Neapolis University Pafos

Predicting NBA Rookie Impact from NCAA Statistics: A Machine Learning Approach Using Regularized Adjusted Plus-Minus (RAPM)

Authored by

Maksim Misevich

Supervised by Dr. Andrei Smolensky

May 2026

Neapolis University Pafos

Predicting NBA Rookie Impact from NCAA Statistics: A Machine Learning Approach Using Regularized Adjusted Plus-Minus (RAPM)

Dissertation which was submitted for obtaining an Applied Computer Science degree

Authored by

Maksim Misevich

Supervised by Dr. Andrei Smolensky

May 2026

Table of Contents

Validity Page

Student's Name & Surname: Maksim Misevich

Dissertation Title: Predicting NBA Rookie Impact from NCAA Statistics: A Machine Learning Approach Using Regularized Adjusted Plus-Minus (RAPM)

This Dissertation was prepared in the context of the studies for obtaining a Degree at Neapolis University Pafos and was approved on 22 May 2026 by the members of the Examiners' Committee.

Examiners' Committee:

First Supervisor (Neapolis University Pafos): Dr. Andrei Smolensky

Committee member: Dr. Avgousta Kyriakidou-Zacharoudiou

DECLARATION

I, Maksim Misevich, being fully aware of the consequences of plagiarism, declare responsibly that this paper entitled "Predicting NBA Rookie Impact from NCAA Statistics: A Machine Learning Approach Using Regularized Adjusted Plus-Minus (RAPM)", is strictly a product of my own personal work and all sources used have been duly stated in the bibliographic citations and references. Where I have used ideas, text and/or sources of other authors, they are clearly mentioned in the text with the appropriate citation and the relevant reference is included in the bibliographic references section with a full description.

The Declarant

Maksim Misevich

Abstract

Projecting which college basketball players will succeed in the NBA remains one of the most challenging tasks in professional sports. This study utilizes publicly available NCAA statistics to investigate whether machine learning models can predict players' rookie window impact, using Regularized Adjusted Plus-Minus (RAPM) over their first four NBA seasons as the proxy metric.

Using data on 546 drafted players and undrafted free agents from 2008 to 2021, the analysis applies extensive feature engineering including era-adjusted efficiency metrics, per-100 possessions standardization, team quality ratings and team strength of schedule. A temporal train (2008-2018) / test (2019-2021) split is employed to evaluate whether the model can generalize on new, unseen data. Two models are compared: a regularized linear model (ElasticNet) and a gradient-boosted tree ensemble (XGBoost), with hyperparameter tuning via cross-validation grouped by draft year.

Both models explain approximately 7-8% of the variance in the rookie window RAPM on the held-out test set. However, despite the modest R^2 values, both models demonstrated meaningful ranking ability, outperforming the aggregated pre-draft media consensus board (pooled Spearman ρ of 0.265 for XGBoost and 0.233 for ElasticNet vs 0.050 for consensus). Feature importance analysis shows that both models converge on defensive stops per 100 possessions, free-throw percentage, assists to turnover ratio and draft age as the strongest predictors of early career NBA impact.

The findings indicate that while college statistics provide a limited but real signal for NBA projection, interpretable models can still add value as a supplementary tool to traditional scouting by identifying undervalued or overvalued prospects relative to consensus expectations.

Keywords: NBA Draft, player projection, RAPM, machine learning, sports analytics, college basketball, regression.

List of Figures

List of Tables

Chapter 1 Introduction

The NBA Draft is the primary mechanism through which new talent enters the league. Held annually since 1947, it allows teams to select eligible players in a predetermined order. That order is generally determined by the previous season's standings, with the intention of promoting competitive balance by giving struggling teams the earliest opportunity to select high-upside talent.

The selection of new talent is one of the most important decisions the NBA franchises make every year. Teams allocate years of salaries and development resources to players they believe will succeed. The difference between a good and a poor selection can be the difference between a team transitioning into title contention and one that has spent millions of dollars and several seasons on a player who does not deliver an adequate return on that investment. Despite the enormous competitive and financial stakes, projecting which players will succeed remains an extremely difficult task. The history is full of high-profile “busts”, players selected near the top who failed to meet expectations, and of “steals”, players whom many teams passed on who eventually developed into All-Star tier players. A recent example of the latter is Jalen Brunson selected by the Dallas Mavericks as the 33rd overall pick in the second round of the 2018 NBA Draft, who has since become a multiple-time All-Star and a clear instance of front offices collectively misjudging a prospect. Such cases illustrate that, even with extensive scouting resources, professional evaluators can still make errors.

Before entering the draft, players demonstrate their abilities by competing in other leagues, the most prominent of which are the NCAA, the EuroLeague and Australia's National Basketball League. The persistent difficulty of identifying the right players to select has motivated a data-driven approach. If a measurable signal exists in the statistical record a player accumulates before entering the league, then a trained model could supplement traditional scouting and help quantify what is otherwise an intuition-led process. Because most draftees come from the NCAA, the study restricts its scope to that sample and aims to investigate whether the pre-draft NCAA statistics carry a usable signal for predicting players’ NBA impact and which aspects of college performance translate most reliably to the NBA.

Specifically, this study addresses three questions.

1. Can machine learning models accurately predict a player’s rookie window impact from his pre-draft college statistics?
2. Which college statistics carry the most predictive signal?
3. Can the model rank prospects more accurately than the established pre-draft consensus?

By addressing each question, the study aims not to produce a finished scouting tool, but to establish whether predictive signal exists in pre-draft data and explain it.

Chapter 2 Literature Review

2.1. Advances in Measuring Individual Player Impact

Early approaches to NBA player evaluation attempt to infer individual value by aggregating box-score statistics into a single continuous figure. Metrics such as Player Efficiency Rating (PER) (Basketball Reference, n.d.-a) and Win Shares (Basketball Reference, n.d.-b) transform observable box-score statistics, like Points, Rebounds, Assists, and other counting stats into single-number estimates of player contribution. Both metrics have well-documented limitations. Win Shares is heavily influenced by team win totals, which do not necessarily reflect individual contribution to those wins. PER largely measures offensive performance and is regarded as unreliable when it comes to measuring defensive impact. More fundamentally, box-score metrics cannot account for contribution that stat keepers do not record, such as off-ball movement, gravity, screen-setting on offense, or positioning, rim deterrence, box-outs and timely rotations on defense.

In response to these limitations, plus-minus-based metrics were introduced. Rosenbaum (2004) introduced Adjusted Plus-Minus (APM), which applies least-squares regression to lineup-based point differentials to isolate individual contributions from the surrounding lineup context. While APM offers a more complete picture of player impact than raw +/-, it still suffers from several flaws, as it is sensitive to outliers, affected by multicollinearity between players who frequently share the floor, and produces unstable estimates when lineup sample sizes are small. To address these issues, Regularized Adjusted Plus-Minus (RAPM) replaces APM's ordinary least-squares regression with ridge regression (Sill, 2010). The ridge penalty shrinks coefficient estimates towards zero, reducing variance and improving out-of-sample reliability.

A related but methodologically distinct family of metrics is Statistical Plus-Minus (SPM). Unlike PER or Win Shares, whose weights are largely chosen by their designers, SPM is a regression model, where box-score statistics serve as input features, and the model is trained to predict a player's historical RAPM as the target. SPM therefore produces a plus-minus-style estimate of player value derived entirely from box-score statistics. The most prominent SPM variant is Daniel Myers's Box Plus-Minus (BPM), whose coefficients are empirically fitted rather than hand-chosen (Myers, 2020).

To reduce the noise even further, more recent metrics like Estimated Plus-Minus (EPM) use a hybrid approach of combining the SPM and RAPM components together. Rather than allowing the ridge regression to default toward zero when the signal is weak, EPM makes the ridge default towards each player's SPM estimate, effectively using SPM as a Bayesian prior (Dunks & Threes, n.d.). A low-minutes player's EPM therefore leans on the SPM estimate, while a high-minutes player's EPM is mostly informed by their RAPM signal. A comparison of all-in-one player metrics by Dunks & Threes (2020) found that metrics combining RAPM with a Bayesian prior, such as EPM, achieved the lowest prediction error in a retrodiction test. The analysis further suggested that this advantage came primarily from the direct use of RAPM rather than from the accuracy of the statistical prior.

2.2. NCAA-to-NBA Modelling Research

One of the first papers that addressed this problem is Greene (2015), who applied multiple linear regression to first-round NBA draft picks from 1985 to 2005 in order to attempt to predict PER, Win Shares and Win Shares per 48 (WS/48) minutes using college box-score statistics, physical measurements and contextual variables like All-American honors and NCAA tournament seed. The regression explained only a small share of variance in NBA outcomes, with the WS/48 reaching an R^2 of just 0.11. Field Goal %, blocks and first-team All-American selections emerged as the most consistent predictors of all three target features. Despite the low explanatory power, Greene found that the rankings produced by the WS/48 model did a better job at ranking the players than the actual NBA draft order in most of the seasons studied, whereas the PER and Win Shares models did not.

While Greene (2015) relied on traditional linear regression, Kannan et al. (2018) reframed the task as one of supervised classification and applied a machine learning approach to it. Their study used the participants in the 2009-2014 NBA draft combines who had completed the biometric measurements, yielding a final sample of 194 players. The outcome was a binary indicator of success, based on whether a player had accumulated 174 NBA games. This number corresponds to the average number of games played across the dataset. Three classifiers were evaluated. The random forest performed best across precision, recall and F1-score and was used for further analysis. The classifier was then split into two models: a reduced model containing only combine biometrics and player’s age, and a full model that additionally incorporated college statistics, the regular-season ratings, draft selection order and positional indicators. The reduced model showed little predictive power, whereas the full model marginally improved. Upon a feature importance inspection, it was determined that draft order was by far the strongest predictor, followed by college box-score statistics like points and rebounds per game. Kannan et al. (2018) concluded that collegiate performance and draft position were the strongest predictors of NBA success, while the biometric data collected at the combine had little predictive power.

A more recent contribution by Costa (2024) extended this line of work by focusing on interpretability, arguing that models must not only be accurate, but also comprehensible to the decision makers. Working with the NCAA dataset covering the 2009-2021 period, the authors framed the task as a classification problem and used CART, C4.5, C5.0 and Logistic Regression as white-box models, compared against two black-box models, the Support Vector Machine and Multilayer Perceptron. Each classifier was paired with a genetic algorithm used for feature selection, with the goal of reducing the feature set while preserving or improving predictive performance. The accuracy of white-box models remained close to that of the black-box baselines, proving that interpretability does not need to be sacrificed for accuracy. The C4.5 tree had the highest raw accuracy, however its precision and recall collapsed. Logistic Regression combined with the genetic algorithm delivered the most balanced performance across accuracy, precision, recall and F1-score. A SHAP analysis of that model identified that composite metrics such as dporpag and BPM contributed the most to the prediction.

2.3. Key Limitations and Research Gaps

While the previous studies establish that college performance carries signal for projecting NBA outcomes, they share several limitations. First, much of the prior work focuses on classification rather than regression. This discards the magnitude of a player's professional impact, treating a marginal bench player and an All-Star as equivalent successes. Second, the studies that do model a continuous outcome regress onto box-score aggregate metrics. Greene (2015) used PER, Win Shares and WS/48, the very metrics whose limitations were pointed out in Section 2.1. Third, prior work pays little attention to era and pace effects. Greene (2015) used college counting statistics spanning two decades without performing any era adjustments, essentially treating two decades worth of seasons as a constant environment. Fourth, the existing literature largely neglects the team context in which college statistics are produced. A player's output is influenced by the strength of his team and the difficulty of its schedule, yet the existing literature ignores this aspect. Two further limitations concern the robustness and design of the prior work. Greene's (2015) study scope was older picks, from 1985 to 2005, and a great deal has changed since then, so findings drawn from that earlier player population may not necessarily generalize today. More significantly, Kannan et al. (2018) included draft selection order as one of the predictors, and it was found to be the most important feature in their model. The issue with this approach is that draft position is not a college statistic but rather an aggregate forecast of NBA success, formed by the collective judgement of NBA front offices. Using it as a predictor introduces a circularity, as the model is trying to predict NBA success from existing professional opinions of that success rather than from college performance. This inflates the apparent predictive power of the model, and it leaves the model of little practical use to anyone seeking to project a prospect's success before the draft itself has taken place.

2.4. Contributions from the Present Study

To address the limitations identified in Chapter 2.3, this study makes several contributions to the modelling of NCAA-to-NBA player projection. The principal contribution is the choice of the target variable, instead of predicting PER or Win Shares, or reducing the outcome to a binary label, the study regresses college performance onto RAPM. As established in Section 2.1, RAPM isolates individual player impact from the surrounding lineup more reliably than box-score metrics and provides a better ground truth against which to learn. Framing the task as a regression problem further preserves the magnitude of professional impact, allowing the model to distinguish a high-impact player from just a decent bench piece rather than treating both as equals.

A second contribution lies in the construction of the feature set. College basketball is not a vacuum for every player, the environment is different for everyone, whether it’s the level of players across the entire country, or pace of play, or team help, or difficulty of opponents. This study therefore adjusts statistics relative to season league averages, normalized production to a per-100 possessions basis and incorporates explicit team and opponent context features. These adjustments put college

production from different years under the same scale, an aspect that has been largely ignored in prior literature.

A third contribution is methodological. The feature set deliberately excludes draft selection order. As argued in Section 2.3, draft position is already a forecast of NBA success and including it would allow a model to predict NBA outcomes based on professional judgements, rather than from college performance. By excluding it, this study ensures that its predictors rest entirely on college statistics and team context, and that the measured predictive power is attributable to information available prior to the NBA draft.

A fourth contribution concerns evaluation. The data is portioned using a temporal train/test split, with earlier draft classes used for training and later classes held out for testing, a random split would allow information about a given era appear in both test and training sets. A temporal split mirrors a real-world application of such model, where it gets trained on older players to predict player it has not seen yet. This study additionally evaluates its models against the pre-draft consensus of professional analysts and scouts, providing an external benchmark for whether the model can match or exceed the common view.

Lastly, this study compares an interpretable linear model with a more flexible non-linear model and examines whether they agree on feature selection, predictive power of those features and their importance. This combination allows the findings to be interpreted with confidence, if two model families converge on the same conclusion, then those conclusions are more likely to carry some truth.

Chapter 3 Data

3.1. Data Sources

This study utilizes four publicly available data sources. NCAA player and team statistics were sourced from barttorvik.com (Torvik, n.d.), covering Division I players and programs across the 2007-08 to 2020-21 seasons. NCAA league average statistics for the same period were collected from sports-reference.com (Sports Reference, n.d.-a) and later used to adjust player efficiency statistics for different eras, as described in Section 3.4. Official NBA draft results from 2008 to 2021 and the exact draft dates were obtained from basketball-reference.com (Sports Reference, n.d.-b), which operates under the same Sports Reference network. RAPM values were manually collected from nbarapm.com (databallr, n.d.), with each player's rookie window value recorded according to the procedure explained in Section 3.3.

3.2. Dataset Construction

The raw barttorvik datasets have 65,649 entries consisting of players who have logged any NCAA minutes from the 2007-08 to 2020-21 seasons. According to research done by the NCAA, it is estimated that 3.4% of draft-eligible Division I players were selected in the 2024 draft (NCAA, 2025), showcasing how rare it is to transition from the NCAA to the NBA and how many players need to be filtered out from the raw datasets. To construct a dataset of players' pre-draft NCAA statistics and NBA outcome data, a multi-step filtering and merging procedure was followed.

The full barttorvik dataset was first loaded for each season. The first step was to identify notable undrafted free agents (UDFAs) who signed NBA contracts between 2008-2021, extract and save their college statistics separately. Only those UDFAs who entered the league within two years of going undrafted were included. This approach ensures that their NBA performance can still be attributed to their final college statistics captured in the dataset, rather than professional development in other professional leagues like the EuroLeague, or the NBL.

The next step was to filter out those players who went undrafted but also are not part of the UDFA dataset. By removing all the rows where the draft pick value was missing, only the players who eventually got selected via the NBA draft were kept.

The new dataset of drafted-only players contained duplicate entries caused by returning college players. For example, Ja Morant appeared in two different seasons, 2017-18 and 2018-19. To resolve such cases, all the drafted players were cross-referenced with official NBA results from basketball-reference.com. By keeping only players whose names appear in the corresponding draft year's results, each player is represented exactly once, using their final college season. Additional name discrepancies between barttorvik and basketball-reference datasets were identified and resolved through manual mapping on a year-by-year basis.

3.3. Target Variable: Rookie Window RAPM

The target variable for this study is a player's Regularized Adjusted Plus-Minus (RAPM) during their rookie window, defined as the four-year window beginning from their first NBA rookie season. RAPM was selected over traditional box-score metrics like Player Efficiency Rating (PER) or Win Shares due to its ability to isolate individual player impact on team point differential while controlling for factors like quality of teammates and strength of opponents (Sill, 2010). Unlike counting statistics, RAPM does not reward players whose stats look good on the surface but don't translate to winning basketball. Likewise, it does not penalize those with ordinary stats due to limited roles.

RAPM values for players in the Section 3.2 dataset, both drafted and signed UDFAs, were manually sourced from nbarapm.com. During this manual collection process, players with extremely limited NBA appearances were not included in the dataset as their RAPM estimates are unreliable. These players therefore were excluded at the point of collection rather than during a later filtering stage.

The rookie year is the starting point of the window, defined as the first season in which a player logged at least 48 minutes of total play, which is equivalent to one full NBA game. Seasons that did not meet this threshold were excluded and did not count as the rookie year. This threshold establishes a consistent and objective baseline for all players, avoiding arbitrary judgements about which early-season appearances constitute a true rookie year.

For players who do not have four complete seasons during their rookie window, the following rules were applied to keep things consistent:

- 1 season only: the single-season RAPM value from the 6Factor page is used.
- 2 seasons: the 2Y RAPM for the active span is used.
- 3 seasons: the 3Y RAPM for the active span is used.
- Gap seasons within the window: if a player missed one or more intermediate seasons within the defined span, the appropriate multi-year RAPM for the active span is used, with missing seasons ignored. E.g., if a player played seasons 1, 2 and 4, but not season 3, then his 4Y RAPM is taken.

Upon completion of the RAPM dataset, 13 players had their rookie window adjusted due to falling below the 48-minute threshold. In these cases, their window was shifted forward until the first season met the requirement without violating the two-year constraint mentioned in Section 3.2. One player, Jaden Springer, was removed entirely from the dataset since his window shift put his rookie season in what would be equivalent to the 2022 draft cohort, which was excluded from the study due to not having a complete four-year window at the time of data collection. After applying these adjustments and removing Jaden Springer, the initial RAPM database consisted of 595 players.

Lastly, to filter out noisy entries from the RAPM database, a minimum minutes threshold was applied. Players who played fewer than 500 total NBA minutes throughout their rookie window were excluded. The 500-minute threshold was determined empirically after examining the distribution of total minutes across the initial dataset. Players below this threshold carry high variance and noise, and their RAPM does not necessarily reflect true player ability. This filter removed 49 players (8.2% of the initial 595-players sample), leaving 546 players who would be fed to the models. It should be noted that total minutes values vary in precision for different players across the database. Minutes for the earlier seasons were sourced from basketball-reference.com, which provides exact figures, while later

season minutes were sourced from nbarapm.com, which rounds minute totals for players above 1000 minutes. This inconsistency is a consequence of the manual data collection process, and it does not affect the application of the 500-minute threshold in any meaningful way. Table 3.1 presents descriptive statistics for total rookie window minutes before and after filtering. Figure 3.1 and Figure 3.2 show the distributional shift via histogram and boxplot respectively.

Table 3.1: Descriptive statistics of total rookie window minutes before and after the 500-minute filter

Statistic	Before Filtering (n=595)	After Filtering (n=546)
Mean	4,361	4,729
Std	2,888	2,729
Min	20	515
50th pctl	4,092	4,355
75th pctl	6,596	6,800
90th pctl	8,398	8,500
95th pctl	9,400	9,403
99th pctl	11,171	11,280
Max	12,759	12,759

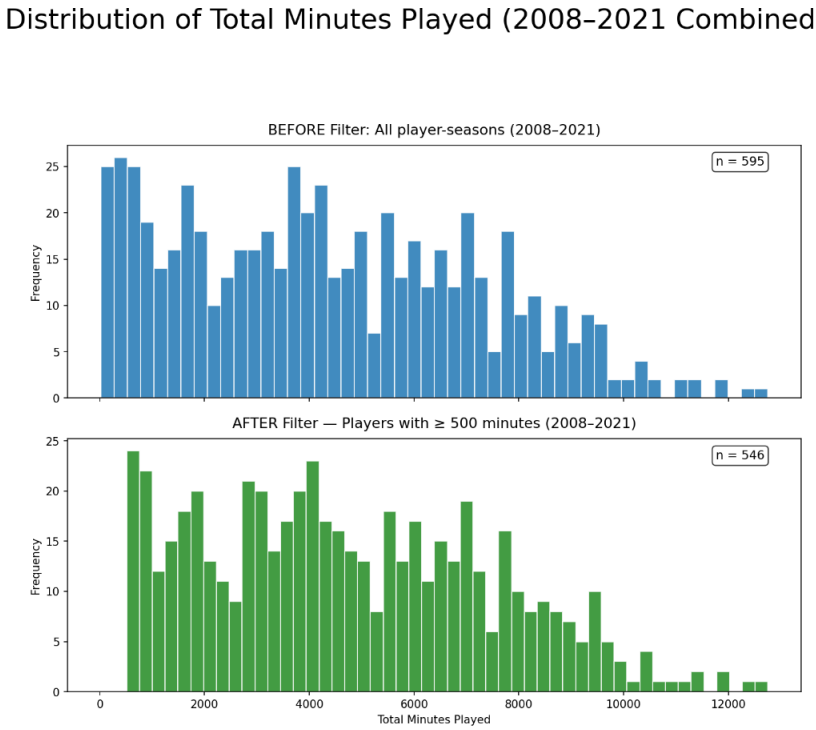


Figure 3.1: Distribution of total rookie window minutes before and after the 500-minute filter

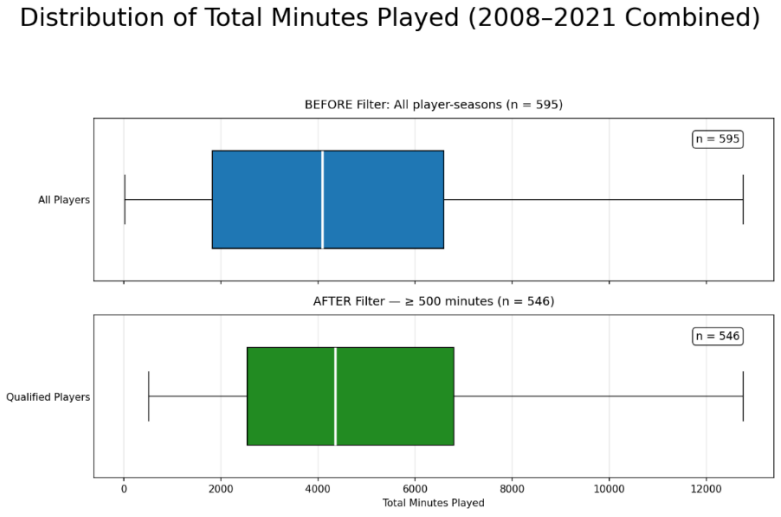


Figure 3.2: Boxplot of total rookie window minutes before and after the 500-minute filter

3.4. Feature Engineering

Raw college statistics capture absolute player production within their season; however, they do not account for the fact that college basketball is not a static environment. Over the 14-season study window, the game has changed and evolved in many different aspects. As shown in Figure 3.3 and Figure 3.4, volume and counting statistics show drift during the study window. For instance, teams started taking more 3-point shots, with 3PA rising from league average 19.1 attempts per game in 2007-08 to 21.7 by 2020-21. Rules also changed; a notable example is the NCAA extending the three-point line to the international distance of 22 feet, 1¾ inches (6.75 meters) (Johnson, 2019), which resulted in a league average drop off in 3-point shooting as players had to adjust to the greater distance. Figure 3.5 illustrates that efficiency statistics over the same period remained relatively stable, though subtle trends are still visible.

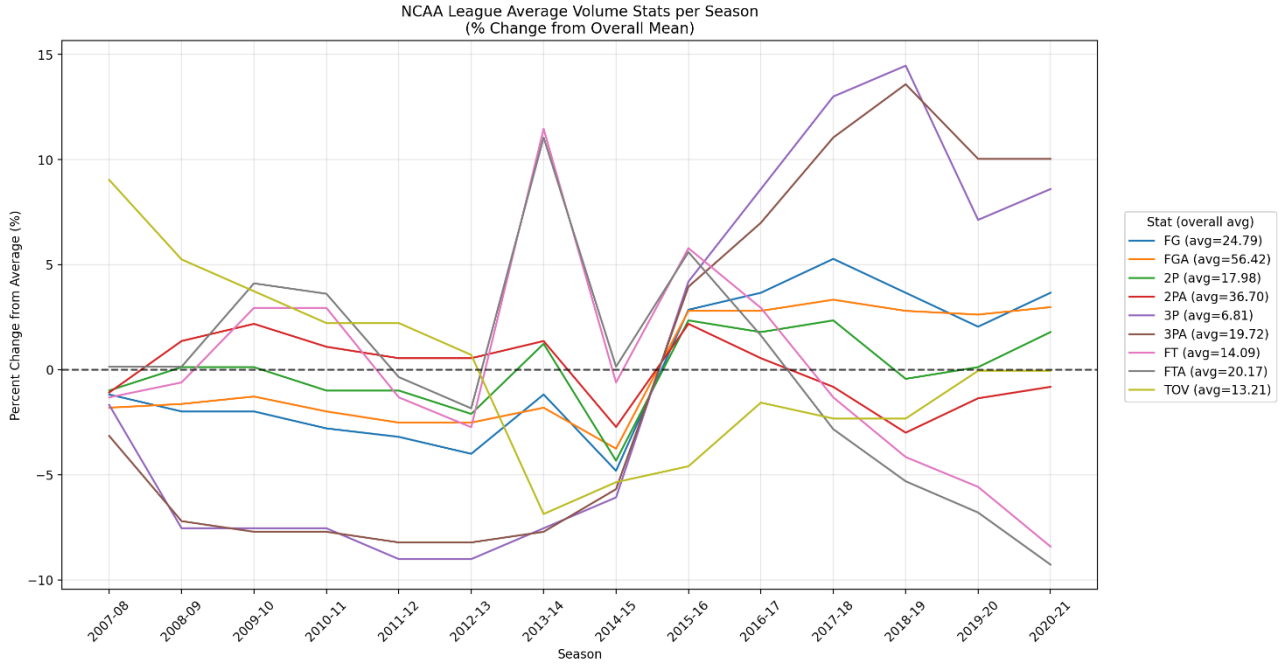


Figure 3.3: NCAA league average volume stats per season (2007-08 to 2020-21)

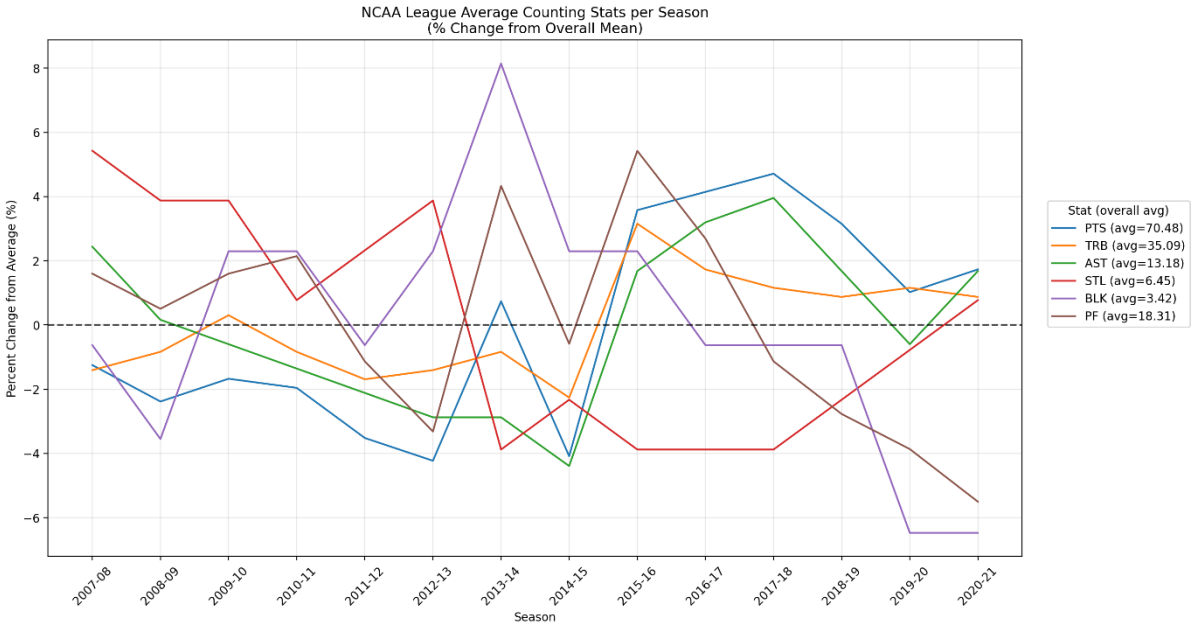


Figure 3.4: NCAA league average counting stats per season (2007-08 to 2020-21)

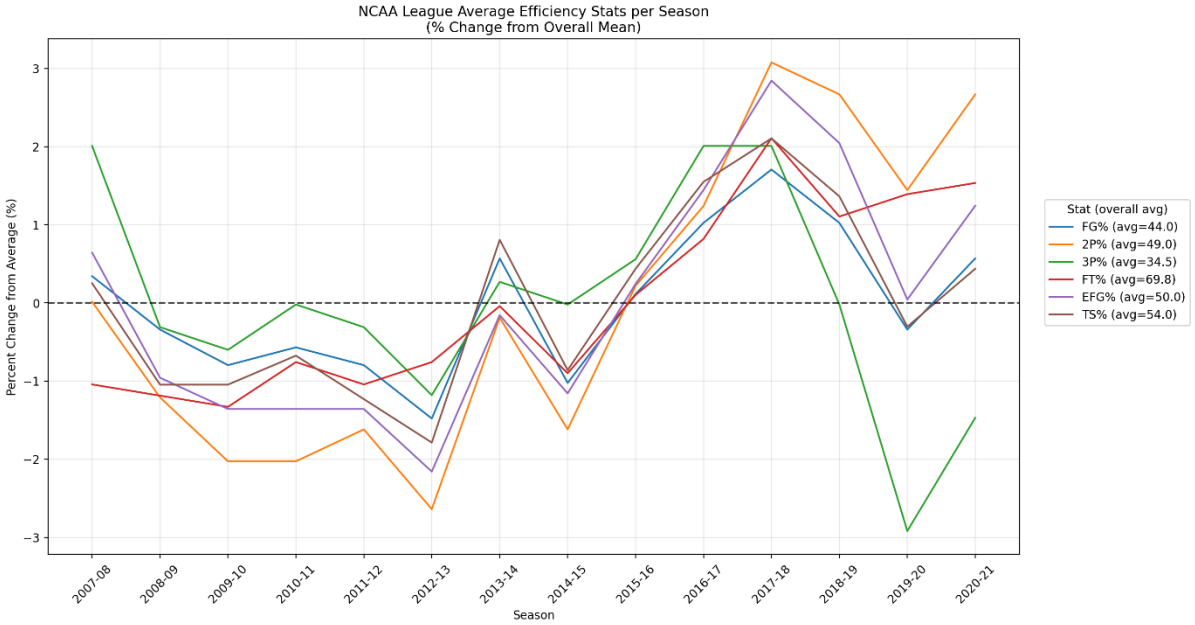


Figure 3.5: NCAA league average efficiency stats per season (2007-08 to 2020-21)

To illustrate the practical implications of these shifts, consider the following comparison between Steven Adams and Miles Bridges. Steven Adams ended the 2012-13 season with a True Shooting percentage of 55.46%, whereas Miles Bridges had 57.26% in 2017-18. At face value Bridges was more efficient; however, when their true shooting percentages are expressed as values relative to the league average for their respective seasons, Steven Adams comes out on top with a relative True Shooting percentage of +2.46%, against Bridges' +2.16%, as shown in Figure 3.6. By adjusting their percentages, it is revealed that Steven Adams was the more efficient player relative to his competitive environment, a distinction that raw absolute statistics otherwise obscure.

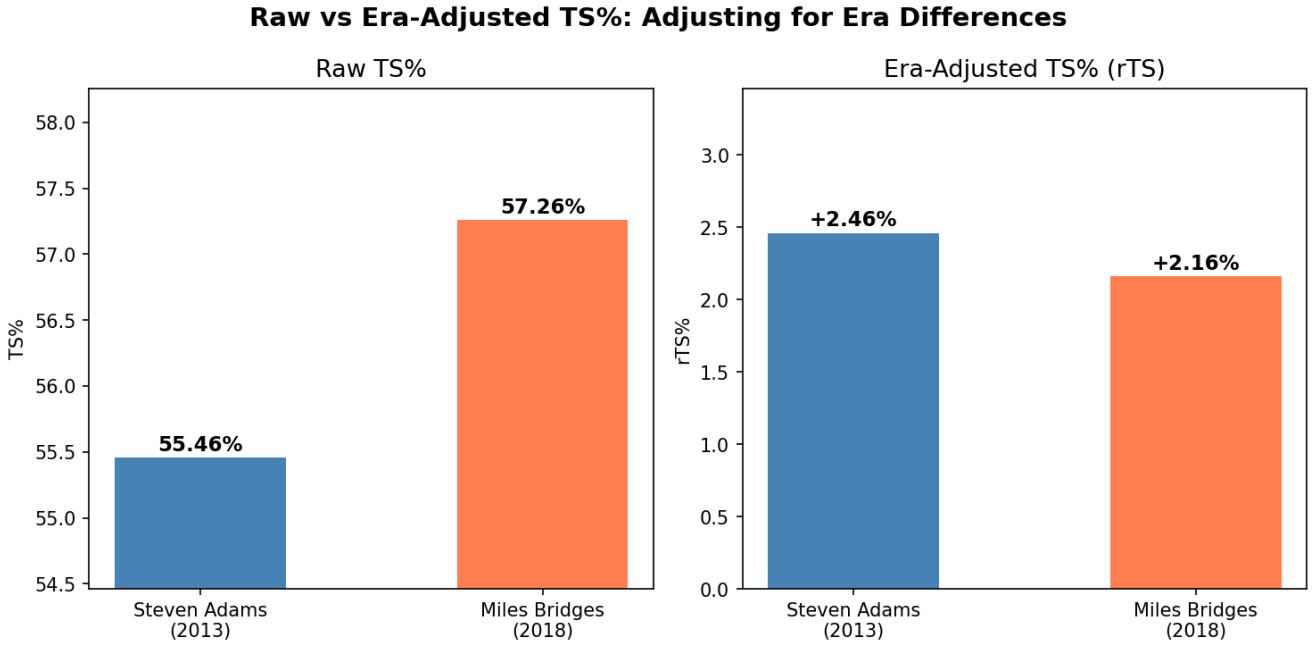


Figure 3.6: Raw vs era-adjusted TS% for Steven Adams (2013) and Miles Bridges (2018)

3.4.1 Adjusting For Different Eras

To formally assess the magnitude of the drift in league average statistics, a Coefficient of Variation (CV) was computed for each efficiency metric across the 14 seasons. League average TS%, eFG% and 2P% were not directly available in the table provided by sports-reference.com and were therefore computed from the league totals using the standard formulas:

$$\text{TS\%} = \frac{\text{PTS}}{2 \times (\text{FGA} + 0.44 \times \text{FTA})} \times 100$$

$$\text{eFG\%} = \frac{\text{FGM} + 0.5 \times \text{3PM}}{\text{FGA}} \times 100$$

$$\text{2P\%} = \frac{\text{FGM} - \text{3PM}}{\text{FGA} - \text{3PA}} \times 100$$

Results aligned with the previous figures: efficiency statistics were relatively stable, with CVs ranging from 0.009 (FG%) to 0.019 (2P%). Volume statistics showed considerably more drift, with three-point attempts and makes being at 0.087-0.089. Full CV tables for all three statistic groups are provided in Appendix A.

The most direct evidence of era change is visible in the immediate impact of the NCAA three-point line being extended. League average 3P% fell from 34.5% in 2018-19 to 33.5% in 2019-20, the first season under the extended line. Era adjustment was therefore applied uniformly across all efficiency statistics rather than selectively. Adjusting a stable metric has no analytical downside; while failing to adjust a slightly drifting one introduces bias.

For each player, new efficiency features were created by subtracting the league average from their absolute values:

- $\text{rTS} = \text{TS\%} - \text{League Average TS\%}$
- $\text{rEFG} = \text{eFG\%} - \text{League Average eFG\%}$
- $\text{r2P} = \text{2P\%} - \text{League Average 2P\%}$
- $\text{r3P} = \text{3P\%} - \text{League Average 3P\%}$
- $\text{rFT} = \text{FT\%} - \text{League Average FT\%}$

3.4.2 Possession Estimation and Per 100 Possessions Features

Players across different teams don't play the same number of minutes, nor do their teams play at the same pace. Statistics such as points, rebounds, assists, as well as many others are directly affected by how many possessions a team runs per game. A player who sees a lot of minutes on a fast-paced team will naturally have higher totals, however that alone does not necessarily mean that the player is more productive on a possession-to-possession basis. Expressing these statistics per 100 possessions puts every player under the same scale.

The barttorvik dataset does not directly provide possession count, however the individual player possessions can still be estimated using the following formula:

$$\text{Possessions} = \text{FGA} + 0.44 \times \text{FTA} + \text{TOV} - \left(\frac{\text{Team OR\%}}{100}\right) \times \text{Missed FGA}$$

Each term in this formula represents a way a possession can end: a field goal attempt, a possession ending free throw trip (the 0.44 coefficient approximates how many free throws end a possession, accounting for attempts that don't end the possession, such as the first free throw of a multi-shot trip and "and-one"), or a turnover. The final term subtracts missed shot attempts that were recovered by the offensive team, since those offensive rebounds extended the possession.

Of the inputs required to compute estimated possessions, barttorvik provides FTA in its player dataset and Team OR% in the team-level four factors dataset. FGA and missed FGA can be derived by summing 2PA + 3PA and subtracting 2PM + 3PM respectively. The last remaining input, TOV, is not directly provided by barttorvik and must therefore be estimated. Two estimation methods were evaluated. The first derives TOV from the AST/TOV ratio, by first computing total assists as:

$$\text{AST}_{\text{total}} = \text{AST per game} \times \text{GP}$$

And then:

$$\text{TOV} = \frac{\text{AST}_{\text{total}}}{\text{AST/TOV}}$$

The second method rearranges barttorvik's TO% definition: Given that:

$$\text{TO\%} = 100 \times \frac{\text{TOV}}{\text{FGA} + 0.44 \times \text{FTA} + \text{TOV}}$$

TOV can be recovered as

$$\text{TOV} = \frac{\text{TO\%} \times (\text{FGA} + 0.44 \times \text{FTA})}{100 - \text{TO\%}}$$

Each turnover estimate was plugged into the possession formula and evaluated by reconstructing each player's 3PA/100 and comparing against barttorvik's own 3PA/100 values across all 14 seasons. Results are shown in Table 3.2.

Table 3.2: Summary of possession estimation methods validation against barttorvik 3PA/100

Method	Mean Correlation	Mean MAE
AST/TOV	0.9968	0.5591
TO%	0.9961	0.7290

Both methods achieved near-identical correlations with barttorvik's own 3PA/100 (0.9968 vs 0.9961); therefore, MAE is the more informative metric for differentiating the two methods. The AST/TOV based estimation scored a mean MAE of 0.559 across 14 seasons compared to 0.729 for the TO% based estimation, a reduction of approximately 23%. This gap is consistent across all 14 seasons (see Appendix B). The AST/TOV method likely performs better because TO% uses barttorvik's own internal possession estimation assumptions, whereas AST/TOV is a simple ratio of two counting statistics. Given these results, the AST/TOV method was chosen for estimating possessions.

An important caveat is that the full possession formula requires Team OR% which must be sourced from a separate barttorvik team four factors endpoint. Researchers without access to the team level data can apply a simplified version of the formula by omitting the offensive rebounds adjustment (the final term of the formula):

$$\text{Possessions} = \text{FGA} + 0.44 \times \text{FTA} + \text{TOV}$$

For completeness, both turnover methods were also evaluated using the simplified formula. The simplified formula produces meaningfully higher mean MAE, confirming that the offensive rebound adjustment significantly improves the accuracy of possession estimation. The simplified formula is not recommended where Team OR% can be obtained but remains a viable alternative for data-limited applications. The results can be seen in Table 3.3.

Table 3.3: Summary of possession estimation methods using the simplified possession formula

Method	Mean Correlation	Mean MAE
AST/TOV	0.9964	0.9770
TO%	0.9956	1.1257

The next step was to estimate how many team possessions occurred whenever a player was on the floor using this formula. The possessions estimated above represent the number of possessions the player personally ended, not the total team possessions. Because USG% is defined as the percentage of team possessions by a player, team possessions can be recovered by rearranging this formula:

$$\text{USG\%} = 100 \times \frac{\text{Player possessions used}}{\text{Team possessions while player on floor}}$$

such as:

$$\text{Team possessions} = \frac{\text{Player possessions}}{\text{USG\%}/100}$$

This scaling was validated by confirming that implied on-court possessions per game consistently fell in the 60-75 range, which lines up with expected college basketball possessions. The following player statistics were normalized per 100 possessions using the newly calculated team possessions, ensuring direct comparability for players from different seasons:

- $\text{PTS}/100 = \frac{(\text{Points Per Game} \times \text{GP})}{\text{Estimated On-court Possessions}} \times 100$
- $\text{2PA}/100 = \frac{\text{Two-point attempts}}{\text{Estimated On-court Possessions}} \times 100$
- $\text{FTA}/100 = \frac{\text{Free Throw attempts}}{\text{Estimated On-court Possessions}} \times 100$
- $\text{Stops}/100 = \frac{\text{Stops}}{\text{Estimated On-court Possessions}} \times 100$

A comparison between raw stops and stops/100 with rookie window RAPM showed a stronger correlation for the pace-adjusted variation of this stat (0.194 vs 0.109).

The remaining features in the candidate set (USG%, FTR, AST%, AST/TOV, ORB%, DRB%, STL%, BLK% and PFR) required no transformation as they were already expressed as rates or percentages.

Several limitations of this possession estimation approach should be acknowledged. The exact formula that barttorvik uses to calculate their possessions is unknown, so the adopted formula does not perfectly match theirs. Although the strong validation against barttorvik's 3PA/100 (Table 3.2) indicates close alignment in practice, a comparison against an independent possession estimate revealed a small systematic disagreement; therefore, the adopted possession estimates should be treated as reliable approximations rather than exact possession counts.

3.4.3 Team Context Features

Having adjusted most numerical features by expressing the efficiency stats as relative to their seasons and standardizing the counting stats per 100 possessions, the next step is to address team and conference features, which are categorical. Currently each player has a team and a conference column; it would be natural to assume that these features carry signal, as players on strong teams in bad conferences may have more inflated statistics relative to their true level of play, while players on weaker teams in stronger conferences may be underrated.

A standard approach of feeding categorical data to machine learning models is to one-hot encode it, where each unique team and conference becomes a binary feature. However, this approach treats teams and conferences as stable environments, even though this is not true in college context due to roster turnover, player development and coaching changes.

To address this limitation, two team-level contextual features, sourced from barttorvik team ratings for each season from 2007-08 to 2020-21 were merged onto each player row based on their college team:

- **Team Barthag** is barttorvik's power rating for college teams, derived from barttorvik's team adjusted offensive and defensive efficiency using a Pythagorean expectancy formula (Torvik, 2018). It ranges from 0 to 1, representing a team's probability of beating an average Division I opponent.
- **Team SOS** is the team's strength of schedule rating, capturing the average quality of opponents faced throughout the season. It is computed by barttorvik across a team's full schedule, which includes conference play, non-conference play, as well as the playoffs.

Figure 3.7 illustrates the average Team SOS across the five power conferences over the study window. In 2008 the SEC's teams had the lowest average SOS of 0.642, however in 2021, the SEC

ranked as third-best conference with the average team SOS of 0.746. This improvement alone directly contradicts the one-hot encoding assumption that the conferences and teams remain fixed over the years.

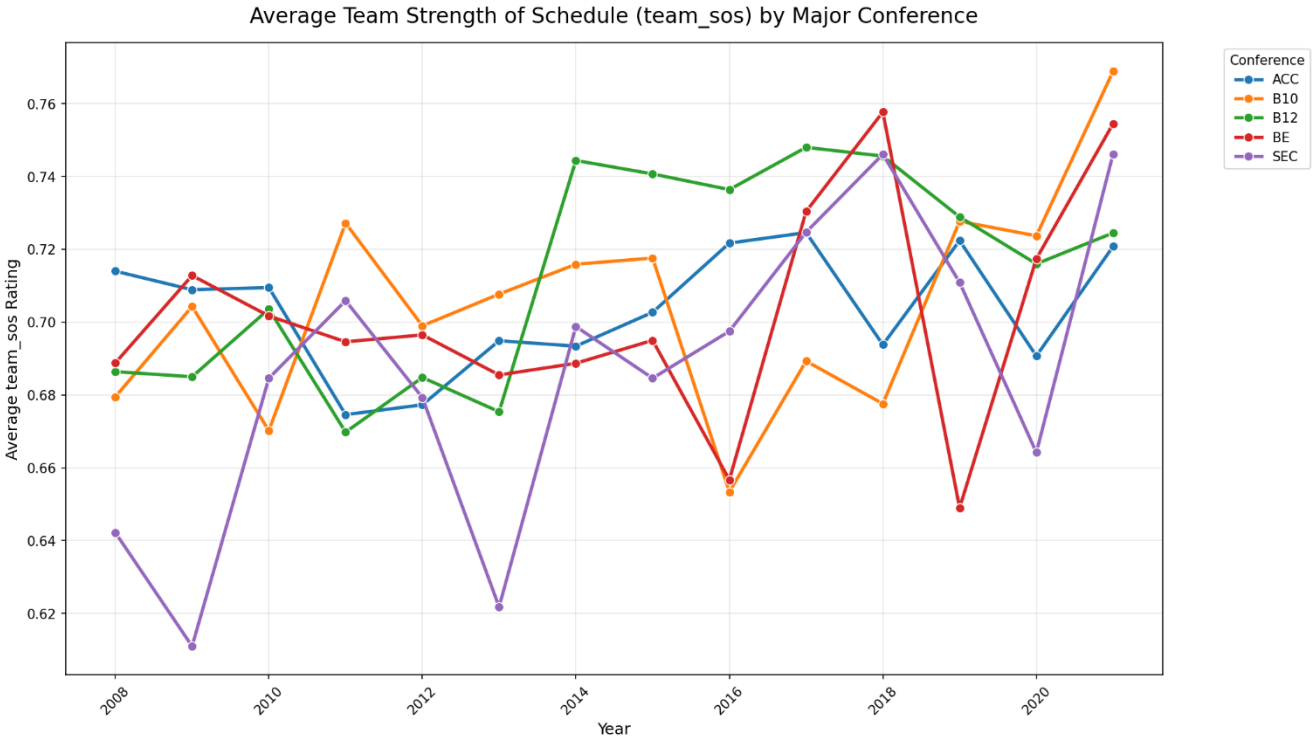


Figure 3.7: Team Strength of Schedule of ACC, B10, B12, Big East and SEC from 2008 to 2021

To further visualize the practical implications of this model, consider these two players as an example: Anthony Randolph and Cameron Thomas. If their team and conference columns were one-hot encoded, then the model would see these features as essentially a human equivalent of this:

Player	Team	Conference	Year
Anthony Randolph	LSU	SEC	2008
Cameron Thomas	LSU	SEC	2021

However, by introducing the Team Barthag and Team SOS, the model now has access to more context:

Player	Team Barthag	Team SOS	Year
Anthony Randolph	0.687805	0.641048	2008
Cameron Thomas	0.893923	0.729604	2021

3.4.4 Draft Age

A player's age at the time of the NBA draft was included as a feature to capture maturity as well as physical development of players. Two players with almost identical college productions may not necessarily translate to the NBA the same: a 23-year-old senior averaging 20 points per game in college inherently benefits from four additional years of experience, physical and technical development, whereas an 18-year-old freshman averaging the same stats has a much bigger window of improvement due to their lower baseline. Without an age feature, the model cannot distinguish between these cases.

To demonstrate that college production is indeed correlated with players' age, four composite college metrics were examined (BPM, OBPM, DBPM and PORPAG). These metrics aggregate multiple box-score components into a single all-in-one metric, like the NBA's all-in-one RAPM which serves as the target feature of this study. The results can be seen in Figure 3.8.

Player age explains a small proportion of the variance in college production, with R^2 values of 0.071 for OBPM, 0.077 for PORPAG, and 0.029 for BPM. This confirms that older college players are expected to perform better than those with less experience and that statistics carry different information depending on the player's age.

DBPM presents a nearly flat trend, with an R^2 of 0.004, indicating virtually no correlation with age in this dataset. This is consistent with findings from Engelmann (2026a), who observed a similar pattern, where age explains points, rebounds, assists, steals and 3P% but not blocks. It is worth noting that the formula for defensive BPM is simply BPM-OBPM (Myers, 2020). Another metric that better captures defense may have a stronger correlation with player age, given the previously mentioned findings where age has a strong correlation with steals.

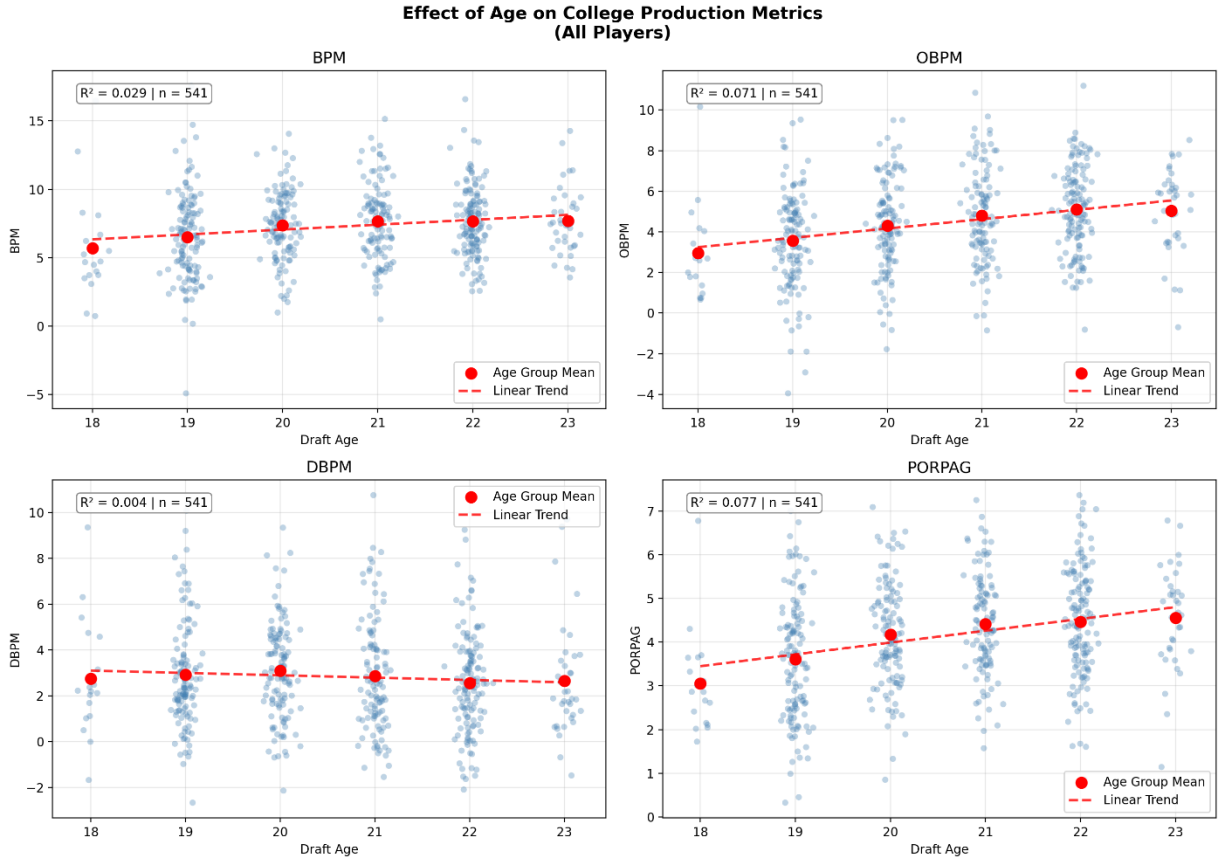


Figure 3.8 Correlation between age and all-in-one metrics (BPM, OBPM, DBPM, PORPAG)

Draft age for each player was computed using the following formula:

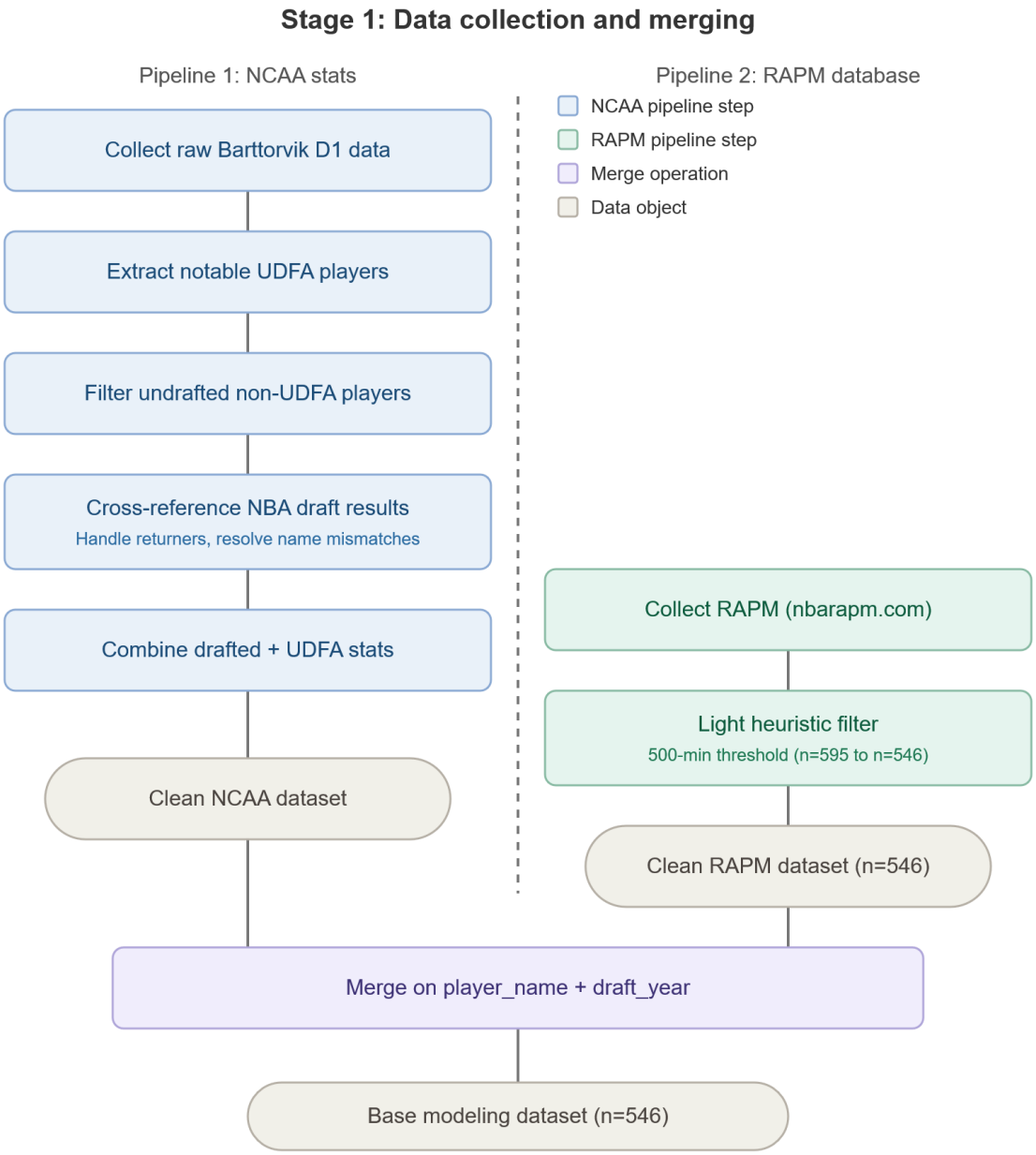
$$\text{Draft Age} = \frac{\text{Draft Date} - \text{Date of Birth}}{365.25}$$

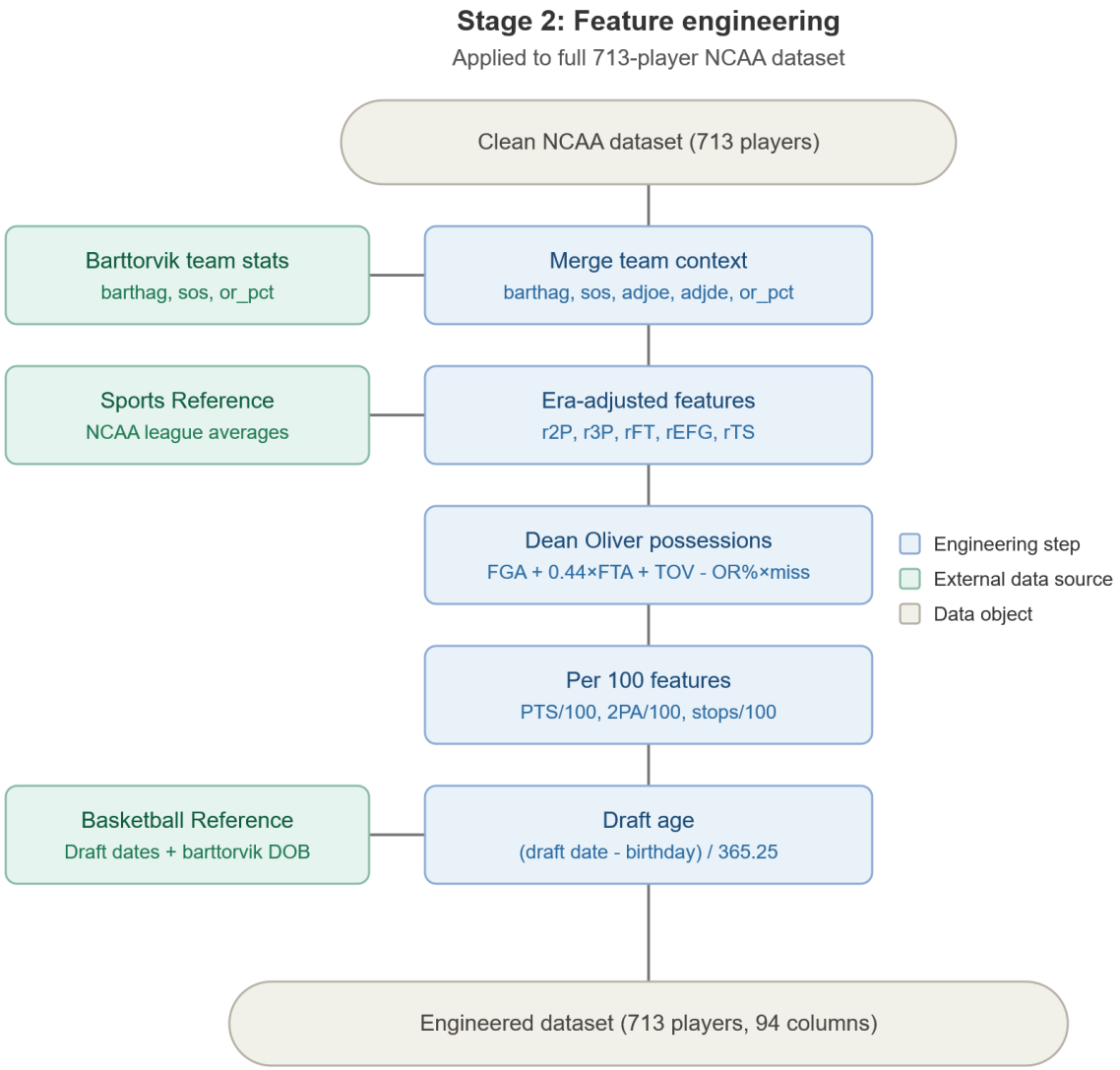
Date of birth was sourced directly from the barttorvik player dataset, and exact draft dates for each year were manually collected and verified from basketball-reference.com's individual draft pages.

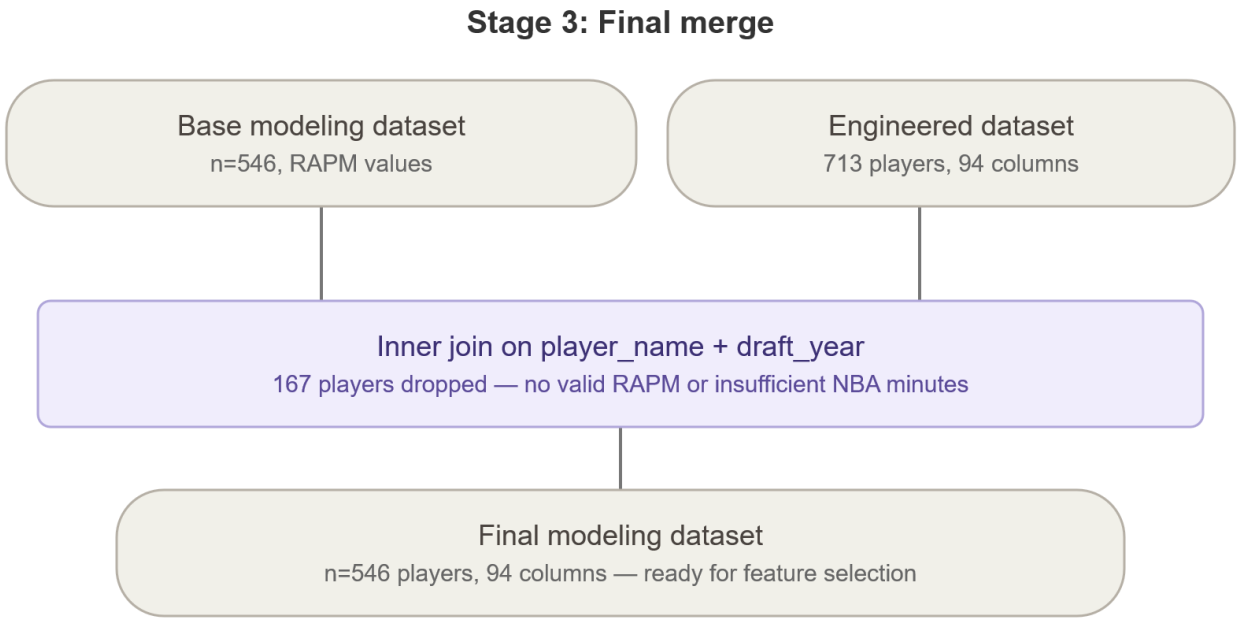
It should be noted that, according to this study's pre-defined rookie seasons, a small number of players had their NBA debut a year or two after draft day, and their draft age does not necessarily represent their age at the first NBA appearance. This distinction is unlikely to affect the model given the small number of affected players; however, it is acknowledged for transparency.

3.5. Data Pipeline Flowchart

The summary of previous steps is illustrated in the data pipeline flowchart figure below. The link to the final dataset can be found in Appendix C.




Stage 2: Feature engineering
Applied to full 713-player NCAA dataset
Clean NCAA dataset (713 players)
Barttorvik team stats
barthag, sos, or_pct
Merge team context
barthag, sos, adjoe, adjde, or_pct
Sports Reference
NCAA league averages
Era-adjusted features
r2P, r3P, rFT, rEFG, rTS
Dean Oliver possessions
FGA + 0.44×FTA + TOV - OR%×miss
Per 100 features
PTS/100, 2PA/100, stops/100
Basketball Reference
Draft dates + barttorvik DOB
Draft age
(draft date - birthday) / 365.25
Engineered dataset (713 players, 94 columns)
Engineering step
External data source
Data object


Stage 3: Final merge
Base modeling dataset
n=546, RAPM values
Engineered dataset
713 players, 94 columns
Inner join on player_name + draft_year
167 players dropped — no valid RAPM or insufficient NBA minutes
Final modeling dataset
n=546 players, 94 columns — ready for feature selection

Chapter 4 Methodology

4.1. Problem Framing

This study frames NBA rookie impact prediction as a supervised regression problem. Each player is represented by a feature vector of pre-draft NCAA statistics from their final college season, as described in Chapter 3. The output is a single continuous value, the player's RAPM throughout their rookie window, as defined in Section 3.3.

The choice of regression over classification reflects the nature of the target variable. As implied by the formulation of Regularized Adjusted Plus-Minus, RAPM is a continuous metric that captures players' impact on a full spectrum, from significantly negative to significantly positive. Classifying players into categories such as "bust" or "star" would discard the magnitude of players' impact. The difference between a player whose RAPM is -0.5 and one whose RAPM is -2.0 is practically significant, even though both might fall into the same category. Additionally, predicting the RAPM value allows for player comparisons and simulated draft rankings described in Section 4.7.

It is important to note that the main purpose of this study is to determine whether predictive signal exists in pre-draft NCAA statistics and identify which features carry the most signal, rather than to produce a scouting tool.

4.2. Train/Test split

The dataset was split temporally into a training set consisting of draft classes from 2008 to 2018 (422 players) and a test set consisting of draft classes from 2019 to 2021 (124 players), representing an approximate 77/23 split. A temporal split was chosen over random splitting primarily because the 2019-20 season was the first season played under the extended NCAA three-point line, making the test set meaningfully different from the training data and providing a more rigorous test of the model's ability to generalize across rule changes. A temporal split also preserves the integrity of the draft class comparison described in Section 4.7, which evaluates model rankings against consensus draft order over the same years. A random split would distribute players from the same draft years across both the training and test sets, making this comparison impossible.

4.3. Feature Selection

The feature selection process followed a multi-stage protocol designed to maximize model interpretability by avoiding collinear features and minimizing noise. The 70 candidate features were reduced to 20 final modelling features through domain expertise and redundancy analysis.

4.3.1 Domain-Based Exclusions

The first stage of the protocol involved excluding features based on domain knowledge and data quality considerations:

Composite metrics such as Box Plus-Minus variants (BPM, OBPM, DBPM, GBPM, OGBPM, DGBPM), offensive and defensive ratings (ORTG, AdjOE, DRTG, AdjDE), and PORPAG were excluded from the final feature list as they are derived from the same raw statistics used in the model.

Raw counting stats, such as points, rebounds, assists, steals per game were discarded in favor of per-100 possession rates.

Player anthropometric data was excluded on data quality grounds. Research by Hausch (2020) found that collegiate athletes tend to report themselves as taller and lighter than they were when officially measured, indicating systematic bias in self-reported height and weight data. Although NCAA roster measurements are typically recorded by coaching staff rather than players themselves, there is evidence suggesting similar inaccuracies in official listings. For example, an ESPN analysis comparing heights listed in college against official NBA combine measurements revealed that players

were often listed 1-2 inches taller than their actual measured height (Medcalf, 2014). While the NBA does measure players at the combine, participation was voluntary until recently, meaning not all players have recorded measurements.

Categorical position labels (Wing F, Pure PG, Wing G, Combo G, C, PF/C, Stretch 4, Scoring PG) were excluded, as they are arbitrary and fail to capture a player's actual functional role. For example, Pascal Siakam is listed as a center, even though in reality he was a Power Forward. Instead, a player's role is implied from their statistical profile, allowing the models to identify roles through statistical interactions like USG% and AST%.

4.3.2 Feature Redundancy Analysis

A correlation matrix was computed for the remaining candidate features and visualized as a heatmap to identify pairs exceeding the threshold of |r| > 0.80. Relative True Shooting (rTS%) and relative Effective Field Goal (rEFG%) exhibited high correlation (r = 0.93). Both features were removed in favor of the underlying shooting components (r2P%, r3P% and rFT%). This follows Engelmann (2026b), who suggested that 2P% and FT% are more stable predictors of professional translation. FTA/100 and FTR were found to be highly redundant (r = 0.83). FTR was kept as it is inherently adjusted for a player's field goal volume, making it a cleaner measure of player's ability to draw fouls, regardless of their usage. And while USG% and PTS/100 share a high correlation (r = 0.88), removing either of these features does not make basketball sense, as USG% measures player's workload and PTS/100 measures the productivity. An argument could be made that the efficiency stats provide enough context for PTS/100, however efficiency alone does not account for offensive burden a player might carry, or their gravity which might make them more resilient to double teams. Because the final modelling pipeline utilises ElasticNet (L1 and L2 shrinkage) and XGBoost, models that can handle correlated features, the presence of both should not cause any problems. That said, adding anthropometric features alongside the existing set would have introduced many additional predictors relative to the training sample size (n = 422), which can inflate variance and hurt generalisation regardless of model choice. The resulting correlation matrix for the finalized feature set is presented in Figure 4.1, illustrating the relationships between the selected predictors and the RAPM target variable, and Table 4.1 lists the final set of features ready for modelling.

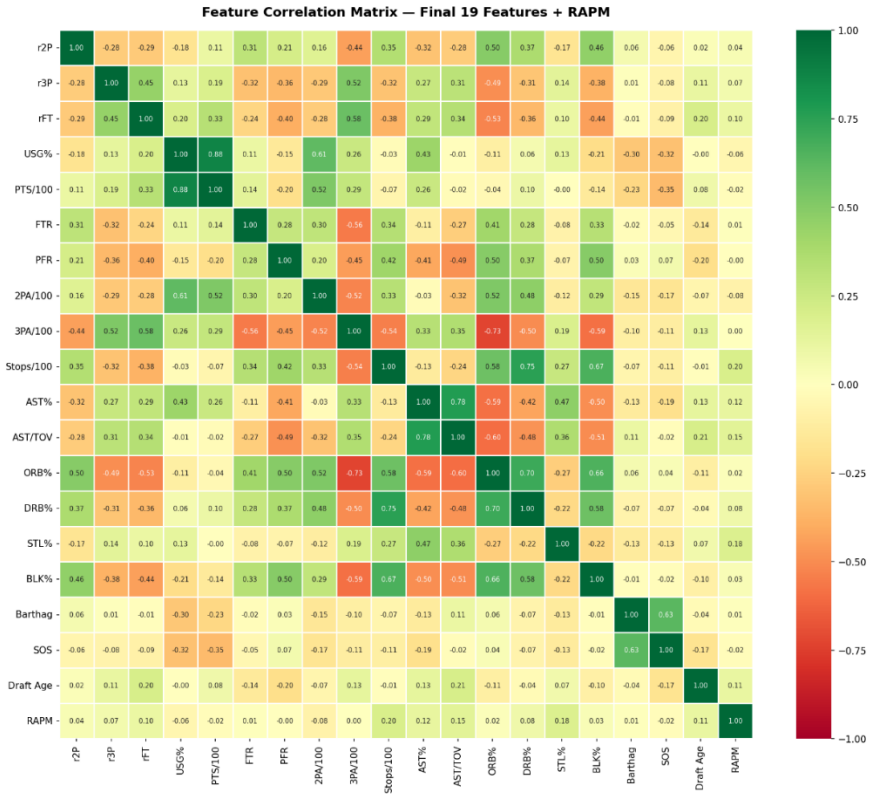


Figure 4.1: Correlation Matrix Heatmap for the final feature set and rookie window RAPM, with values ranging from -1.0 (strong negative correlation, red) to +1.0 (strong positive correlation, green).

Table 4.1: Final 20-Feature Modelling Set

Category	Feature	Description
Efficiency	r2P	2-point % vs league avg.
	r3P	3-point % vs league avg.
	rFT	Free Throw % vs league avg.
	AST/TOV	Assists to Turnovers ratio
Usage & Volume	USG%	Possession Usage Rate
	PTS/100	Points scored per 100 possessions
	2PA/100	2-point attempts per 100 possessions
	3PA/100	3-point attempts per 100 possessions
	AST%	% of teammate field goals a player assists
	FTR	Free Throw Rate
Rebounding	ORB%	% of offensive rebounds a player gets
	DRB%	% of defensive rebounds a player gets
Defense	Stops/100	Defensive stops per 100 possessions
	STL%	Steal %
	BLK%	Block %
	PFR	Fouls Comitted per 40 minutes
Team & Individual context	Barthag	Team quality
	SOS	Strenght of Schedule (incl. non-conference)
	Draft Age	Player's age at the time of the NBA Draft
	Total Minutes	Player's total minutes played in a season

4.4. Modelling

The primary goal of the modelling phase was to obtain the most accurate out-of-sample prediction of rookie RAPM from college statistics, as well as identify which statistics carry the most signal for that prediction, so that the resulting model can provide the additional value of being interpretable in terms of recognizable basketball skills, as opposed to being just a black box trained to predict players' NBA impact. With these goals in mind, two models were selected for this task: ElasticNet is a regularized linear regression that combines L1 (lasso) and L2 (ridge) penalties. The L1 regularization performs feature selection by shrinking the less important features to exactly zero, while the L2 component shrinks the coefficients by adding a penalty proportional to the sum of squared weights to the model's loss function. Together, they reduce overfitting and improve the model's interpretability and generalization on unseen data. The second model is XGBoost (Extreme Gradient Boosting), which is an ensemble of gradient-boosted decision trees. Unlike ElasticNet, XGBoost can capture nonlinear feature effects and interactions through tree splits, at the cost of less interpretability. With a small training sample (n=422), tree-based methods are prone to overfitting, so careful hyperparameter tuning is essential. Using two models allows a falsifiable test of whether tree-based nonlinearity meaningfully improves predictive power of college statistics and whether the two models agree on which statistics carry signal.

4.4.1 Baseline Model Training and Hyperparameter Tuning

Both models, ElasticNet and XGBoost, were trained on the 20-feature baseline set, which can be found in Section 4.3. Hyperparameters for each were selected through 5-fold cross validation on the training data only (2008-2018, n=422), with the test set (2019-2021, n=124) held out for final evaluation. The cross validation was done with scikit-learn's `GroupKFold` scheme which grouped players by draft year, ensuring that each fold held out an entire year of players and prevented temporal leakage between training and validation folds. All input features were standardized to zero mean and scaled to unit variance using sklearn's `StandardScaler`.

ElasticNet hyperparameters were tuned using scikit-learn's `ElasticNetCV`, which performs an internal grid search over the lambda regularization strength (scikit-learn's `alpha` parameter) and L1 mixing ratio (scikit-learn's `l1_ratio` parameter, the proportion of the penalty allocated to the L1 component). The search grid consisted of 50 log-spaced `alpha` values between 0.0001 and 100,

combined with 9 evenly spaced `l1_ratio` values between 0.1 and 0.9, yielding 450 candidate configurations. The selected hyperparameters were `alpha = 0.0655` and `l1_ratio = 0.9`, indicating a strongly Lasso-dominated penalty. On the held-out test set, this baseline ElasticNet scored an R^2 of 0.0714, a Spearman rank correlation of 0.2793 with actual rookie RAPM, an RMSE of 2.236 and an MAE of 1.753. Training R^2 was 0.1323.

XGBoost hyperparameter selection was done with scikit-learn's `RandomizedSearchCV`, which, in contrast to `ElasticNetCV`, evaluates a fixed number of randomly sampled configurations from user-defined ranges, rather than systematically evaluating all predefined combinations. The selected hyperparameters for the baseline XGBoost model, as well as the hyperparameter search space can be found in Table 4.2.

Table 4.2: XGBoost hyperparameters search space and selected optimal values for the model from 500 configurations combined with 5 cross-validation folds (2,500 model fits).

Hyperparameter	Range (in python)	Selected Value
max_depth	randint(2, 6)	4
learning_rate	uniform(0.005, 0.095)	0.0184
n_estimators	randint(100, 800)	172
subsample	uniform(0.6, 0.4)	0.6615
colsample_bytree	uniform(0.6, 0.4)	0.813
min_child_weight	randint(1, 20)	14
reg_lambda	uniform(0, 10)	0.2289
reg_alpha	uniform(0, 5)	1.6126
gamma	uniform(0, 2)	1.4428

On the held-out set, XGBoost achieved an R^2 of 0.0827, a Spearman rank correlation of 0.3095, an RMSE of 2.222, and an MAE of 1.766. Training R^2 was 0.4919, which is significantly higher than the test value. This gap reflects the model's natural tendency to fit the training-set patterns more closely than a linear model would. Whether either of these baseline models can be improved by augmenting the feature set is examined in the next section.

4.4.2 Feature Set Search

The baseline models established in Section 4.4.1 used a small 20-feature set of college statistics and demonstrated limited predictive capability, as evidenced by their low R^2 scores. To determine whether the models' predictive power was capped due to a suboptimal feature set, an exhaustive feature search was performed for both ElasticNet and XGBoost. Eight additional candidate feature groups were defined, each containing thematically related features that could plausibly contribute to predictive signal beyond the baseline feature set. The groups and their feature members are listed in Table 4.3.

Table 4.3: List of additional groups and their features that were evaluated on top of the 20-feature baseline set

Group	Features	Count
Composite metrics	porpag, dporpag, gbpm, ogbpm, dgbpm	5
Advanced shooting stats	r_ts_per, r_efg_per	2
Raw shooting stats	ts_per, efg, ft_per, twop_per, tp_per	5
Raw boxscore stats	pts, ast, oreb, dreb, treb, stl, blk, stops	8
Advanced team stats	team_adjoe, team_adjde, team_or_pct	3
Anthropometric Measurements	ht_inches	1
Lineup stats	ortg, drtg, adjoe, adrtg	4
Playing time	gp, mp, min_per	3

The search ranged over all $2^8 = 256$ possible on/off combinations of the eight additional groups, with the all-zero combination corresponding to baseline alone. For each candidate combination, hyperparameters were re-tuned from scratch using the same cross-validation protocol described in Section 4.4.1 (5-fold `GroupKFold` on draft year). The two models differed in their per-combination search capacity. ElasticNet was tuned with `ElasticNetCV` over the same 50 x 9 = 450

`alpha/l1_ratio` grid used for the baseline. XGBoost was tuned with `RandomizedSearchCV` over the same parameter distribution used for the baseline, however, the number of sampled configurations was reduced from 500 to 20 per combination, yielding 256 x 20 x 5 = 25,600 total model fits across the full search. The configuration count was reduced for computational tractability, as 500 configurations per combination would have yielded 640,000 fits. Despite the reduced configuration sampling, the procedure still identified clear performance patterns.

Across the 256 combinations examined by ElasticNet, cross-validated R^2 values were narrowly distributed (mean = 0.0895, std = 0.0049, min = 0.0785, median = 0.0914, max = 0.0978). The best combination was baseline + composite metrics + raw box-score stats + playing time (36 features), with a CV R^2 of 0.0978. The CV improvement over the baseline-only feature set (CV R^2 = 0.0888) was +0.0090. At face value, this suggests that the additional features provide a modest but real benefit. To further verify whether this CV gain reflected actual signal rather than fold-specific noise, the CV best combination was retrained on the full training set with its own tuned hyperparameters and evaluated on the held-out test set. The test R^2 was 0.0537, a decrease of 0.018 compared to the baseline test R^2 of 0.0714. The CV best combination's test Spearman correlation was 0.2925, which is higher than baseline's 0.2793. This suggests that the additional features may have captured slightly better ranking order at the cost of the point-prediction accuracy.

XGBoost produced a similarly narrow CV R^2 distribution (mean = 0.0658, std = 0.0076, min = 0.0492, median = 0.0658, max = 0.0881). The CV best combination was baseline + advanced shooting stats (22 features) with a CV R^2 of 0.0881. The CV improvement over the baseline-only feature set (CV R^2 = 0.0860) was just +0.0021, substantially smaller than the ElasticNet search's improvement. Baseline-only feature set ranked second best out of all 256 combinations, right behind the baseline + advanced shooting stats combination. When the CV best combination was retrained on the full training set and evaluated on the held-out test set, the test R^2 was 0.0751, a dip of 0.008 compared to baseline's 0.0827. The test Spearman correlation dropped as well, from 0.3095 to 0.2868.

These findings suggest that the modest R^2 values are not the result of suboptimal feature selection, as additional features did not improve generalization. Possible explanations for the performance ceiling are discussed in Chapter 5. Full ranked tables of all 256 combinations for both models are provided in Appendix C.

4.4.3 Feature Importance Comparison

To determine which baseline features contributed most strongly to the outputs of each model, a feature importance analysis was performed. Interpreting the features of an ElasticNet regression is straightforward, as each coefficient has both a sign (positive or negative) and a magnitude. A positive coefficient means that the corresponding feature tends to increase the predicted outcome, whereas a negative sign indicates a decrease in the prediction. The magnitude of a coefficient reflects the strength of the feature's influence, larger coefficient values have a stronger impact on the prediction, and those features that ElasticNet considers uninformative are assigned a coefficient of exactly zero. As can be seen in Figure 4.2, ElasticNet kept only 8 features for this regression task.

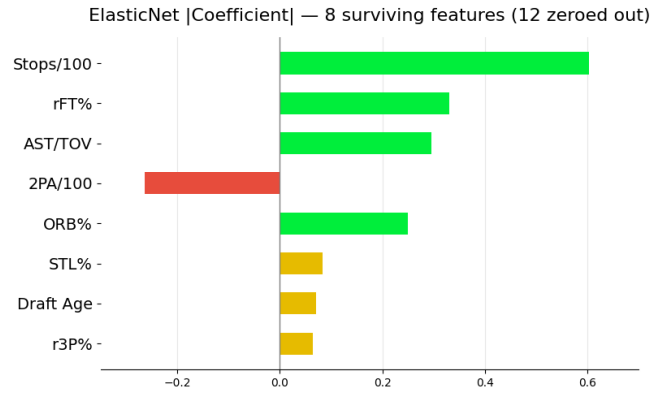


Figure 4.2: ElasticNet coefficients sorted by their absolute magnitudes

Defensive stops per 100 possessions is in a tier of its own when it comes to projecting NBA impact, there are three implicit ways to get a defensive stop; steal the ball, draw an offensive foul or block a shot that your team recovers. It makes sense that being good at causing turnovers, especially live ball ones that allow you to run in transition and preventing second chance opportunities would translate to RAPM, which is an inherently possession-based metric. The second most important feature is rFT%, which aligns with a common view that it is a stronger indicator of NBA shooting than college 3P%. This is followed closely by assists to turnovers ratio, which reflects the ability to generate high quality looks for your teammates while maintaining strong ball security. 2-point attempts per 100 have a negative impact on RAPM, suggesting that too much 2-point shot making reliance is a negative trait. Generating “extra possessions” from Offensive Rebounds is the 5th most important skill. The next feature is Steal %, which despite a marginal drop in its coefficient magnitudes, still seems to have some signal. A steal reinforces the previously mentioned idea of defensive stops, however the fact that it was not zeroed out despite being one of the three components of defensive stops warrants a deeper dive into what’s happening. The overall correlation between Steal % and Stops/100 across the full dataset is just 0.273. However, stratifying by player position group reveals within group correlations of 0.786 (guards), 0.700 (wings) and 0.543 (bigs), substantially higher than the aggregate level correlation, as can be seen in Figure 4.3.

Figure 4.3: Relationship between Steal % and Stops/100 possessions. The top panel shows the overall sample, while the bottom panels display the relationship by position (Guard, Wing, Big). Dashed lines represent linear fits; Pearson’s r and sample sizes (n) are shown in each panel.

This pattern could be explained by the fact that in basketball different positions usually have different defensive assignments, guards and wings accumulate stops through steals, whereas bigs do so through blocks. ElasticNet likely retained Steal % as it captures the live-ball turnovers that result in successful transition offense that Stops/100 alone does not fully distinguish. The last two remaining features are players draft age and r3P%, with older age implying easier initial transition to the NBA due to more maturity and better physical development. r3P% suggests that players who shot well in college should bring more immediate impact, as they do not have to develop their shot after being drafted.

XGBoost on the other hand, does not have such single-number summary. The XGBoost library has three different feature importance metrics, each measuring a different aspect of feature use; gain is the most common metric for XGBoost, it is the average reduction in loss across splits involving the feature. Weight is a frequency-based metric that reports the number of times a feature is used as a split point across all trees. And cover is the average number of training samples affected by splits on the feature. Table 4.4 reports all three metrics, their values and individual ranks; the features are sorted by their gain.

Table 4.4: XGBoost feature importance metrics comparison (Gain vs Weight vs Cover).

Feature	Gain	Weight	Cover	Gain Rank	Weight Rank	Cover Rank
stops_100	33.826	122	167.541	1	1	1
ast/tov	29.174	103	151.136	2	2	4
stl_per	28.011	46	157.544	3	18	3
r_ft_per	25.384	102	166.402	4	3	2
ast_per	25.289	61	122.279	5	11	11
drb_per	23.651	52	143.423	6	14	5
usg	22.904	61	135.820	7	11	8
pfr	22.791	32	115.563	8	20	14
orb_per	22.697	95	137.800	9	4	6
twopa_100	22.571	47	115.660	10	17	13
total_minutes	22.072	62	117.726	11	10	12
blk_per	21.820	33	115.455	12	19	15
ftr	21.531	79	79.709	13	6	20
r_2p_per	20.729	74	131.365	14	8	9
draft_age	20.253	77	137.065	15	7	7
team_sos	19.677	52	103.423	16	14	17
r_3p_per	19.673	50	93.420	17	16	19
pts_100	19.401	81	131.111	18	5	10
threepa_100	16.858	56	93.625	19	13	18
team_barthag	16.456	63	107.683	20	9	16

All three XGBoost metrics agree with ElasticNet on stops being the most predictive college statistics, they also broadly agree on assists/turnovers and rFT%. Beyond these three statistics, however, the metrics diverge in their opinions. The most notable example is Steal %; it ranks just 18th in weight as it was used in just 46 tree splits, yet despite that ranks 3rd in gain and cover, meaning those 46 splits were highly informative. The opposite case is PTS/100, which ranks 5th by weight but only 18th by gain, meaning XGBoost split on it frequently yet the splits did not contribute that much. These disagreements reflect the fact that a feature can be used heavily during training, yet contribute very little to actual predictions, or vice versa. The important caveat here is that these metrics describe how features were used during the training phase, none of them explain how each feature contributes to the prediction.

To resolve this, SHAP (SHapley Additive exPlanations) values were computed on the training set. SHAP values attribute each feature's contribution to the difference between the model's prediction for a given player and the average prediction across the dataset (Lundberg & Lee, 2017). They are computed as the average marginal contribution of each feature across all possible feature combinations, and the result is a signed value expressed in the same units as the prediction itself. SHAP values can be computed exactly for tree-based models using the `TreeExplainer` algorithm (Lundberg, et al., 2020), which exploits the tree structure to compute exact SHAP values in polynomial time. When SHAP values are aggregated, they provide interpretation at both global and individual level. Taking the mean of the absolute SHAP values over the training data gives a single importance score per feature, like the absolute magnitude of an ElasticNet coefficient. For an

individual player, the SHAP values can break down how exactly each feature contributes to their predicted RAPM. Table 4.5 reports each feature's mean absolute SHAP value alongside its rank.

Table 4.5: Baseline features ranked by their mean absolute SHAP values.

Rank	Feature	Mean |SHAP|
1	stops_100	0.3091
2	ast/tov	0.2599
3	r_ft_per	0.2336
4	orb_per	0.1696
5	ast_per	0.1196
6	stl_per	0.1121
7	r_2p_per	0.0940
8	usg	0.0918
9	twopa_100	0.0897
10	draft_age	0.0721
11	pts_100	0.0660
12	r_3p_per	0.0615
13	drb_per	0.0602
14	threepa_100	0.0593
15	ftr	0.0564
16	total_minutes	0.0498
17	team_barthag	0.0476
18	team_sos	0.0466
19	pfr	0.0379
20	blk_per	0.0313

The SHAP values closely resemble the ElasticNet coefficients shown in Figure 4.2. To quantify the agreement, the absolute ElasticNet coefficients and the mean absolute SHAP values were compared directly. Comparing the 8 features that ElasticNet retained, the rank correlation between the two importance orderings is 0.90 (Spearman ρ), indicating strong agreement between the two methods, as can be seen in Figure 4.4.

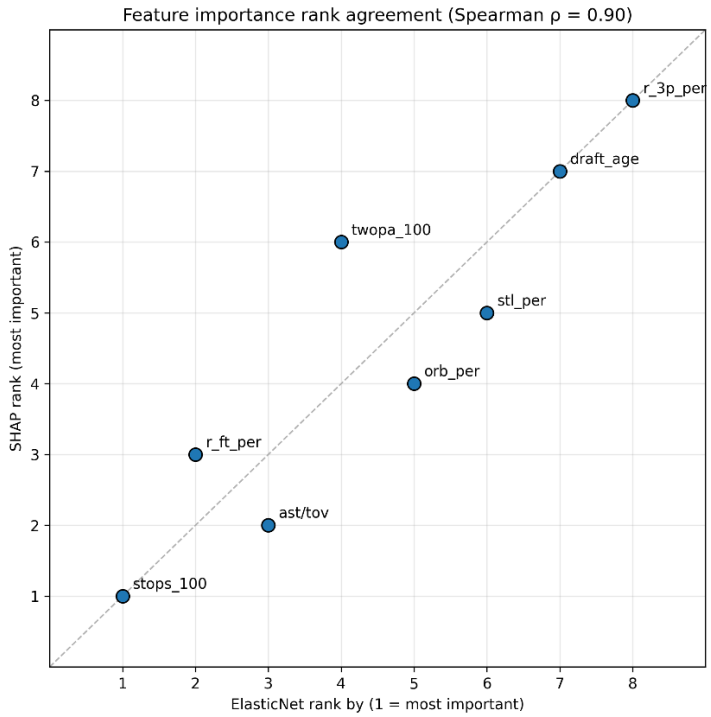


Figure 4.4: Relationship between ElasticNet and SHAP feature importance rankings.

The mean absolute SHAP, however, only captures the magnitude of each feature's contribution. To recover the direction in which features push the prediction, Figure 4.5 displays the SHAP summary plot, where each point represents a player from the training set. The horizontal axis shows the players' signed SHAP value for that feature, where positive values push the predicted RAPM up, while negative values push it down. The color encodes the players' values for those features, red representing high values and blue representing low. Features are stacked by their mean absolute SHAP values.

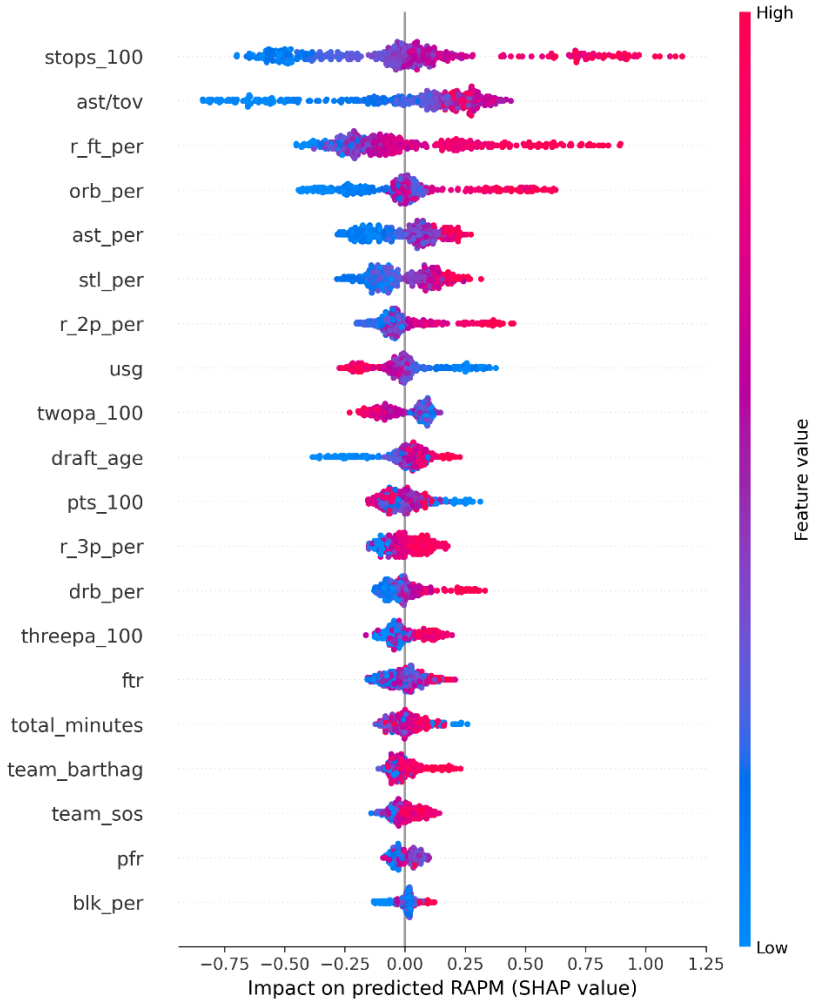


Figure 4.5: SHAP summary plot. X-axis shows SHAP value (impact on prediction), y-axis shows features ranked by importance. Point color indicates feature value (red = high, blue = low)

The summary plot also enables observations that are otherwise invisible in magnitude only rankings. For example, high Usage % values (red points) tend to negatively affect predicted RAPM, this phenomenon could be explained by the fact that most high-usage college players are point guards, and their transition to the league is notoriously harder compared to low-usage role players, whose role is usually spotting up in the corner to shoot threes, which also allows them to preserve more energy for defense.

Despite the structural differences between ElasticNet, which is an additive linear regression and XGBoost, an ensemble of nonlinear gradient-boosted trees, the two models agree on which college statistics carry the strongest predictive signal for rookie window RAPM. This convergence in the opinions suggests that despite the modest predictive accuracy reported earlier, machine learning might still have practical applications in basketball. Teams and front offices could use such models to enable more efficient and data-driven talent scouting, by analyzing which KPIs distinguish high-impact prospects from less impactful ones. Another use case of these models is that players can make informed decisions on which areas to improve in, as well as how to adjust playstyles so that better align with the skills that translate to the NBA.

4.4.4 Comparison Against Pre-Draft Consensus

As reported in Section 4.4.1, the two models explain approximately 7-8% of the variation in rookie window RAPM. Although the number is quite low, the feature importance analysis performed in Section 4.4.3 demonstrated that both models still identified a coherent and meaningful set of features as the strongest predictors of rookie window RAPM. This raises the question: can these models rank prospects more accurately than the media consensus produced by journalists, analysts and scouts?

To answer this question, a comparison between the models' predictions and the media consensus was performed by using the trained ElasticNet and XGBoost models to generate predictions on the held-out test set (2019-2021), ranking the players by their predicted RAPM and comparing the model-generated rankings against the consensus big board, with actual rookie window RAPM serving as the ground truth. The consensus pre-draft ranking was sourced from the consensus big board published by Lichtenstein (n.d.), which compiles overall rankings from twelve major NBA outlets (The Athletic, Bleacher Report, CBS Sports, ESPN, Hoops Prospects, NBA Draft Net, The Ringer, NBA Scouting Live, Sports Illustrated, Sporting News, Yahoo, and 247 Sports) into a single big board. The boards for the 2019, 2020 and 2021 draft classes were retrieved using the year selector on the same page. The comparison was restricted to players appearing in both the held-out test set and the consensus big board, resulting in 111 players across the three years (n = 42, 32, 37 for 2019, 2020, and 2021 respectively). The excluded players are a mix of international players who did not play in the NCAA and players who did not meet the 500-minute NBA playing-time threshold. The point of this comparison is to determine which method orders a pre-defined list of likely-to-be drafted players more accurately.

The agreement of each method with realized RAPM rankings was measured using the Spearman rank correlation, computed for each draft class individually, with a pooled value obtained by averaging the three per-year correlations. As shown in Table 4.6, both models meaningfully outperform the consensus big board across the pooled comparison, with XGBoost achieving the highest pooled correlation ($\rho = 0.265$), followed by ElasticNet ($\rho = 0.233$), with consensus trailing them far behind at $\rho = 0.050$. However, the individual-year correlations reveal that the 2020 draft class proved difficult for all three methods, with the consensus big board outperforming both models that year ($\rho = 0.063$ vs 0.017 for ElasticNet and -0.026 for XGBoost). This may reflect the disruption COVID-19 caused to the 2020 NCAA season, which shortened schedules, cancelled the tournament, and introduced unusual circumstances that the models were not equipped to account for.

Table 4.6: Spearman rank correlation between each method's player ordering and the actual 4-year rookie window RAPM ordering

Year	n	Consensus ρ	ElasticNet ρ	XGBoost ρ
2019	42	0.067	0.253	0.334
2020	32	0.063	0.017	-0.026
2021	37	0.020	0.429	0.488
Pooled	-	0.050	0.233	0.265

This finding reinforces the conclusion drawn in Section 4.4.3 that, despite their limited predictive power, machine learning models still provide valuable input beyond what the current consensus offers.

Chapter 5 Discussion

5.1. Interpretation of Results

The main empirical result of this study is that pre-draft NCAA statistics carry a real but limited signal for projecting rookie window RAPM. Both models explained roughly 7-8% of the variance in the held-out test set. Augmenting the baseline features with eight additional thematic groups did not boost the performance, suggesting that the performance ceiling has several contributing causes. For example, rookie RAPM is still a noisy target variable due to a limited number of possessions some rookies play, factors like injuries, coaches' trust or player development curves cannot be accounted for by college statistics. Another plausible reason is that the training dataset is not large enough, and more years worth of data would improve the model.

What the models did recover, however, was a coherent and sensible ordering of which college skills carry signal. Defensive stops per 100 possessions emerged as the single strongest predictor across both models, which is intuitive given that RAPM is a possession-based metric and getting stops on the defensive ends shaves a lot of points off your opponent. Being a good free throw shooter is also seemingly more important than shooting well from the 3-pt line, which lines up with the notion that free-throw shooting is a more stable indicator of prospect's shooting talent than college three-point percentage. Both models agree that high two-point volume, ball-dominant college players translate less reliably than efficient, lower usage role players. This finding aligns with the well-documented difficulty that high-usage college guards face when adjusting to the NBA.

5.2. Cross-Model Agreement

Because of the fundamental differences between ElasticNet and XGBoost, as one is an additive linear regression, and the other is an ensemble of non-linear trees capable of capturing interactions, the convergence in their findings strengthens confidence in the robustness of the results. Despite their contrasting differences, both models produced nearly identical predictive performance, with the added non-linearity of XGBoost providing no significant difference in out-of-sample accuracy over the linear model. Moreover, the models showed strong agreement regarding which features carried predictive signal, evidenced by a Spearman rank correlation of 0.90 between the absolute ElasticNet coefficients and the mean absolute SHAP values across the eight retained features. If the more flexible XGBoost model had meaningfully outperformed ElasticNet, it would suggest the linear approach was limited by its inability to exploit non-linear interactions. Instead, a relatively simple linear model appears capable of extracting nearly as much predictive power from the publicly available college statistics as its non-linear alternative. The convergence between the models also strengthens the confidence in the feature importance findings, as predictors identified independently by two fundamentally different models are less likely to be artefacts specific to either model.

5.3. Comparison to the Literature

The findings of this study both align with and differ from the prior work reviewed in Chapter 2. The low R^2 values of both models echo Greene (2015), whose regression onto box-score metrics explained only a small share of variance, the WS/48 model reaching just 0.11. Another parallel is that both models covered in this study seem to outperform the media consensus when it comes to ranking players, Greene found similar results, where his WS/48 model outperformed some of the actual NBA drafts, despite its low R^2. In both cases, modest R^2 scores coexist with useful ordinal signal, suggesting that high explained variance is not required for a practical prospect ranking.

Where Greene (2015) relied on unadjusted raw statistics spanning two decades, this study went through the process of adjusting them for different eras and standardizing for pace, and Chapters 3.4.1 and 3.4.2 provide evidence that it is good practice to put the stats under the same scale when comparing players from different seasons.

Unlike Kannan et al. (2018), this study did not use the draft position as a feature, which has important implications for how the results should be interpreted. The draft position reflects aggregated expert judgment rather than independent college performance. By omitting it entirely, the model reflects the signal contained within the actual player performance, as opposed to trying to predict how the NBA teams will think at the draft. This makes the model useful for making its own player rankings, that may have an edge over the consensus.

5.4. Practical Applications

The primary purpose of this study was not to construct a finished scouting system, but to determine whether the predictive signal exists in NCAA statistics, and identify which features carry the signal. While the models are not sufficiently accurate for precise point prediction of a prospect's future rookie RAPM, as the modest R^2 values indicate a lot of unexplained variance, the models still have some practical use cases, as both models ranked prospects more accurately than the aggregated media consensus across the pooled comparison sample. This indicates that the models may serve as a useful supplementary input to traditional scouting processes, particularly when used to identify undervalued or overvalued prospects relative to consensus expectations. Importantly, this should not be treated as evidence that these models can replace scouting expertise, but rather they can be used as tools on top of the existing practices.

5.5. Limitations

Several limitations exist in this study. The most fundamental one being the target variable itself. Rookie window, while arguably a more defensible measure of individual impact than box-score aggregates, remains a noisy estimate, particularly for players with limited playing time. Although the 500-minute filter tackles this issue, it does not eliminate it completely. The models are attempting to predict a target that itself has a substantial measurement error, which likely caps the predictive power of these models.

Second, the sample size is modest. With 422 training observations, the data supports only limited model complexity, while the test set of 124 observations spanning three draft years can be sensitive to characteristics of individual classes. This is illustrated by the poor performance of both models on the 2020 draft class. Given the small number of evaluation years, that result should be interpreted cautiously rather than treated as a definitive finding.

Third, the study models only a player's final college season and excludes anthropometric measurements on data-quality and availability grounds. While both decisions are methodologically defensible, they also mean that potential informative signals were omitted from the modelling process.

Finally, the consensus comparison was restricted to players appearing both in the test sample and within consensus big boards, meaning the conclusions regarding model vs consensus performance apply specifically to the overlapping subset of likely to be drafted players, and not to the entire draft pool.

Chapter 6 Conclusion

This study approached the prediction of NBA rookie impact as a supervised regression problem, training both a regularized linear model and a gradient-boosting tree ensemble to predict rookie window RAPM from era-adjusted and pace-normalized NCAA statistics. Although both models explained only a modest share of the variance in rookie impact, they both identified a coherent and basketball-relevant set of predictors and converged closely on their performance and ranked prospects more accurately than the aggregated pre-draft media consensus, despite the structural differences of both models.

These findings suggest that the primary limitation on the predictive performance may stem less from the model and more from the noisiness and informational limits of available data and the target variable. At the same time the results demonstrate that interpretable predictive signal does exist within the NCAA statistics and machine learning may offer practical value on top of traditional scouting processes.

A natural direction for future research would be to evaluate alternative target variables with lower measurement noise, such as hybrid impact metrics that combine RAPM with statistical priors. Such targets may enable a more stable learning signal and potentially improve predictive performance, especially if trained on a larger sample size of players.

Chapter 7 References

Basketball Reference. (n.d.-a). *Calculating PER*. Retrieved May 20, 2026, from https://www.basketball-reference.com/about/per.html

Basketball Reference. (n.d.-b). *NBA Win Shares*. Retrieved May 20, 2026, from https://www.basketball-reference.com/about/ws.html

Costa, D. d. (2024). *A machine learning approach comparing predictive performance and interpretability of models for predicting success of NCAA basketball players to reach NBA.* Retrieved from https://dspace.sti.ufcg.edu.br/handle/riufcg/38052

databallr. (n.d.). *NBA player impact analytics*. Retrieved May 20, 2026, from https://www.nbarapm.com/

Dunks & Threes. (2020, June 17). *NBA Player Metric Comparison*. Retrieved from https://dunksandthrees.com/blog/metric-comparison

Dunks & Threes. (n.d.). *Estimated Plus-Minus (EPM) and Estimated Skills*. Retrieved May 20, 2026, from https://dunksandthrees.com/about/epm

Engelmann, J. (2026a, March 27). *How to adjust college stats — featuring Caleb Wilson vs. Cameron Boozer*. Retrieved from 5x5 | Royce Webb: https://www.roycewebb.com/p/how-to-adjust-college-stats-featuring

Engelmann, J. (2026b, April 15). *How to build an NBA draft model*. Retrieved from 5x5 | Royce Webb: https://www.roycewebb.com/p/how-to-build-an-nba-draft-model

Greene, A. C. (2015). *The Success of NBA Draft Picks: Can College Careers Predict NBA Winners?* St. Cloud, MN: St. Cloud State University.

Hausch, J. (2020). *Accuracy of self-reported height and weight in collegiate athletes.* Eastern Michigan University, Health Promotion and Human Performance. Ypsilanti, MI: Eastern Michigan University. Retrieved from https://commons.emich.edu/honors/689/

Johnson, G. (2019, June 5). *Men's college basketball 3-point line extended to international distance*. Retrieved from NCAA: https://www.ncaa.com/news/basketball-men/article/2019-06-05/mens-college-basketball-3-point-line-extended-international

Kannan, A., Kolovich, B., Lawrence, B., & Rafiqi, S. (2018). Predicting National Basketball Association Success: A Machine Learning Approach. *SMU Data Science Review, 1*(3). Retrieved from https://scholar.smu.edu/datasciencereview/vol1/iss3/7/

Lichtenstein, J. (n.d.). *NBA Draft Consensus Big Board*. Retrieved May 20, 2026, from jacklich10: https://jacklich10.com/bigboard/nba/

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems 30 (NIPS 2017).* Long Beach, CA. Retrieved from https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., . . . Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*. Retrieved from https://www.nature.com/articles/s42256-019-0138-9

Medcalf, M. (2014, May 15). *College stars shrink at NBA draft combine*. Retrieved from ESPN: https://www.espn.com/blog/collegebasketballnation/post/_/id/99002/college-stars-shrink-at-nba-draft-combine

Myers, D. (2020, February). *About Box Plus/Minus (BPM)*. Retrieved from Basketball Reference: https://www.basketball-reference.com/about/bpm2.html

NCAA. (2025). *Probability of Competing in College and Beyond*. Retrieved May 18, 2026, from NCAA: https://www.ncaa.org/sports/2013/12/17/probability-of-competing-beyond-high-school.aspx

Rosenbaum, D. T. (2004, April 30). *Measuring How NBA Players Help Their Teams Win*. Retrieved from 82games: https://www.82games.com/comm30.htm

Sill, J. (2010). Improved NBA Adjusted +/- Using Regularization. *MIT Sloan Sports Analytics Conference.* Boston, MA. Retrieved from https://supermariogiacomazzo.github.io/STOR538_WEBSITE/Articles/Basketball/Basketball_Sill.pdf

Sports Reference. (n.d.-a). *College Basketball Stats and History*. Retrieved May 20, 2026, from https://www.sports-reference.com/cbb/

Sports Reference. (n.d.-b). *Basketball Stats and History*. Retrieved May 20, 2026, from https://www.basketball-reference.com

Torvik, B. (2018, September 17). *T-Rank Methodology Update*. Retrieved from Adam's WI Sports Blog: https://adamcwisports.blogspot.com/2018/09/t-rank-methodology-update.html

Torvik, B. (n.d.). *T-Rank: Customizable College Basketball Tempo Free Stats*. Retrieved May 20, 2026, from https://barttorvik.com/

Appendix

Appendix A: Coefficient of Variation for NCAA league average statistics across 2007-08 to 2020-21

Statistic	CV	Statistic	CV	Statistic	CV
Volume		**Counting**		**Efficiency**	
FG	0.034	PTS	0.031	FG%	0.009
FGA	0.027	TRB	0.016	2P%	0.019
2P	0.019	AST	0.025	3P%	0.014
2PA	0.016	STL	0.034	FT%	0.012
3P	0.089	BLK	0.038	EFG%	0.015
3PA	0.087	PF	0.032	TS%	0.012
FT	0.050				
FTA	0.052				
TOV	0.043				

Appendix B: TOV estimation methods validations against the simplified and full possession formulas

	Simplified Formula				Full Formula			
Year	TO%_corr	TO%_mae	AST/TOV_corr	AST/TOV_mae	AST/TOV_corr	AST/TOV_mae	TO%_corr	TO%_mae
2008	0.9965	0.9476	0.9971	0.8144	0.9979	0.466	0.9974	0.6196
2009	0.9963	1.2618	0.997	1.0844	0.9976	0.6316	0.9969	0.8393
2010	0.9969	0.8608	0.9972	0.7367	0.9974	0.4106	0.9972	0.5548
2011	0.997	1.2366	0.9975	1.0722	0.9977	0.6371	0.9974	0.8264
2012	0.9969	1.1248	0.9976	0.9837	0.9981	0.615	0.9975	0.7791
2013	0.9968	1.126	0.9972	0.9766	0.9976	0.5282	0.9973	0.6992
2014	0.9962	1.1576	0.997	1.0243	0.9975	0.6464	0.9966	0.798
2015	0.9966	1.0743	0.9972	0.9426	0.9974	0.5622	0.9968	0.7096
2016	0.995	1.0403	0.9962	0.8972	0.9966	0.4872	0.9955	0.6497
2017	0.9922	1.115	0.9936	0.9818	0.9946	0.5922	0.9932	0.7448
2018	0.9933	1.2284	0.9942	1.0687	0.9954	0.5664	0.9946	0.7481
2019	0.9933	1.2851	0.9948	1.1051	0.9951	0.6342	0.9937	0.8339
2020	0.9968	1.1671	0.9974	1.0148	0.9977	0.5191	0.9972	0.6946
2021	0.9938	1.1555	0.9949	0.9942	0.9949	0.5407	0.994	0.7222
Mean	**0.9955**	**1.1272**	**0.9964**	**0.9783**	**0.9968**	**0.5598**	**0.9961**	**0.7300**

Appendix C: Supplementary Data and Materials

https://zenodo.org/records/20315721?token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjNlMzU0NDViLTcwNWMtNGMyMy1hZWM1LTRkMTcyN2M5MDU0ZCIsImRhdGEiOnt9LCJyYW5kb20iOiJjY2FmNDI2N2JmY2MxNmE3YjlhYWIyOTQwNDZiYTlmYSJ9.fbyZdWpV2As6oQFqjQEeObXRn6vGfpXjsUWPoSAaF6xvziAbPQQYc2OY18Jp7KcqqyZJzcGFAqld56SrRq_e7w