢ NBA success depends heavily on:
  - Development environment
  - Opportunity & team fit
  - Coaching and post-draft growth

OBJECTIVES

Primary Goals

1. Binary Classification: Predict whether a college player will be drafted to the NBA
2. Regression Analysis: Predict specific draft position (picks 1-60)
3. Diamond Detection: Identify undervalued players who become NBA stars

Research Questions

- ➢ What college statistics best predict NBA draft status?
- ➢ How accurately can machine learning predict draft outcomes?
- ➢ Can we detect diamond players from college performance alone?
- ➢ What insights can improve NBA talent evaluation processes?

DIAMOND PLAYERS

The Diamond Player Phenomenon 💎

Definition: Players drafted in the 2nd round (picks 31-60) or undrafted **who achieved:**

- ➢ Career WAR (Wins Above Replacement) ≥ 10.0
- ➢ Career minutes played ≥ 2,000 (~1-2 seasons)

Results

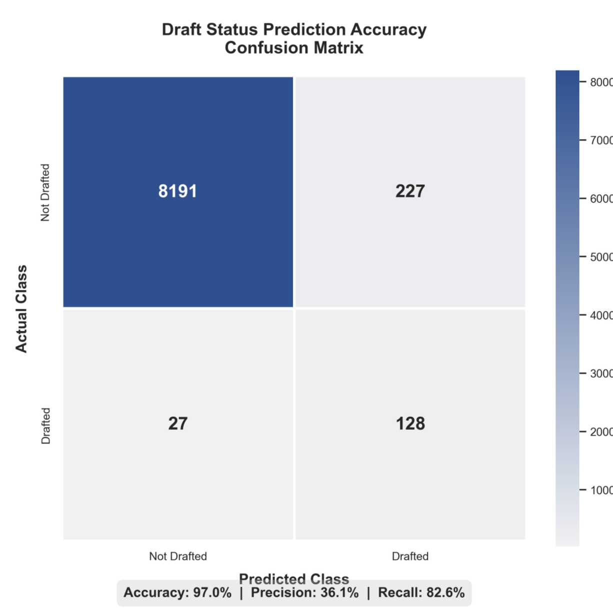
📊 Draft Classification

ROC-AUC: 0.97 → excellent discrimination

Draft Position Prediction

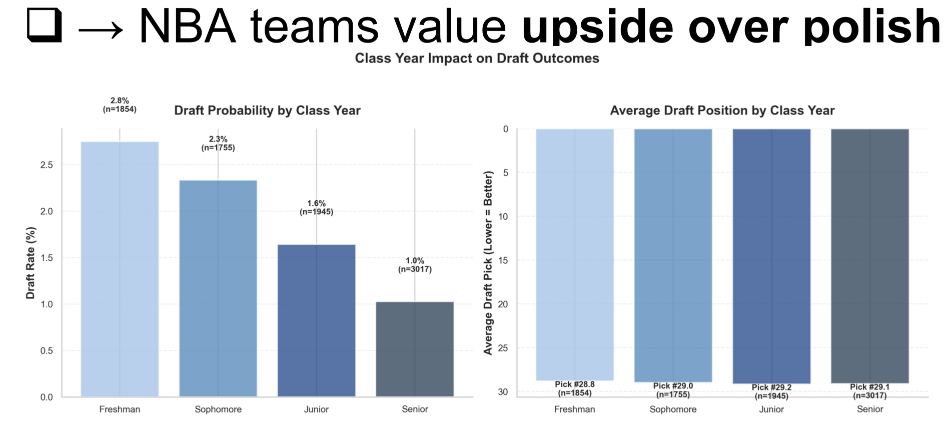
MAE: ~12 picks

51.6% within 10 picks, comparable to NBA scouting benchmarks



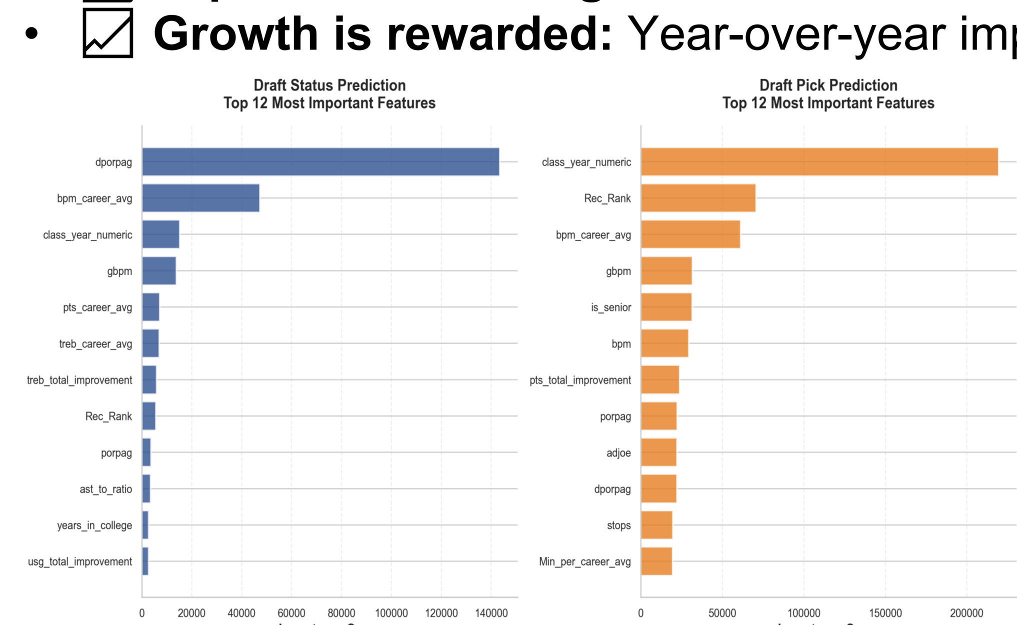
🎓 "Senior Penalty":

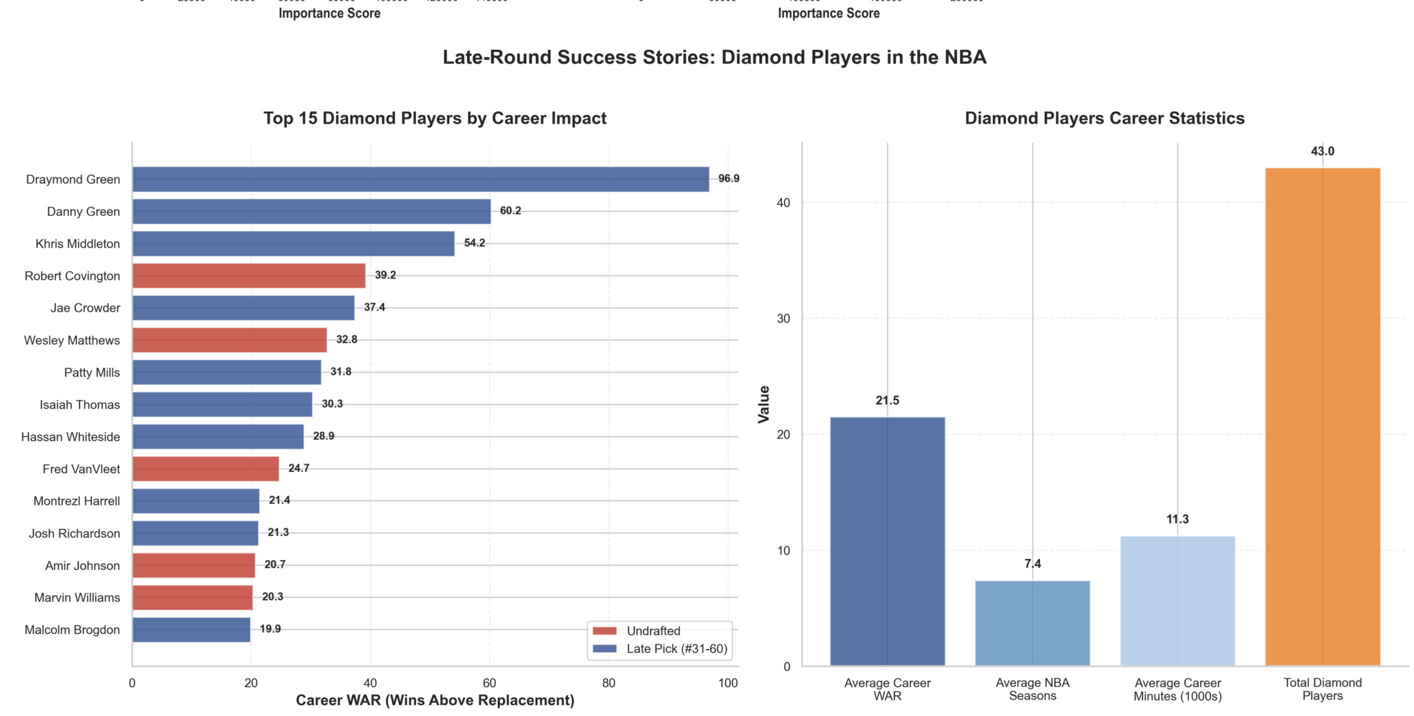
- ❑ Seniors drafted slightly more often
- ❑ Selected **~16 picks** later than freshmen on average
- ❑ → NBA teams value **upside over polish**



🔍 Feature Importance & Insights

- 🛡 **Defense matters most**: Defensive rating is the top predictor of draft status.
- 📐 **Impact over scoring:** BPM-based metrics outperform raw point totals.
- 📈 **Growth is rewarded:** Year-over-year improvement strongly increases





METHODOLOGY

🗁 DATA AND SETUP

The dataset includes **25,216 NCAA players (2009–2021),** combining college box score and advanced statistics, official NBA draft results, and **NBA career outcomes (FiveThirtyEight RAPTOR, career WAR).**

Time-based split avoids data leakage **(training: 2009–2018; testing: 2019–2021).**

💎 **Diamond players** are defined as **second-round or undrafted players with ≥10 career WAR and ≥2,000 NBA minutes (43 total).**

FEATURE ENGINEERING (PRIMARY CONTRIBUTION)

A total of **138 engineered features** capture player quality beyond raw scoring. These include:

- **Production and usage** (points, rebounds, assists, minutes load)
- **Efficiency metrics** (TS%, eFG%, points per shot, assist-to-turnover ratio) **Advanced Impact Statistics** (BPM, OBPM, DBPM, PER)
- **Defensive value** (defensive rating, steal and block rates, defensive win shares)

Contextual features measure competition quality using **strength of schedule and conference strength**, while **development trajectory features** model **year-over-year improvement in scoring, shooting, and playing time**, enabling detection of late bloomers rather than single-season performers. **Height** was excluded due to corruption in the source data.

⚙ MODELING

Models were built using **LightGBM gradient boosting**, selected for its ability to handle **large, imbalanced datasets** and **non-linear feature interactions**. Two tasks were modeled: **draft classification** (drafted vs. undrafted) and **draft position regression** (picks 1–60).

Hyperparameter Tuning:

- Grid search over learning rate, max depth, number of leaves
- Cross-validation for optimal parameter selection
- Regularization to prevent overfitting
- Early stopping to improve generalization

APPLICATION: MOCK DRAFT ANALYSIS

Model Performance on 2021 Draft:

- ✓ Successfully identified 6 of top 10 actual lottery picks
- ✓ Correctly predicted draft status for 78% of eventual draftees
- ✓ Average prediction error: 8.2 picks for top prospects