CS 559: NBA Player Points Analytics

Gabriel Castillo and Adrian Chacon
Stevens Institute of Technology
gcastill@stevens.edu, achacon1@stevens.edu
Fall 2025

Abstract

The goal of this project is to be able to properly predict the number of points a player will score against an opposing team in an NBA game on any given day. Basic statistics that are easily accessible, such as a player's average points per game, shot attempts, and two- and three-point shooting percentages, provide a general overview of scoring ability. However, these numbers alone do not capture how a player may perform in specific matchups or how strong the opposing team's defense is.

This project focuses on incorporating both individual offensive statistics and opposing team defensive statistics into a machine learning framework to better estimate player scoring outcomes. By accounting for matchup-specific factors through statistical modeling, the predictions aim to be more informative than simple averages. These results can benefit fans seeking deeper insight, coaches and staff evaluating player performance, and analysts whose work centers on understanding and optimizing basketball statistics.

1 Introduction

Predicting how many points a basketball player will score in a given game is an important problem in modern basketball analytics. Scoring is one of the most visible measures of player performance and often plays a large role in how players are evaluated by teams, analysts, and fans. While season averages give a quick snapshot of a player's scoring ability, they do not explain why a player may score significantly more or fewer points in certain games.

There are many factors that influence a player's scoring on a given night. These include how efficient the player is as a shooter, how many shots they typically take, and the defensive quality of the opposing team. Some teams are better at limiting scoring opportunities, contesting shots, or slowing the pace of the game. As a result, a player who performs well against one team may struggle against another, even if their season averages suggest otherwise.

The goal of this project is to move beyond simple averages by accounting for both player-specific offensive statistics and opponent-specific defensive statistics. By combining these factors, we aim to build a model that can better predict a player's point total in an individual game. The scope of this project focuses on statistical data that is publicly available and machine learning methods covered in class, with an emphasis on understanding how different features contribute to scoring outcomes.

2 Related Work

The use of machine learning and statistical analysis in basketball has become increasingly common in recent years. Many existing projects focus on predicting player performance using historical averages, recent game trends, or basic box score statistics. These approaches often rely on linear regression or similar models to estimate future performance based on past results.

One limitation of many existing approaches is that they primarily focus on individual player statistics while ignoring the opposing team. This can lead to inaccurate predictions when matchup difficulty varies significantly. For example, a player facing a strong defensive team may score fewer points than expected, even if they are performing well overall.

Professional analysts and odds-makers often account for match-ups, defensive strength, injuries, and other contextual factors when evaluating expected player performance. While the exact methods used by professionals are not publicly available, this highlights the importance of including opponent-related features in predictive models. This project aims to take a similar idea and apply it using transparent, data-driven methods based solely on publicly available statistic

3 Methodology

In this project, predicting the number of points an NBA player will score in a game is formulated as a supervised learning regression problem. The goal is to learn a function that maps a vector of numerical features to a continuous target values representing points scored in a single game.

Each data sample corresponds to one player-to-game instance. Let $x_i \in \mathbb{R}^d$ denote the feature vector for the i-th sample, where the features combine player offensive statistics and opposing team defensive statistics. Player features include shooting volume and efficiency metrics such as field goals attempted, three-points attempted, free throws attempted, and shooting percentages. Opponent features include team-level defensive measurements such as points allowed and opponent shooting percentages. The target variable $y_i \in \mathbb{R}^d$ represents the actual number of points scored by the player in that game.

All numerical features are standardized prior to training so that each feature has zero mean and unit variance. This is a necessary step to ensure that the features do not vary in scale and ensure stable and efficient optimization when using full-batch gradient descent as our learning method.

As a baseline approach, standard linear regression is used. The model assumes a linear relationship between the input features and the target value, given by:

$$\hat{y}_i = w^T x_i + b$$

where w is the weight vector and b is the bias term. The model parameters are learned by minimizing the mean squared error loss over the training data:

$$L(w) = \tfrac{1}{N} \textstyle\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

This optimization is performed using full batch gradient descent. At each iteration, the loss function gradients are calculated using the entire training dataset and the model parameters are updated accordingly.

The primary model used in this project is ridge regression, which extends linear regression by introducing L2 regularization on the model weights. Ridge regression minimizes the following objective function:

$$L(w) = \tfrac{1}{N} \textstyle\sum_{i=1}^{N} (y_i - w^T x_i)^2 + \lambda \|w\|_2^2$$

The regularization term $\lambda||w||_2^2$ acts as a form of weighting on the model coefficients by penalizing large weight values. This effectively shrinks the learned coefficients to zero, reducing the influence of any single feature and improving the stability of the model. In the context of basketball statistics, this weighting is important because many features are correlated and unregularized regression can produce unstable or overly large coefficients.

Ridge regression in this project is also trained using full batch gradient descent. During each update step, gradients are computed over the entire training set, and the regularization term directly influences the gradient updates by discouraging large parameter values.

4 Experimental Setup

The experiments are structured to evaluate whether incorporating opponent defensive statistics and regularization improves the accuracy of player point predictions compared to simpler baseline methods. Specifically, the experimental setup focuses on assessing how well the ridge regression model predicts player point totals for individual games and how its performance compares to standard linear regression, season averages, and betting market expectations. By evaluating model predictions against actual game outcomes, the experiments aim to determine whether the proposed approach provides a more informative and reliable estimate of scoring performance across different players and matchup.

4.1 Data

The data used in this project was collected from publicly available NBA statistics sources using a Python-based data pipeline implemented in a Jupyter notebook. The dataset includes player-level game statistics as well as team-level defensive statistics across multiple NBA seasons. Each data point represents a single player's performance in one game.

After pre-processing, the dataset contains several thousand player-game instances with 25 features per instance. Basic data cleaning was performed to remove incomplete records, and all numerical features were standardized before training to ensure fair comparisons across different scales.

4.2 Evaluation Metrics

To evaluate how well the model predicts player point totals, we use mean squared error (MSE), which is a standard evaluation metric for regression problems. MSE measures the average squared difference between the true number of points scored by a player and the number of points predicted by the model. By squaring the errors, larger prediction mistakes are penalized more heavily than smaller ones. MSE is defined as:

$$\text{MSE} = \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

A smaller MSE indicates that the model's predictions are, on average, closer to the actual point values, while a larger MSE reflects larger and more frequent prediction errors. This metric is particularly suitable for player point prediction because it captures both the magnitude and consistency of errors across many games.

In this project, MSE is computed on a held-out test set that was not used during training. This ensures that the evaluation reflects the model's ability to generalize to unseen games rather than simply memorizing past performance. MSE is used to compare the baseline linear regression model with the ridge regression model, allowing us to assess whether regularization and opponent-based features improve predictive accuracy.

4.3 Comparison Methods

To evaluate the effectiveness of the proposed machine learning models, predictions are compared against commonly used baseline methods that reflect how player scoring expectations are typically estimated in practice.

The first comparison method is a player's current season average points per game, obtained from official NBA statistics provided by the NBA official website. This baseline represents the most widely available and commonly referenced measure of scoring performance. Using the season average provides a simple benchmark that assumes a player will perform similarly across all games, without accounting for matchup-specific factors.

The second comparison method is the player point total, or "player prop" line, offered by major betting platforms. In this project, betting lines from FanDuel Sportsbook are used as a real-world benchmark. These lines are set by professional odds-makers and are intended to represent the expected number of points a player will score in a given game. Betting lines incorporate a variety of factors, including opponent strength, recent performance, and historical trends, making them a strong practical comparison point.

These comparison methods are chosen because they reflect both statistical baselines and market-based expectations that are widely used by fans, analysts, and industry professionals. By comparing ridge regression predictions against season averages and FanDuel betting lines, we can assess whether the proposed model provides additional value beyond commonly available scoring estimates.

5 Results

This section presents the results obtained from applying the ridge regression model to predict player point totals for individual NBA games. The goal of this analysis is to compare the model's estimated points against the actual points scored and to examine how prediction errors vary across different players and game contexts.

The table below summarizes a sample of player–game predictions, including the opponent, the estimated points produced by the model, the actual points scored, and the resulting differential. Positive differentials indicate that the player exceeded expectations, while negative differentials represent underperformance from the player relative to the model's prediction. These results provide a quantitative basis for evaluating how well the model captures scoring performance and for identifying patterns in prediction error.

Player Name	Opponent City	Estimated Points	Actual Points	Differential
Cade Cunningham	Boston	25	32	7
Jalen Duren	Boston	14	6	-8
Derrick White	Detroit	16	31	15
Payton Pritchard	Detroit	12	12	0
Brandon Ingram	Miami	21	28	7
Scottie Barnes	Miami	19	17	-2
Immanuel Quickley	Miami	17	15	-2
Davion Mitchell	Toronto	8	12	4
Norman Powell	Toronto	18	20	2
Bam Adebayo	Toronto	19	20	1
Cooper Flagg	Utah	18	42 *	24
Klay Thompson	Utah	16	12 *	-4
Lauri Markkanen	Dallas	23	33 *	10
Keyonte George	Dallas	16	37 *	21
Ace Bailey	Dallas	10	7 *	-3
Alperen Sengun	Denver	21	33 *	12
Kevin Durant	Denver	27	25 *	-2
Amen Thompson	Denver	13	14 *	1
Nikola Jokic	Houston	28	39 *	11
Jamal Murray	Houston	22	35 *	13
Peyton Watson	Houston	8	5 *	-3
Ja Morant	La Clippers	22	12	-10
Santi Aldama	La Clippers	12	3	-9
Jaylen Wells	La Clippers	10	16	6
Kawhi Leonard	Memphis	23	21	-2
James Harden	Memphis	21	13	-8
Donovan Mitchell	Chicago	26	32	6
Jarett Allen	Chicago	15	14	-1
Nikola Vucevic	Cleveland	18	20	2
Coby White	Cleveland	20	25	5
Josh Giddey	Cleveland	14	23	9
Matas Buzelis	Cleveland	11	9	-2
Donte Divincenzo	Memphis	14	19	5
Naz Reid	Memphis	14	16	2
Julius Randle	Memphis	21	21	0
Jaden McDaniels	Memphis	5	13	8
Santi Aldama	Minnesota	12	8	-4
Cedric Coward	Minnesota	13	13	0
Jaylen Wells	Minnesota	11	17	6

These results do vary significantly, but that should be expected when dealing with a sport in which conditions change drastically on a daily basis. What is being done here is predicting how a human will perform on a given night using numerical data based on recent performances. However, within the NBA—and sports in general—conditions are constantly changing. This includes the number of days a player has had to rest between games. There are occasions during the season when teams play on consecutive days (back-to-backs), or alternatively receive four or more days of rest. These situations introduce additional factors such as muscle soreness, fatigue, and mental fortitude, all of which can influence performance on a given night. This only scratches the surface of the variables that can manipulate results. Throughout the development and testing process, the status and relevant news of all 30 NBA teams were tracked to identify factors that could affect players' roles. Three major influences were found that are difficult to predict using pure statistics alone: games going to overtime, players sitting out, and increased usage for star players within a game.

The first factor is overtime games, which occur when both teams are tied at the end of regulation. Over the past several years, sources estimate that between 4–6 percent of NBA games go to overtime. Given that each of the 30 teams plays 82 games per season, overtime is not a consistent or deciding factor when predicting player performance. However, within the recorded sample size, games marked with an asterisk indicate those that went into overtime, and these games produced some of the largest discrepancies between expected and actual point totals. During overtime periods, top performers typically experience increased usage in an effort to outscore the opponent during the additional five minutes of play. The top players involved in overtime games within this sample include Cooper Flagg, Keyonte George, Alperen Sengun, Nikola Jokic, and Jamal Murray. In these games, each player attempted five or more shots than their usual average, with Flagg in particular scoring well beyond his recent point totals. These performances represent rare outliers that are difficult to predict, highlighting the inherent unpredictability of real-world game scenarios.

The second unpredictable factor is player availability. Players may sit out due to injuries or illness, and when a regular rotation player is unavailable, their minutes and usage are redistributed among the remaining players. NBA rotations typically consist of five to twelve players that coaches rely on consistently. When one or more of these players sit out, others see increased minutes, shot attempts, and offensive responsibility. The final six players in the dataset experienced increased usage due to the absence of their team's star players. For example, Jaylen Wells played without Ja Morant, widely regarded as the face of the Memphis Grizzlies. Without Morant, additional responsibility fell to Wells and other rotation players, resulting in a significant increase in usage and shot attempts. This led Wells to score approximately six points above his expected total, as the model was unaware of Morant's absence. A similar effect was observed with Jaden McDaniels. With Anthony Edwards unavailable for the Minnesota Timberwolves, McDaniels absorbed additional offensive touches, resulting in an outlier performance with an eight-point differential above expectation.

The third factor relates to star players and increased usage. This applies to the most trusted players on each team, such as Cade Cunningham, Derrick White, Brandon Ingram, and Josh Giddey. These players consistently handle the ball, take the most shots, and are relied upon during high-leverage moments. As a result, they occasionally produce performances that exceed expectations. There are also moments within games where players make multiple shots in succession, leading coaches and teammates to continue feeding them the ball to maintain a hot streak. When a player catches fire and establishes dominance, their point total can significantly exceed both their season average and expected output. For example, Cunningham's season average is 27 points, yet he scored 32 due to unusually efficient shooting. Similar patterns were observed with Ingram and Giddey, who both exceeded expectations on nights where they shot with exceptional efficiency.

These cases represent statistical outliers that are difficult to capture using numerical features alone. However, their existence does not indicate that the predictor is inaccurate. The purpose of the model is not to estimate the maximum possible number of points a player could score under ideal conditions, but rather to compute the expected performance under typical circumstances. While real-world conditions constantly fluctuate, the model relies on stable inputs derived from player statistics and opposing team defenses. In cases where the prediction error was within two points—such as with Pritchard, Barnes, Quickley, Powell, Randle, and Coward—the model performed especially well. These players have clearly defined roles within their teams' rotations, resulting in consistent minutes, usage, and shot profiles. Because their responsibilities remain stable from game to game, their scoring output becomes statistically reliable.

Overall, the model performs as intended by providing a consistent baseline expectation for a player's scoring output. It incorporates relevant offensive metrics alongside opposing team defensive performance to produce realistic estimates. While it does not explicitly account for rare or situational factors, such as injuries or overtime, it serves as a dependable foundation that can be supplemented with real-time context when evaluating player performance. Given the inherently dynamic nature of human sports, this approach balances statistical rigor with practical realism.

6 Future Work

There are several directions in which this project could be extended, particularly by addressing the outliers discussed above. Additional features could be introduced to account for increased usage resulting from teammate injuries, as well as methods to model expected performance in overtime based on historical overtime data. Other contextual variables, such as days of rest between games or whether a team is playing at home or on the road, could also be incorporated. These features were not available on the public website used for data collection, so future work would require integrating additional datasets or scraping information from other sources.

Beyond predicting points, this framework could be expanded to estimate other performance metrics such as assists and rebounds. Incorporating opposing team statistics related to steals or rebounding could further enhance predictive accuracy. Since scoring alone does not fully capture a player's impact on the game, extending the model to multiple performance dimensions represents a natural and valuable next step.

Future work can build upon this project by extending the point prediction framework to analyze how players perform relative to their expected scoring output. Using the player-level data collected in this study, predictions from the model can be used as a baseline to categorize performances as scoring below, at, or above expectation. This categorization would allow for a more detailed analysis of the conditions under which players exceed or fall short of predicted outcomes.

By combining these prediction-based performance categories with detailed shooting data, such as shot distribution and shooting efficiency across different shot types, it becomes possible to evaluate how effectively a player operates within a team's offensive system. This approach can help identify whether players are performing efficiently given their roles and whether certain shot selection patterns are associated with outperforming expectations.

In the long term, this type of analysis could be used to study the efficiency of different offensive strategies at both the player and team levels. For example, it could help assess whether emphasizing three-point shooting, attacking the basket, or maintaining a balanced shot profile leads to more consistent scoring outcomes. Such insights may provide valuable guidance for coaching decisions, player development, and system-level optimization in professional basketball.

References

Basketball-Reference.com. Player and team statistical tools used for data collection and feature engineering in basketball analytics reports. :contentReference[oaicite:1]index=1

FanDuel Sportsbook. Betting lines and player proposition totals from major sportsbook analytics.

MadanThevar. "NBA Statistics Analysis Project," a GitHub repository illustrating the application of machine learning and data visualization techniques to NBA player data and predictive analytics tasks. :contentReference[oaicite:2]index=2

NBA.com. Official NBA player statistics and traditional player data for the 2025–26 season, including scoring and defensive metrics. National Basketball Association.

The Data Scientist. "The Science Behind AI-Driven NBA Props Predictions," an article discussing the use of artificial intelligence and machine learning techniques for predicting NBA player performance outcomes. The Data Scientist. :contentReference[oaicite:0]index=0

Zelaya, A. (2025). How oddsmakers set NBA lines: A behind-the-scenes look. *Cardinal Media*. This article discusses how professional oddsmakers use data, experience, and market dynamics to set betting lines for NBA games.