Building an Expected Points Model for Basketball Using Shot-Level Data and Contextual Features

Dimitrios Kotsis

May 2025

Abstract

Sports analytics is a popular research area. With the substantial growth of the sports betting market, modeling the characteristics of a sports game using statistical modeling and machine learning methods has attracted the interest of analysts and stakeholders. Most of this interest has focused on football and basketball. Using shots-level data and team stats data, this project develops a machine learning-based Expected Points (xPts) model. The model estimates the expected value of each shot based on spatial location, temporal game context, and custom defensive metrics. It incorporates psychological featues such as player form and player fatigue, extending beyond traditional field goal percentage analysis. An XGBoost classifier was used to predict the shot-making probability. The evaluation metrics were accuracy and AUC scores. Results were compared to a PPG baseline and a random forest classifier. XGBoost clearly outperformed both. In addition, the SHAP framework was used to determine feature importance, providing a more interpretable measure of how each feature contributed to the model's predictions. Then, the predicted probability outcomes were multiplied by the shot value to calculate expected points (xPts). This newly proposed metric for the game of basketball enables new applications in player evaluation, tactical strategy, and simulation for both analysts and coaches.

Contents

1 Introduction

Sports analytics has become a very popular research area, for multiple reasons. The substantial growth of the sports betting market (Garnica-Caparrós, Memmert, and Wunderlich 2022), interest in sports media coverage (Garnica-Caparrós, Memmert, and Wunderlich 2022; Baker and McHale 2013; Kovalchik 2016; Štrumbelj and Šikonja 2010) and predicting sports provides a better understanding of the characteristics of the sport (Garnica-Caparrós, Memmert, and Wunderlich 2022; Heuer and Rubner 2009; Erik Štrumbelj and Vračar 2012, Wunderlich and Memmert 2018) has attracted the attention of many researchers and stakeholders. With the rapid development of artificial intelligence and machine learning methods, it is more than ever easier to accurately predict player and team performance (Chandru, Kaushik, and Jaiswal 2025, Ouyang et al. 2024)

On the other hand, basketball is an attractive sport in terms of available data because every game generates rich, structured information: shots, assists, rebounds, fouls, steals, turnovers, and free throws, among many other stats (Lampis et al. 2023). This allows experts to have a better judgment on player and team performances, and enhance decision making (Chandru, Kaushik, and Jaiswal 2025)

Football and basketball, and especially the European Premier League and the national Basketball Association (NBA) respectively have dominated the literature thanks to their popularity (W.-J. Chen et al. 2021). Much of this work focuses on predicting game results (win/loss) rather than modeling events inside the game, even though shot-by-shot modeling can offer a more nuanced view of quality (W.-J. Chen et al. 2021). In basketball, the purpose of an offense is to generate the highest possible quality of shot. The higher the quality, the more likely to score and win the game (Skinner 2012). Thus, investigating the different features that affect shot quality will give us insights as to what increases the probability of shot success. Our goal is to build on this line of research by predicting shot success and aggregating those probabilities into an expected-points (xPts) framework.

By leveraging a rich shot-level dataset spanning 1997 to 2020 and modern classification techniques, the model incorporates spatial, temporal, and contextual features. These include not only shot location and type, but also opponent defensive strength and the offensive form of the shooter. Beyond simply predicting outcomes, the model aggregates predictions to the game and player level, revealing trends, efficiencies, and mismatches otherwise obscured by traditional basketball statistics.

The objectives of this research project are threefold: Develop a high-accuracy predictive model, use feature importance to understand what factors of the game have the biggest influence in shot success, and use model outputs to produce and expected points (xPts) framework and comment on its versatility.

2 Literature Review

Basketball shot prediction has traditionally relied on aggregated statistics such as FG% and effective FG%, but recent advances in data availability and analytics have enabled stronger approaches. The prediction of basketball games is a topic many researchers have delved into over the years, and most notably Boulier and Stekler (1999), Caudill (2003), Loeffelholz, Bednar, and Bauer (2009), Rosenfeld et al. (2010)), Stekler, Sendor, and Verlander (2010) and Erik Štrumbelj and Vračar (2012) (Manner 2016). In more detail, Erculj and Štrumbelj (2015) demonstrated how different shot types vary in success rate, while Skinner (2012) framed shot selection as a decision problem influenced by time constraints. Vencúrik et al. (2022) extended this to show that shooting efficiency is context-dependent and often suppressed by defensive pressure.

A wide range of machine learning methods have also been tried for basketball and other sports outcome prediction. Studies report results with classical classifiers such as Logistic Regression, Naïve Bayes, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and tree-based models, as well as neural networks (Ouyang et al. 2024, Houde 2021). For example, Rodrigues and Pinto (2022) used Naïve Bayes, KNN, Random Forest, SVM, and Artificial Neural Networks on five seasons of English Premier League data to predict match outcomes, and related work has applied hybrid approaches—such as SVMs combined with decision trees—to NBA games (Ouyang et al. 2024). Leicht, Gómez, and Woods (2017) built logistic regression and Conditional Inference (CI) trees for men's

Olympic basketball (2004–2016), finding that logistic regression delivered higher accuracy while the CI tree offered better practical interpretability for coaches by handling non-linear effects transparently (Ouyang et al. 2024).

Within basketball specifically, most research targets the binary win/loss outcome for NBA games, reflecting the league's data quality and the natural framing of the problem (W.-J. Chen et al. 2021). Beyond classification, some studies aim to predict scores using modern function-approximation methods such as multivariate adaptive regression splines (MARS), KNN, extreme learning machines (ELM), XGBoost, and stochastic gradient boosting (SGB) (W.-J. Chen et al. 2021)]. In parallel, comparative papers often frame the task as a binary classification problem and benchmark several algorithms side-by-side—typically Logistic Regression, Random Forests, KNN, SVM, Gaussian Naïve Bayes, and XGBoost—to understand trade-offs between accuracy, non-linear capacity, and interpretability (Houde 2021).

Two themes cut across these studies. First, basketball's data richness enables both statistical models and machine learning methods that capture non-linear interactions and complex context (Lampis et al. 2023). Second, there is value in combining models with expert priors or beliefs, for example via the mathematical theory of evidence, so that decisions and downstream strategies reflect both objective signals and informed judgment (Chandru, Kaushik, and Jaiswal 2025). Our work follows this direction but shifts the prediction target from game-level outcomes to shot-level success, which we then aggregate into xPts. This connects the event-level modeling seen in other sports to the NBA setting, while still aligning with the broader tradition of comparing multiple classifiers and leveraging tree-based ensembles like XGBoost when non-linear structure matters (Ouyang et al. 2024), Houde 2021).

Expected goal is a new approach at looking at the performances of a team and a player. Rather that just looking at historical traditional player stats, data scientists and statisticians look at each shot independently. But getting contextual features of each shot, they give an evaluation score, i.e., rating the shot quality. This number can then be aggregated over an entire game, or season to evaluate over or under-performance from a team. The model each-self is not set, with different sports companies offering their own approach and contextual features.

This philosophy is new, and has very rarely been incorporated into other sports. In basketball, similar models are less common and typically limited in scope. Most fail to incorporate dynamic defensive context or player-level offensive trends. Our xPts model addresses this by including a custom defensive rating, season-long and recent offensive performance metrics, and validating both micro (shot-level) and macro (game-level) predictive power.

Brechot and Flepp (2020) introduced expected goals in football as an alternative method to form judgments about a team's performances. They argue that by directly comparing expected goals versus actual results over a game or a season, allow more objective judgment of true performance. Rathke (2017) and Brechot and Flepp (2020) showed how xG captures performance trends obscured by randomness. These models typically combine spatial and event-level features with probabilistic models such as logistic regression.

Low scoring games have a level of difficulty when analyzing team and player performance. Randomness and scarcity of goals make it much more limit a researcher's ability to judge performance. The philosophy was also applied in hockey, and the National Hockey League more specifically, by Macdonald (2012). No attempt has been done to adapt the expected scoring philosophy in basketball. We argue that, regardless the increased difficulty of prediction modeling, aggregating predictions to the game and player level, will reveal strengths, weaknesses and areas of improvement.

3 Methodology

Building an expected points model in basketball requires access to detailed, play-by-play shot-level data, which can be challenging to obtain. Major sports analytics companies collect such datasets but typically use them for their own models, limiting public access. Fortunately, the National Basketball Association (NBA) is highly transparent in its data reporting, and a substantial portion of its statistical record is made available to the public. For this project, data were sourced from a publicly shared

dataset on Kaggle.com (originally compiled from official NBA statistics), which contains the necessary shot-level and contextual information required to develop the model (jonathangmwl 2019; Moore n.d.). This ensured both accessibility and reliability, while allowing for reproducibility of the modeling process.

3.1 Data Sources

The dataset used in this study came from two different sources: a shot-level dataset and a team statistics dataset. The shot-level dataset contains information on all shots taken during regular season and playoff games in the NBA from 1997 to 2021. Specifically, the information extracted from the shot-level dataset and used as features in the model included:

- **Shot location and distance:** Distance from the basket, calculated in Euclidean terms using the coordinates of the player at the time the shot was taken.
- **Shot zone:** The attacking half of the court can be divided into a 3x3 grid of zones. Shot zone indicates the zone in which the player was located when they took the shot.
- **Shot type:** There are over 50 recognized shot types in basketball, each with its own level of difficulty.
- **Time left (in seconds):** Calculated by adding the time remaining in the current quarter to the time remaining in all subsequent quarters.
- **Home/Away indicator:** Whether the offensive player was playing at their home arena or away.

Each of the above features can significantly influence shot quality. Starting with shot location and distance, it is common knowledge that the further away a player is from the basket, the more difficult the shot becomes. Shot distance from the basket is widely agreed to affect both shot efficacy and form (França, Gouveia, Gomes, et al. 2022) This is reflected in the NBA's three-point rule: if a player is positioned at or beyond 7.24 m from the center of the basket and makes a shot, they are awarded three points rather than two. Furthermore, a player's position relative to the basket affects the probability of success for a variety of reasons, such as reduced visibility of the hoop, inability to use the backboard for assistance, and altered shooting angles. The decision to shorten the three-point distance to 6.71 m in the corners is further evidence that facing the basket from a sideways angle increases shot difficulty. How distance and location relevant to the basket affects shot quality and success can be observed in Figure 1.

Erčulj and Erik Štrumbelj (2015) suggest that the shooting technique has also strong predictive power for shot success. This is influenced by several factor, such as player type and angle of shot. The angle at which the ball leaves the hands of the shooter and the angle approaching the basket are crucial for shot success. Ideally, the ball needs to approach the hoop vertically (Miller and Bartlett 1996). Each shot type require different technique, positions, angle of shot, and body position relevant to the basket. Therefore we argue that shot type is imperative when attempting to predict shot success and when measuring shot quality.

Time remaining on the game clock is another critical factor influencing both shot difficulty and probability of success. One of the most valued traits in elite players is the "clutch factor" — the ability to score under high-pressure situations. As the remaining game time decreases, each possession becomes more significant, increasing psychological pressure on players. This often leads to suboptimal shot selection and reduced shot quality. Additionally, the opposing team's defense tends to become more organized and intense in these moments, further lowering the likelihood of a successful shot.

Lastly, the home/away indicator captures the well-established home advantage effect observed across major team sports. Playing at home, in front of supportive fans, often provides players with a competitive edge through increased confidence and familiarity with the environment.

The shot-level dataset went through thorough data cleaning to account for missing values duplicates and formatting issues. Specifically, column not relevant to the model were removed and remaining

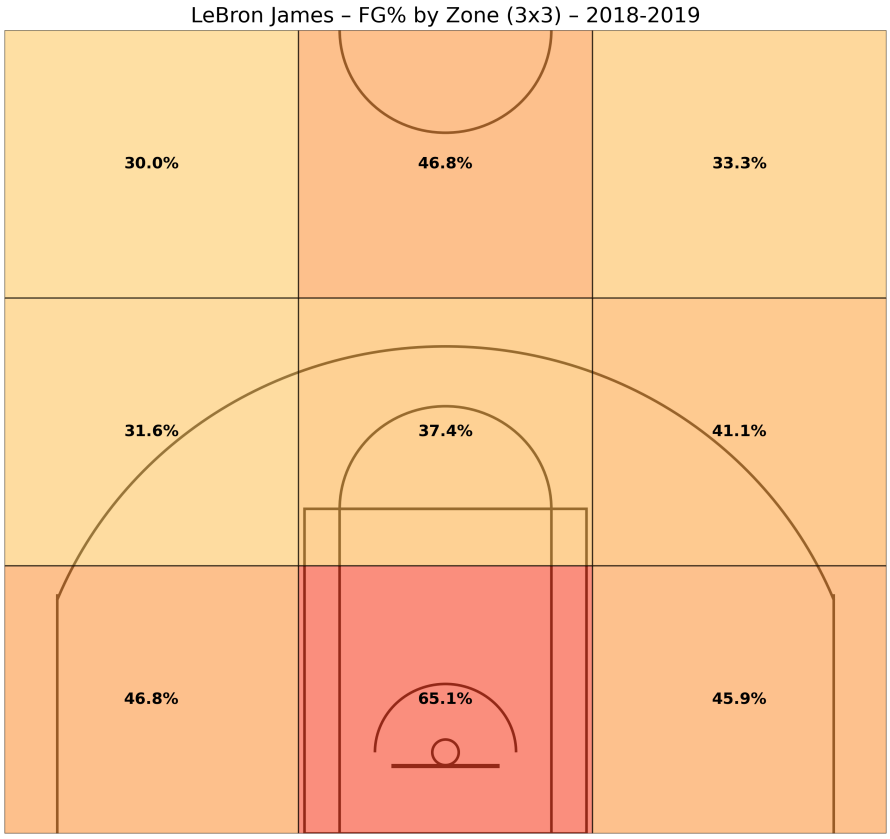

Figure 1: Field goal percentage (FG%) by court zone for LeBron James during the 2018–2019 NBA season. The variation between zones highlights the importance of both distance and location relevant to the basket in modeling shot quality and success.

columns were renamed to make the processing easier. Finally, some teams went through re-branding. These team names and team abbreviations were changed to be consistent. What exactly you cleaned and how (removed rows, imputed values, standardised column names, encoded categories, etc.).

A variety of other contextual and more complex features were added to the model. These will be discussed separately in the following subsections.

3.2 Offensive Rating

Although expected goal models in football typically do not take into account who takes the shot, I argue that in basketball the difference in quality between players can be too significant to ignore. Football and basketball are two sports of very different natures, and what works in one does not necessarily guarantee success in the other. Therefore, this model incorporates historical performance data to assign players an offensive rating. Each player's offensive rating is estimated at a season-long level, based on three key statistics from the previous season: average points per shot, field goal percentage (FG%), and turnover rate. Average points per shot and FG% are widely regarded as two of the most basic yet informative traditional metrics for evaluating a player's offensive performance. Turnovers, defined as how often a player loses possession to the opposing team, are an important indicator of struggles on the offensive end and negatively impact the offensive rating. These three statistics, along with certain defensive metrics, are aggregated into a $+/-$ score, an overall measure of a player's net contribution during a game. The offensive rating used in this study is calculated by extracting the above categories from the shot-level dataset for the previous season.

3.3 Defensive Rating

Shot quality is influenced not only by the offensive abilities of the player taking the shot and other contextual factors, but also by the defensive ability of the opposing team. That is, how effectively the

defending team makes it difficult for the opponent to score. Gaetano et al. (2016) suggest that the caliber of the defender also affects shot quality. Given the nature of the sport, defense is harder to quantify. Basketball is a 5 on 5 sport, with defense dynamics constantly shifting between man-to-man and zone defense. This raises the question on whether defensive rating should be taken on a player or a team level (Stiles 2024). There are many defensive rating formulas available online, including official NBA metrics. However, since this project focuses specifically on predicting shot success, I modified existing defensive rating systems. In particular, I removed steals and turnovers, elements that do not directly lead to a shot, and adjusted the formula accordingly. This ensures that the defensive rating reflects only those qualities that affect shot quality. The custom team-level defensive metric was calculated as follows:

$$FM_{wt} = \frac{\text{DFG\%} \cdot (1 - \text{DOR\%})}{\text{DFG\%} \cdot (1 - \text{DOR\%}) + (1 - \text{DFG\%}) \cdot \text{DOR\%}} \tag{1}$$

$$\text{DefRating} = \text{BLK} \cdot FM_{wt} \cdot (1 - 1.07 \cdot \text{DOR\%}) + \text{DRB} \cdot (1 - FM_{wt}) \tag{2}$$

Here, the following abbreviations are used: **DFG%** – Opponent field goal percentage, **DOR%** – Opponent offensive rebound percentage, **BLK** – Blocks per game, **DRB** – Defensive rebounds per game.

Using the team statistics dataset, I calculated a defensive rating for each team for a given season. This rating was then used as a defensive feature for the subsequent season (Moore n.d.).

3.4 Psychological Factors

As previously mentioned, one of the main objectives of this project was to implement psychological factors that are rarely, if ever, considered in sports modeling. The first factor is current form, how well a player has been performing over the last few games. The concept of momentum is a very popular in sports with most people operating under the assumption that such an effect does exist (Arkes and Martinez 2011). Sports psychologists view momentum as a form of motivation that leads to an increase of psychological and physical performance (W.-J. Chen et al. 2021). It is an elusive and challenging topic to investigate and there is a significant gap in empirical evidence to support it (Arkes and Martinez 2011, Crust and Nesti 2006). Momentum usually goes hand-in-hand with 'hot hand', an effect of similar nature, mostly discussed for basketball. This term refers to a player having an increased probability of shot success had they been successful over the previous few shots (Arkes and Martinez 2011). This concept can be applied to a game-level. Good and poor performances over the past few games leads to higher and lower probability of performing well in subsequent games, respectively. However, it is important to note that hot streaks may be due to randomness (Arkes and Martinez 2011; Crust and Nesti 2006). We argue that that the effects of momentum and 'hot hand' affects the overall performance of the player. The exact number of preceding games is arbitrary, with the optimal number found using trial and error. It is difficult to capture the true representation of a player's offensive abilities by only looking at the previous season performances. Current form captures the fluctuations of performances and gives a more complete picture of where a player's offense stands.

The second psychological factor implemented to the model is player fatigue. Player fatigue is difficult to define, or rather very objectively defined. Athletes are very often put in significant physical demands due to a large number of games, and training sessions. As a results, players may see temporary or permanent decrease in their performances (Shcherbak et al. 2023). Although psychological causes of fatigue have been studied in the past, very little research has been done to find a way to measure player fatigue and incorporate it to predictive models. Given the dataset available, in this model fatigue was measured by how much a player has been traveling and playing consecutive away games. The thinking behind it is that a lot of traveling, leads to fatigue which negatively impacts player performance.

3.5 Modeling

I experimented with several algorithms to identify the most suitable approach for this problem. Logistic regression was a viable option, as it produces a probability for each outcome, which can then be

used to classify the result. Another widely used machine learning algorithm in sports analytics is the random forest classifier. Such simple models are widely applied due to their ease of understanding and strong interpret-ability (Ouyang et al. 2024). However, they fail to achieve high predictive accuracy. The algorithm I ended up using was the XGBoost classifier. XGBoost is a scalable implementation of gradient boosting framework by Friedman J (J. Friedman, Hastie, and Tibshirani 2000; J. H. Friedman 2001; T. Chen, He, et al. 2015), widely used by data scientists (T. Chen and Guestrin 2016). It is also very efficient. This additional advantage of XGBoost over a random forest classifier was very important given the size of the dataset, which spans more than 20 years of games and includes over four million recorded shots.

The modeling process was implemented in Python using the `xgboost`, `scikit-learn`, `pandas`, and `numpy` libraries. Model development followed these steps:

- **Train/test split:** The dataset was split into training (80%) and testing (20%) sets to allow for model evaluation on unseen data. The split was stratified by shot outcome to maintain class balance.
- **Evaluation metrics:** Model performance was assessed using both accuracy and the area under the receiver operating characteristic curve (AUC). Accuracy measures the proportion of correct predictions, while AUC evaluates the model's ability to distinguish between successful and missed shots across all probability thresholds.
- **Feature importance:** The SHapley Additive exPlanations (SHAP) framework was used to determine feature importance, providing a more interpretable measure of how each feature contributed to the model's predictions.

  SHAP is a method to compute feature importance. It leverages game theory concepts to train a classification model using all features, and then computes SHAP values, ranking them in the process (Wang et al. 2024). Evidence has shows that SHAP outperforms other widely used feature selection methods such as ANOVA, Mutual Information and Recursive Feature Elimination, in terms of the Area Under the Receiver Operating Characteristic Curve (AUC) metric (Wang et al. 2024; Marcílio and Eler 2020). It does however require more computation time.
- **Hyperparameter tuning:** Grid search with cross-validation was used to identify the optimal set of hyperparameter for the XGBoost model. The search space included parameters such as learning rate, maximum tree depth, number of estimators, subsampling ratios, and colsample bytree.

Please refer to Figure 2 for a diagram of the model: data sources, features, chosen algorithm and outputs

3.6 Expected Points

Using the model outputs, expected points (xPts) are calculated as:

$$\text{xPts} = P(\text{success}) \times \text{Shot Value}$$

where $P(\text{success})$ is the model's predicted probability of the shot being successful, and *Shot Value* corresponds to the scoring value of the attempt (2 or 3 points, or 1 point for free throws). For example, a 2-point jump shot with $P(\text{success}) = 0.6$ yields 1.2 xPts, regardless of whether the shot is made.
To evaluate performance, xPts can be aggregated at various levels:

- **Player-level:** sum of all xPts from an individual's shot attempts.
- **Team-level:** sum of all xPts from a team's shots in a game or season.
- **Game-level:** comparison of total xPts between teams in a single match.

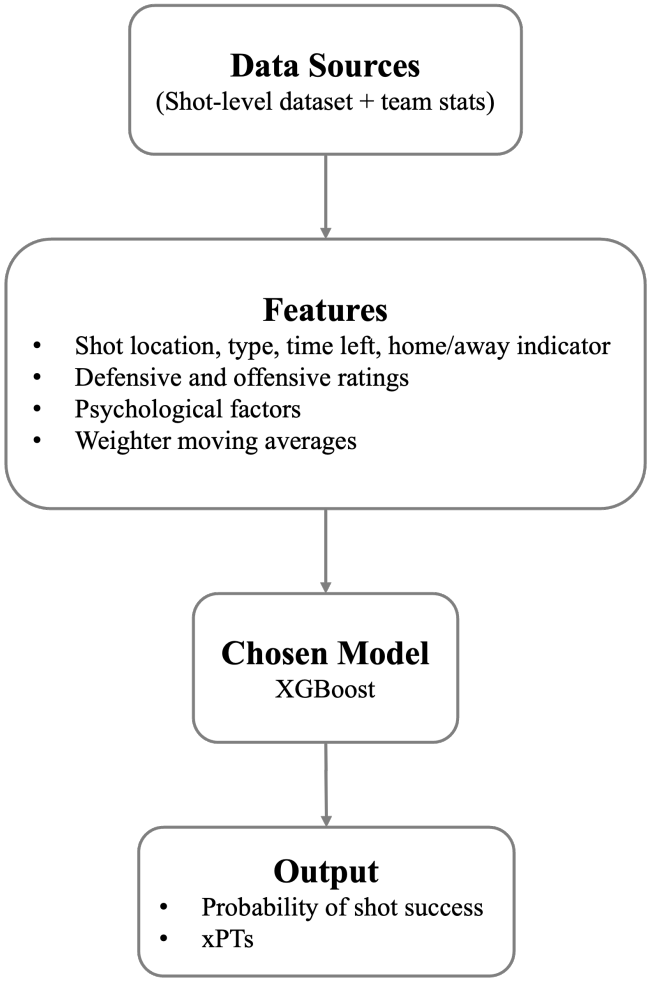

Figure 2: Caption

From these aggregates, a **delta points** metric is derived:

$$\text{Delta Points} = \text{Actual Points} - \text{Expected Points}$$

This captures the degree to which a player or team over- or under-performed relative to the quality of their shot opportunities.

4 Results

4.1 Model Performance

Using the GridSearchCV python library, grid search with cross-validation was used to obtain the optimal set of hyperparameters for the XGBoost model. The optimal number of estimators were 200, max depth 5, the learning rate 0.1, subsample 0.8, and colsample bytree 1.

Figure 3 shows a model performance comparison between the models: a PPG baseline, a Random Forest Classifier, and the XGBoost classifier. The three models were assessed using accuracy and AUC as evaluation metrics. XGBoost achieved the highest scores in both metrics, with an accuracy of 0.6425 and an AUC score of 0.6767, outperforming both the PPG baseline (accuracy 0.55, AUC 0.51) and the Random Forest Classifier (accuracy 0.61, AUC 0.63).

4.2 Feature Importance

Feature importance was evaluated using SHAP values, which ranks the impact of each feature on the model's output. Figure 6 presents a SHAP beeswarm plot showing the 20 most influential features

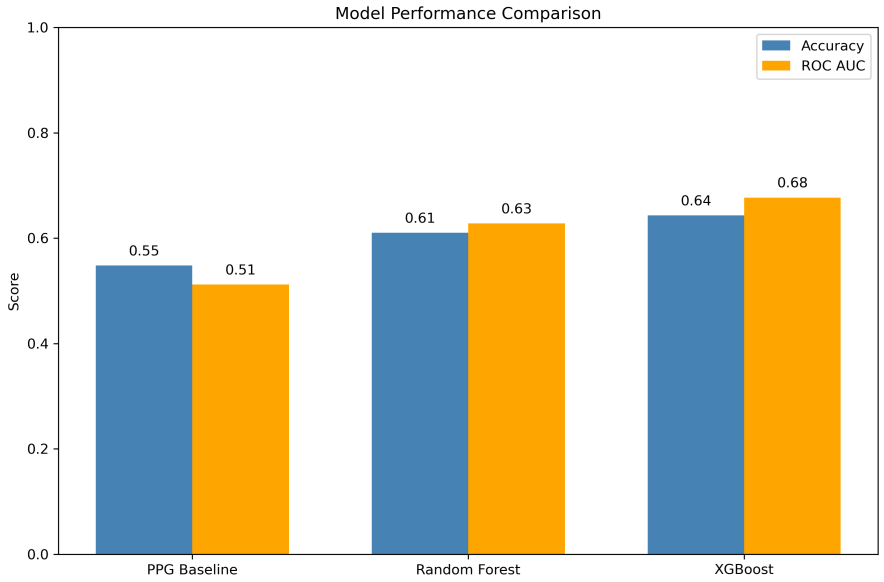

Figure 3: Baseline Comparison: Model performance comparison between a PPG baseline, a Random Forest Classifier, and an XGBoost Classifier. Accuracy and AUC scores were used as evaluation metrics.

in the XGBoost model. The most important variables included shot type (particularly jump shots), followed by shot distance, field goal percentage (FG%), and other types of shots such as layups and slam dunks. Contextual and psychological factors such as time remaining, defensive rating, and recent form also contributed meaningfully to the model's predictions, though to a smaller degree.

4.3 Expected Points Outputs

Using the predicted probabilities of the model, we calculated expected points by multiplying the probability of a make by the shot value. As part of the analysis, we examined the many applications expected points have to offer:

- **Player Evaluation:** Identify players who over-perform or under-perform by directly comparing actual points scored versus expected points. In addition, identify players who produce high-quality shots, regardless of their FG
- **Team Evaluation:** Similar to player evaluations but on a team level, by aggregating expected points over an entire game and/or season for the entire team. Aimed at analyzing areas of strengths and weaknesses of a team.
- **Rankings:** Use results to rank players by efficiency: expected points per 100 possessions

5 Discussion

Before discussing the results in more detail, it is imperative to highlight the importance of a multivariate model. Average points per shot and FG% are widely regarded as two of the most basic yet informative traditional metrics for evaluating a player's offensive performance. Figure 4 shows a chart of all the makes and misses of Lebron James during the entirety of the the 2018-19 season. Yet, makes and misses, FG% and PPG lack spatial information. To illustrate the added value of contextual information even further, refer back to Figure 4 that shows a shot chart for LeBron James over a season. While his FG% and PPG offer a high-level view of his scoring efficiency, the spatial pattern of shot success rates reveals a much richer story, demonstrating why raw aggregate metrics fail to capture the nuances of shot quality and context.

All traditional basketball statistics are important, but when presented on their own, they are not as informative. This is further supported by looking at individual Kernel Density Estimates. Figure 5 shows significant overlap between makes and misses and almost no conclusions can be drawn by simply looking at the effects of individual features. To understand what drives shot success, one needs to look at the combined effect of features in a multivariate model, rather than in isolation.

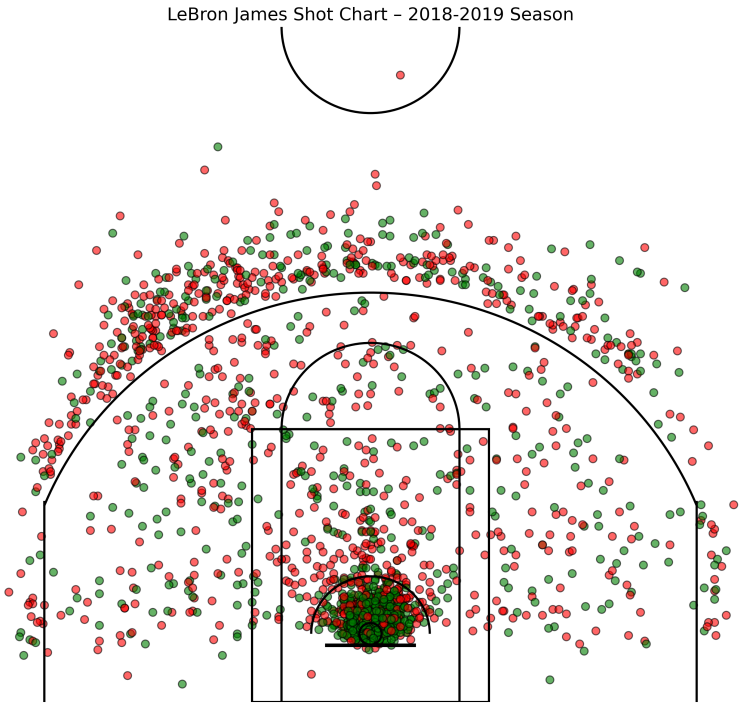

Figure 4: Shot chart for LeBron James during the 2018–2019 NBA season, showing made (green) and missed (red) field goal attempts. The spatial distribution of attempts demonstrates the value of incorporating location-based features.

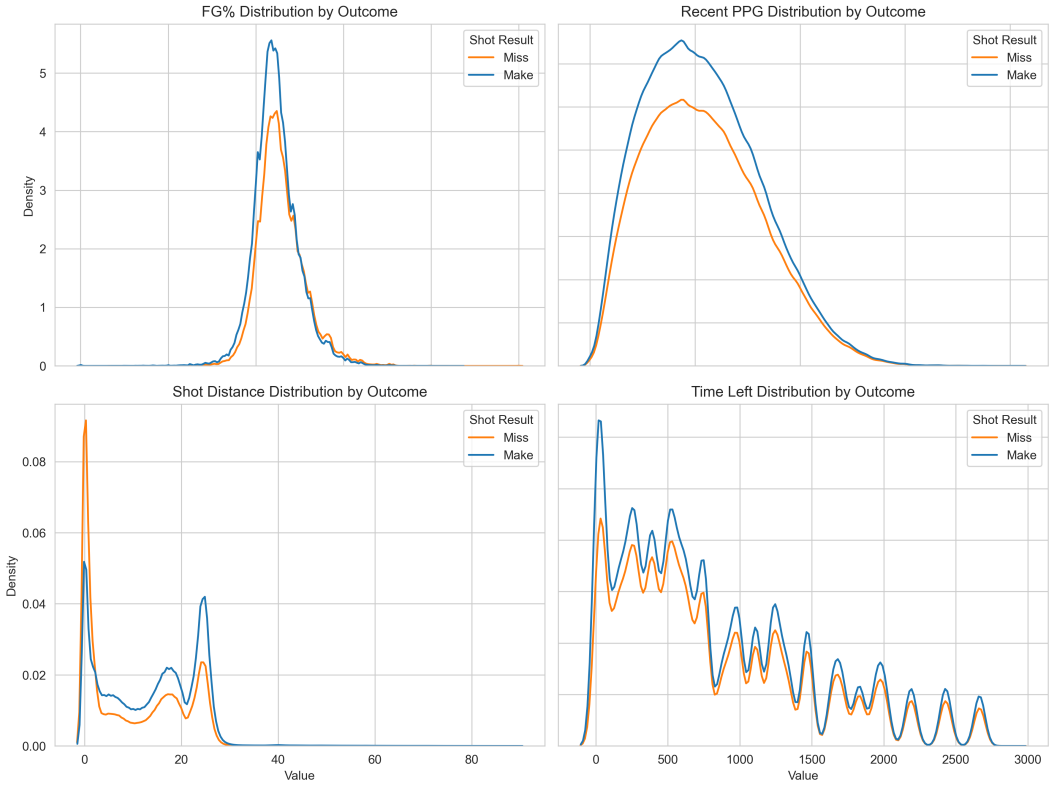

Figure 5: KDE Composite: Kernel Density Estimates from top left to bottom right: FG%. Recent PPG, Shot distance and Time left

5.1 Interpretation of Results

The results demonstrate that the XGBoost classifier outperformed both the points-per-game (PPG) baseline and the random forest classifier across both evaluation metrics. Specifically, XGBoost achieved an accuracy of 0.64 and an AUC of 0.68, compared to the random forest's accuracy of 0.61 and AUC of 0.63, and the baseline model's accuracy of 0.55 and AUC of 0.51. This performance improvement suggests that XGBoost's ability to model complex non-linear interactions between variables, combined with its robustness to a mix of numerical and categorical features, makes it particularly well-suited to predicting basketball shot outcomes. The model's predictive power, as reflected in AUC and accuracy might seem low at first glance. However, in the context of basketball, a fast based game with a high count of shots, and outcomes very noisy, the results are good. In fact, this performance in on par with, or slightly better that previous research done on the field. In addition, xPts remains a valuable analytical tool.

The SHAP feature importance analysis for the XGBoost model supports the inclusion of contextual, spatial, psychological and player ability features. As seen in Figure 6, shot distance was by far the most influential factor, followed by some types of shots and time left on the clock. The beeswarm plot clearly shows that shots taken from a high distance, lower the probability of shot success, whilst shots taken from a short distance increase the probability of shot success. This aligns nicely with the SHAP values for certain types of shots. For example, slam dunks, shot can only be taken from a short distance as they require at least one hand touching the basketball hoop, increase the probability of shot success. However, success rates of types of shot are not only aligned with shot distance, but also the level of difficulty of the shot each-self. This is most commonly influenced by the opponents defensive abilities, forcing a player to do some maneuvering to create some distance from the defender, or take unnatural and uncomfortable shots. A prime example of this is a hook shot, a type of shot where the ball handler stands perpendicular to the basket and shoots the ball in a sweeping motion over their head. To no surprise, according to Figure 6 found that it lowers the probability of shot success. The above observations align with existing basketball analytics literature that emphasizes the importance of shot location and action type in determining shot success probabilities, and further

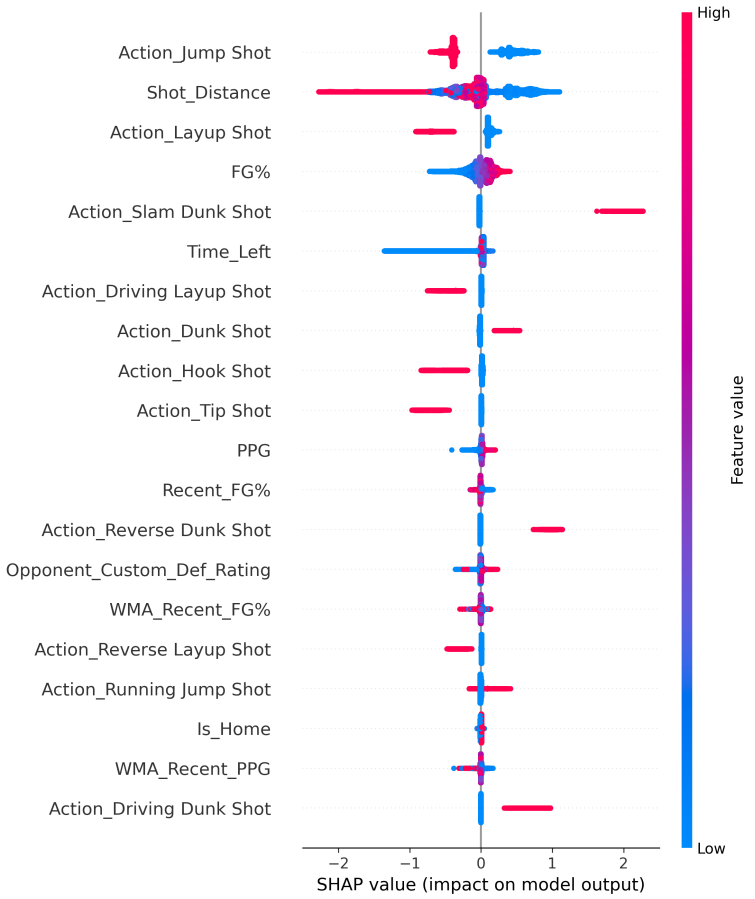

Figure 6: SHAP feature importance beeswarm plot for the XGBoost model, showing the top 20 features ranked by mean absolute SHAP value. The x-axis indicates how much, positively or negatively, each features impacts the model output. Additionally, red means high feature value and blue low feature value. For categorical shot types, feature value corresponds to the binary indicator resulting from one-hot encoding. A high value indicates that the shot was of that type, while a low value indicates it was not.

support the claim we made on the importance of combined effect of features in a multivariate model.

Field goal percentage (FG%) and points per game (PPG) also emerged as strong predictors. Player's with historically higher FG% tend to have a higher probability of shot success, and the other way around. This reflects nicely on the suggestion we made earlier on incorporating player-specific historical performance metrics. Recent form also provided some predictive power, though to a smaller degree. This indicates that this feature complements offensive factors such as FG% and PPG. Furthermore, contextual factors such as time remaining on the clock contributed meaningfully. Low time left value negatively impacted the model output. In other words, as the game approaches towards its end, players take worse shots and tend to miss more. Moreover, the addition of a custom defensive rating provided exploratory power. This indicates that modeling the defensive strengths at a team level complements offensive and other contextual factors.

Overall, these findings validate the methodological choice to move beyond traditional basketball statistics and integrate offensive and defensive, contextual, and psychological factors into a single framework.

5.2 Expected Points Applications

Traditional basketball metrics such as FG% and PPG provide only a partial view of scoring performance. They do not account for the difficulty of shots, defensive pressure, or situational context. The xPts framework addresses these gaps by incorporating multiple contextual and player-specific features

into a single probability-based metric. This allows for fairer comparisons between players and teams, even when their shot profiles differ significantly. The expected points philosophy was applied on multiple levels to examine the usefulness of it, and the many applications it has to offer. First we test on **player evaluations**, by identifying players who consistently generate high-quality shots, regardless of their FG% and PPG. To achieve that, for an entire season, we calculate total points scored, total expected points and a *delta points* metric, which is a direct comparison between actual and expected points. Furthermore, to evaluate each player's efficiency, we also calculate expected, actual and delta points per 100 possessions, i.e., per 100 shots. Table 1 is a list of the top 10 players ranked by xPts per 100 possessions for the season 2015-16.

Table 1: Top 10 players ranked by efficiency: expected points per 100 possessions (season 2015-16)

Player Name	**Total xPts**	**Total Actual**	**xPts/100**	**Actual/100**	**Delta/100**
DeAndre Jordan	683.60	773	140.95	159.38	18.43
Tyson Chandler	411.60	426	131.50	136.10	4.60
Festus Ezeli	265.46	276	127.01	132.06	5.04
Clint Capela	232.05	248	126.80	135.52	8.72
Rudy Gobert	471.99	470	126.20	125.67	-0.53
Tarik Black	244.26	248	124.62	126.53	1.91
Alonzo Gee	212.61	201	123.61	116.86	-6.75
Dwight Howard	487.61	523	123.45	132.41	8.96
Omer Asik	334.13	318	122.84	116.91	-5.93
Andre Iguodala	546.19	572	121.92	127.68	5.76

Notes: xPts/100 = Expected points per 100 shots. Actual/100 = Actual points scored per 100 shots. Delta/100 = Difference between actual and expected points per 100 shots; positive values indicate out-performance relative to model expectations, while negative values indicate under-performance.

There are multiple takeaways from Table 1. DeAndre Jordan topped the table on all metrics. On one hand, he can be considered as the most efficient player during the 2015-16 season as he achieved the highest points total (773) and the highest actual points per 100 possessions (159.38). This indicates that he was a very efficient shot maker, took full advantage of every opportunity to score the basket. In addition, he also produced the highest xPts (683.60) and xPts per 100 possessions (140.95). Therefore, not only did he score the basket efficiently, but he also produced the best shots per 100 possession in terms of quality compared to the rest of the league. However, he also had +18.43 delta points per 100 possessions, the highest in the table. This tells us that despite his incredible season performance and shot selection, he actually over-performed. By directly comparing expected and actual points, we find that DeAnde Jordan scored more that what was predicted, based on the quality of the shots he took. Alonzo Gee on the other hand had a negative delta. That means despite producing high quality shots, he scored less than expected, and therefore under-performed.

A table of such nature can produce quality results for coaches and analysts, and move in many direction based on what metric someone decides to rank players on. This approach, ranking player and/or teams by directly comparing official results with expected outcomes is supported by Brechot and Flepp (2020). Naturally, there are several limitation as well. We need to remember that at the end of the day, basketball is about what team is going to score more than their opponent. The table lists many players that have very few points for a season, and hence it is difficult to make the argument that they are the best players in the league. This is most likely attributed to playing few minutes or focusing primarily on defense. One way to improve results is with positional rankings. Positional rankings based on xPts reveal which roles on the court are generating the highest shot quality. For example, centers may consistently produce high xPts due to shorter shot distances and higher-percentage shot types, whereas guards may rely more on contested jump shots with lower success probabilities. In fact, most players in Table 1 are centers. Unfortunately, in modern basketball, players change positions very often and our dataset does not provide what position each player is.

By aggregating expected points over an entire game, separately for each team, we can analyze a game and investigate which team performed the best and 'deserved' to win. The random game that was analyzed was a game between the Dallas Mavericks (DAL) and the Philadelphia 76ers (PHI).

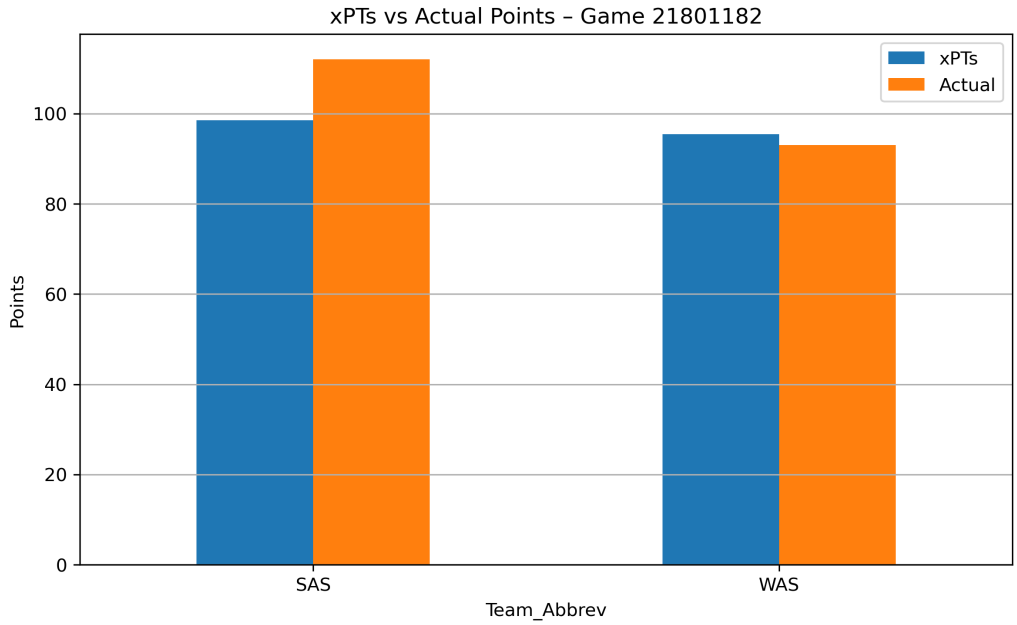

Figure 7: Comparison of expected points (xPts) and actual points for the San Antonio Spurs (SAS) and Washington Wizards (WAS) in game 21801182. While SAS outscored WAS, xPts suggest that WAS generated higher-quality scoring opportunities.

Table 2: Comparison of actual points and expected points (xPts) for Game ID 21801182.

Team	**Actual Points**	**xPts**	**Delta Points**
Dallas Mavericks (DAL)	101	102.856693	-1.856693
Philadelphia 76ers (PHI)	103	85.949316	17.050684

As seen in Figure 7 and Table 2, DAL scored a total 101 points, with their xPts coming at 102.86. PHI scored 103 points, with an xPts of only 85.95. This shows that DAL, scored almsot exactly as much as their xPts suggested. In contrast, PHI scored way more than they were expected. If the points scored from each team were a representation of the quality of their shots, then DAL would have dominated the game. At the team level, aggregating xPts across all games in a season can offer alternative perspectives on performance. Teams that consistently generate higher xPts than their opponents, even in losses, may be playing more efficiently than their win–loss record suggests. This concept can be extended to simulate entire seasons based on xPts-derived results, producing an "expected" league table that reflects underlying performance quality rather than game outcomes alone. This gives coaches a clearer sense of where their team stands, helps them stay confident in their plan and approach, and lets them attribute an unsuccessful run to variance when appropriate. They can also dive into games where they generated low xPts, or allowed opponents to generate higher xPts than their season average, to understand what went wrong offensively and/or defensively.

The Golden State Warriors, during the 2015–16 season, achieved a 73–9 regular-season record. Analysts consider them one of the best teams to ever compete in the NBA, and they were clear favourites to win the NBA Finals. Looking at Table 3, we can compare actual versus xPts results over multiple games for GSW. In a sample of 10 random games, all were actual wins, but three were predicted as losses using expected points. Over the entire season, the xPts-based simulation for GSW gave them 58 wins and 26 losses. Therefore, the expected-points model suggests that this team was not as dominant as the record implied. They ultimately lost the Finals to the Cleveland Cavaliers. So were they as good as people thought, or did they overperform and benefit from a lucky regular season?

Overall, beyond predictive accuracy, the expected points (xPts) framework developed in this project provides a flexible tool for performance evaluation and strategic decision-making. At the player level, aggregating xPts over a season enables the calculation of a *delta points* metric, defined

Team	Opp	Pts	Opp_Pts	Result	xPts	Opponent_xPts	xPts_Result
GSW	NOP	93	81	Win	94.887121	76.775187	Win
GSW	ATL	106	91	Win	117.946038	87.749300	Win
GSW	CHI	105	92	Win	131.085892	102.356335	Win
GSW	NOP	93	77	Win	103.795272	80.836958	Win
GSW	MIL	96	68	Win	80.627555	81.494527	Loss
GSW	NOP	91	76	Win	109.053344	77.521098	Win
GSW	UTA	94	86	Win	89.182722	85.745625	Win
GSW	DET	98	86	Win	114.765554	90.660296	Win
GSW	LAC	84	78	Win	84.690069	86.813323	Loss
GSW	DET	95	88	Win	101.967233	105.038468	Loss

Table 3: Game-level comparison of xPts and actual results for GSW.

as the difference between actual points scored and expected points. Positive deltas indicate over-performance relative to shot quality, while negative deltas suggest under-performance. This can be used to identify players who are either exceeding expectations or failing to capitalize on high-quality shot opportunities.

5.3 Interpretation of Mean and Spread in xPts

The average expected points per shot (xPts) across the dataset is approximately 1.0. At first glance, this might appear to suggest that the model is assigning roughly a 50% make probability to all shots; however, this value is entirely consistent with historical NBA scoring efficiency. In the NBA, the league-average points per shot (PPS) in recent decades has typically fallen between 1.0 and 1.1. This stems from the combination of two main shot types: (1) two-point attempts converted at roughly 50% and (2) three-point attempts converted at approximately 33%. When weighted by their real-world frequency (roughly two-thirds of shots being twos), the overall PPS converges naturally to ~ 1.0.

The more important question is whether the model can differentiate between high- and low-quality shots, rather than the absolute league-wide mean. Examination of the xPts distribution shows a clear spread: the histogram exhibits two main peaks corresponding to different shot profiles (e.g., contested mid-range versus high-percentage paint attempts), with a long tail representing near-certain finishes such as dunks and layups. Across all players, the standard deviation of average xPts per shot is approximately 0.035, indicating meaningful variation despite the league-wide mean being anchored around 1.0. Player-level means range from ~ 0.35 for historically inefficient shooters to values exceeding 1.3 for those taking a high proportion of uncontested close-range attempts.

This reinforces the underlying philosophy of xPts: it is not intended to radically shift the global mean efficiency, but rather to quantify relative shot quality and to enable comparisons across players, teams, and contexts. The model's value lies in highlighting which players or teams consistently create higher-quality opportunities (and whether they convert them), and conversely, which underperform relative to the quality of their looks.

Initial experiments with season-level simulations were explored; however, such results proved highly sensitive to modeling assumptions and data coverage, occasionally producing unrealistic league tables due to systematic overestimation for certain team profiles. Consequently, the analysis in this work focuses on shot-level and aggregated player/team performance, where the metric is most interpretable and directly tied to its design goal: evaluating the quality of the scoring opportunities themselves.

The xPts model confirms that incorporating context improves prediction accuracy. Defensive strength and offensive form both contributed meaningfully, validating the hypothesis that simple FG% does not capture shot quality.

Limitations include lack of player tracking data (e.g., defender proximity), which would enhance precision. Random Forest, while interpretable, may underperform deep models in raw predictive power, though interpretability was a key design goal.

Future work could include real-time prediction via spatiotemporal models or simulations based on xPts to inform in-game strategy. The model could also be extended to women's leagues or European

leagues for comparative analysis.

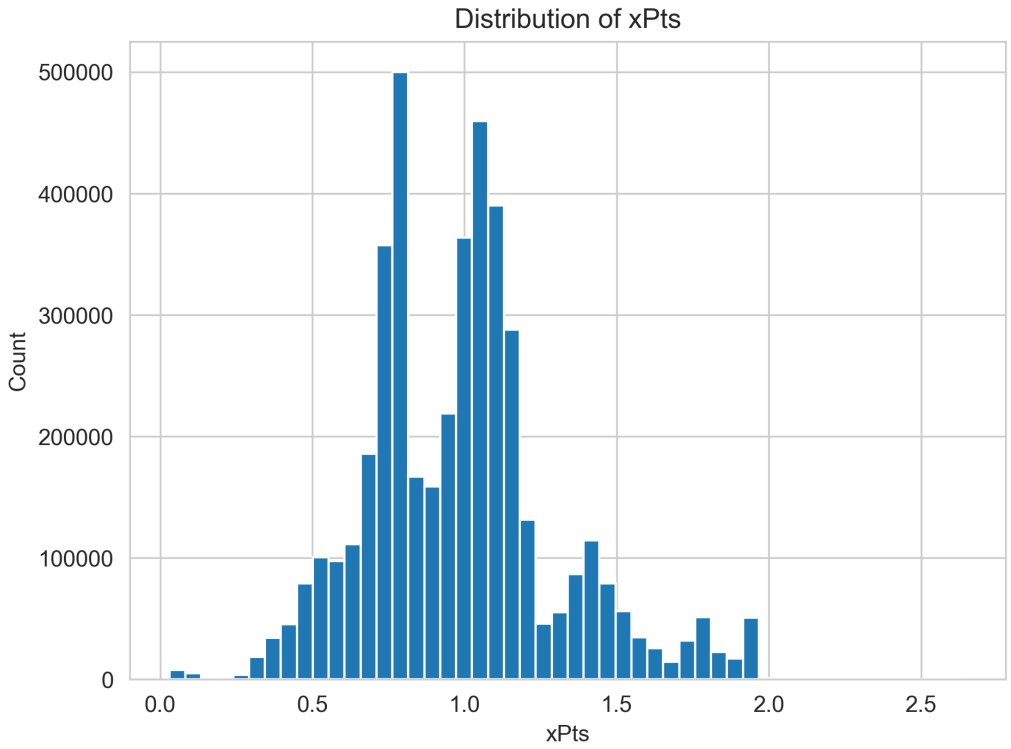

Figure 8: Expected Points (xPts) Histogram.

5.4 Limitations

While the results are promising, several limitations should be acknowledged. First, the dataset does not include player tracking data such as defender proximity, shot contest level, and angle of shot, all of which are known to influence shot difficulty. Second, despite the addition of 'recent form', the offensive ability metric is based primarily on season-level historical averages, which may not capture both short-term form fluctuations, and long term patterns. For example, a player that had a bad season will be attributed a low offensive rating which does not reflect on their true abilities. Third, the model focuses on the probability of shot success of a specific number and does not account for broader possession outcomes, such as assists, turnovers within the same play and second or third chances. Lastly, player fatigue seemed to have very limited predictive power, probably due to the approach I took to define it. Additionally, given the large historical dataset, there is a possibility that certain patterns reflect past playing styles or rule sets that may have shifted over time.

5.5 Future Work

Future research and work could address these limitations in several ways. First, integrating player tracking data would allow for more precise modeling of defensive pressure and shot contesting. Moreover, it would allow to gather information on player movement, total distance covered during a game and total minutes played per game. This additional information will provide a more detailed view on player fatigue. Then, incorporating temporal models, such as recurrent neural networks (RNNs) or long short-term memory (LSTM) networks, could capture the sequential nature of possessions and game flow. Lastly, an addition to the model that could improve the predictive accuracy would be updating offensive ratings dynamically throughout a season. This also applies to player's recent form.

Overall, the findings of this project have practical implications for coaches, analysts, and front offices. Despite several limitations, the xPts framework offers a quantifiable metric for identifying players who are under-performing relative to shot quality, which can inform player development and

shot selection strategies. Positional and team-level xPts analyses can highlight areas for tactical improvement, such as creating higher-quality shots for certain roles or adjusting defensive schemes to limit opponents' high-quality opportunities. In recruitment and scouting, xPts can serve as an additional metric to evaluate prospective players, complementing traditional statistics and subjective assessments. Finally, simulated league tables based on xPts can offer alternative narratives about team performance, potentially influencing strategic decisions in roster management and game planning.

References

Miller, Stuart and Roger Bartlett (1996). "The relationship between basketball shooting kinematics, distance and playing position". In: *Journal of sports sciences* 14.3, pp. 243–253.

Boulier, Bryan L and Herman O Stekler (1999). "Are sports seedings good predictors?: an evaluation". In: *International Journal of Forecasting* 15.1, pp. 83–91.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2000). "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)". In: *The annals of statistics* 28.2, pp. 337–407.

Friedman, Jerome H (2001). "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics*, pp. 1189–1232.

Caudill, Steven B (2003). "Predicting discrete outcomes with the maximum score estimator: The case of the NCAA men's basketball tournament". In: *International Journal of Forecasting* 19.2, pp. 313–317.

Crust, Lee and Mark Nesti (2006). "A review of psychological momentum in sports: Why qualitative research is needed". In: *Athletic Insight* 8.1, pp. 1–15.

Heuer, Andreas and Oliver Rubner (2009). "Fitness, chance, and myths: an objective view on soccer results". In: *The European Physical Journal B* 67.3, pp. 445–458.

Loeffelholz, Bernard, Earl Bednar, and Kenneth W Bauer (2009). "Predicting NBA games using neural networks". In: *Journal of Quantitative Analysis in Sports* 5.1.

Rosenfeld, Jason W et al. (2010). "Predicting overtime with the Pythagorean formula". In: *Journal of Quantitative Analysis in Sports* 6.2.

Stekler, Herman O, David Sendor, and Richard Verlander (2010). "Issues in sports forecasting". In: *International Journal of Forecasting* 26.3, pp. 606–621.

Štrumbelj, E and M Robnik Šikonja (2010). "Online bookmakers' odds as forecasts: The case of European soccer leagues". In: *International Journal of Forecasting* 26.3, pp. 482–488.

Arkes, Jeremy and Jose Martinez (2011). "Finally, evidence for a momentum effect in the NBA". In: *Journal of Quantitative Analysis in Sports* 7.3.

Macdonald, Brian (2012). "An expected goals model for evaluating NHL teams and players". In: *Proceedings of the 2012 mit sloan sports analytics conference*.

Skinner, Brian (2012). "The problem of shot selection in basketball". In: *PloS one* 7.1, e30776.

Štrumbelj, Erik and Petar Vračar (2012). "Simulating a basketball match with a homogeneous Markov model and forecasting the outcome". In: *International Journal of Forecasting* 28.2, pp. 532–542.

Baker, Rose D and Ian G McHale (2013). "Forecasting exact scores in National Football League games". In: *International Journal of Forecasting* 29.1, pp. 122–130.

Chen, Tianqi, Tong He, et al. (2015). "Xgboost: extreme gradient boosting". In: *R package version 0.4-2* 1.4, pp. 1–4.

Erčulj, Frane and Erik Štrumbelj (2015). "Basketball shot types and shot success in different levels of competitive basketball". In: *PloS one* 10.6, e0128885.

Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Gaetano, Raiola et al. (2016). "Analysis of learning a basketball shot". In: *Journal of Physical Education and Sport* 16, pp. 3–7.

Kovalchik, Stephanie Ann (2016). "Searching for the GOAT of tennis win prediction". In: *Journal of Quantitative Analysis in Sports* 12.3, pp. 127–138.

Manner, Hans (2016). "Modeling and forecasting the outcomes of NBA basketball games". In: *Journal of Quantitative Analysis in Sports* 12.1, pp. 31–41.

Leicht, A. S., M. A. Gómez, and C. T. Woods (2017). "Explaining match outcome during the men's basketball tournament at the Olympic Games". In: *Journal of Sports Science & Medicine* 16.4, p. 468.

Rathke, Alex (2017). "An examination of expected goals and shot efficiency in soccer". In: *Journal of Human Sport and Exercise* 12.2, pp. 514–529.

Wunderlich, Fabian and Daniel Memmert (2018). "The betting odds rating system: Using soccer forecasts to forecast soccer". In: *PloS one* 13.6, e0198668.

jonathangmwl (2019). *NBA Shot Locations*. Dataset. Kaggle. URL: `https://www.kaggle.com/datasets/jonathangmwl/nba-shot-locations` (visited on 08/01/2025).

Brechot, Marc and Raphael Flepp (2020). "Dealing with randomness in match outcomes: how to rethink performance evaluation in European club football using expected goals". In: *Journal of Sports Economics* 21.4, pp. 335–362.

Marcílio, Wilson E and Danilo M Eler (2020). "From explanations to feature selection: assessing SHAP values as feature selection mechanism". In: *2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI)*. Ieee, pp. 340–347.

Chen, Wei-Jen et al. (2021). "Hybrid basketball game outcome prediction model by integrating data mining methods for the national basketball association". In: *Entropy* 23.4, p. 477.

Houde, Matthew (2021). "Predicting the outcome of NBA games". In.

França, Cíntia, Élvio Rúbio Gouveia, Beatriz B Gomes, et al. (2022). "A kinematic analysis of the basketball shot performance: impact of distance variation to the basket". In: *Acta of bioengineering and biomechanics* 24.1.

Garnica-Caparrós, Marc, Daniel Memmert, and Fabian Wunderlich (2022). "Artificial data in sports forecasting: a simulation framework for analysing predictive models in sports". In: *Information Systems and e-Business Management* 20.3, pp. 551–580.

Rodrigues, F. and Â. Pinto (2022). "Prediction of football match results with Machine Learning". In: *Procedia Computer Science* 204, pp. 463–470.

Lampis, Tzai et al. (2023). "Predictions of European basketball match results with machine learning algorithms". In: *Journal of Sports Analytics* 9.2, pp. 171–190.

Shcherbak, Tetiana et al. (2023). "Psychological causes of fatigue in football players". In: *Journal of Physical Education and Sport* 23.8, pp. 2193–2202.

Ouyang, Yan et al. (2024). "Integration of machine learning XGBoost and SHAP models for NBA game outcome prediction and quantitative analysis methodology". In: *Plos one* 19.7, e0307478.

Stiles, Dylan J (2024). "Defensive Impact Wins: Developing a New Method to Rate Individual Defense in NBA Games". In.

Wang, Huanjing et al. (2024). "Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods". In: *Journal of Big Data* 11.1, p. 44.

Chandru, Roshan, Abhishek Kaushik, and Pranay Jaiswal (2025). "Enhancing Basketball Team Strategies Through Predictive Analytics of Player Performance". In: *Electronics* 14.11, p. 2177.

Moore, Eoin A. (n.d.). *NBA Dataset – Box Scores & Stats, 1947–Today*. Dataset. Kaggle. URL: `https://www.kaggle.com/datasets/eoinamoore/historical-nba-data-and-player-box-scores` (visited on 08/01/2025).