NBA SHOT VALUE PREDICTIONS
ISYE 6740 - GROUP 1

Tyler Burns, Josh Caplan, Avery Girsky, Spiros Valouxis

ABSTRACT

This project explores the application of machine learning techniques to predict the value of shots taken by NBA players. Leveraging a dataset comprising various features related to player actions and shot circumstances, we trained a predictive model to estimate the expected point value of each attempted shot. Subsequently, we compared these predicted values with the actual points scored by players on their shots. By analyzing the variance between predicted and actual outcomes, we derived a novel metric quantifying a player's proficiency in converting shot opportunities into points. This study not only contributes to the understanding of predictive modeling in basketball analytics but also introduces a practical metric for evaluating shooting efficiency in the NBA. The results provide insights into individual player performance and may have implications for strategic decision-making in team management and player development.

1 INTRODUCTION

In the dynamic realm of professional basketball, understanding the nuanced elements of player performance is paramount for both team success and individual player development. One pivotal aspect of a player's contribution to their team is their ability to convert shot opportunities into points, making accurate shot prediction a crucial undertaking in basketball analytics. This project delves into the multifaceted challenge of predicting NBA shot outcomes and subsequently assessing the value added by players through their shooting proficiency.

The primary objectives of this study encompass two key dimensions: the classification of whether a shot results in a successful basket or not, and the quantification of the inherent value of each attempted shot. To achieve these goals, we employed machine learning techniques, specifically exploring linear regression models and probabilistic approaches based on shot locations to predict the point value associated with each shot attempt. For classifying if the shot goes in or not, we explored decision trees and logistic regression models.

The motivation for this project stems from the desire to enhance our understanding of player performance beyond traditional statistics. By delving into the intricate patterns and circumstances surrounding shots taken, we aim to provide a more nuanced perspective on a player's shooting capabilities. This analytical approach allows us to uncover insights into what makes certain players stand out in terms of shot efficiency and, consequently, contribute the most value as a shooter to their teams.

To undertake this endeavor, we harnessed a comprehensive dataset encompassing every shot recorded in the NBA from the 2003 season to the most recent one DomSamangy. This extensive shot log forms the foundation for our analysis, offering a rich source of information on shot outcomes, player actions, and contextual variables. Additionally, we augmented this dataset by incorporating relevant supplementary data, enabling the creation of new variables that contribute to a more comprehensive understanding of the factors influencing shot success Lauga.

As the NBA continually evolves with advancements in player skills, strategies, and game dynamics, this project seeks to provide a contemporary perspective on the predictive modeling of shot outcomes and the identification of players who significantly elevate their team's performance through their shooting prowess. Through this exploration, we aim to contribute valuable insights to the basketball analytics community and pave the way for more informed decision-making in player evaluation and team strategy.

2 Data Collection and Preprocessing

2.1 Dataset

The original dataset was originally sourced from nba.com. The dataset consisted of every shot taken in the NBA from the 2003-04 season through the 2022-2023 season, with 4,012,561 data points and 26 features. Each observation represents a single shot attempt with data such as player, team, shot location, game situation and more.

2.2 Preprocessing and Additional Features

Within the dataset, we found that the only NaN entries were the POSITION and POSITION_GROUP of approximately 20 players, which we filled by manually finding their positions with Google searches.

Beyond the initial set of features, we sought to create new features that we imagined could be helpful in the modeling stage. The ACTION_GROUP variable cleaned up ACTION_TYPE by reducing the 70 different types of shots in ACTION_TYPE into 8 categories. In order to analyze the timing of shots taken, we created the TIME_LEFT_QUARTER variable using the MINS_LEFT and SECS_LEFT, as well as TIME_LEFT_GAME variable using the QUARTER. From our analysis of time left, we found that the lowest accuracies occurred at the end of each quarter, within the final seconds.

Thus, we created a BUZZER variable to indicate if the shot is a buzzer beater (less than 5 seconds in a quarter remaining) and a CLUTCH variable to indicate if the shot was taken in the clutch (less than 2 minutes in a game remaining). To merge some game-related information for each shot we found another dataset to merge into our current one. We used advanced stats data via the NBA api to merge the defensive ratings of the opponents with our shooting data to get the variable DEFENSIVE_RATING. Additionally, from a games dataset that had the scores of each game over the same period as our shot data, we created a binary variable called 'HOME' to indicate if a shooter is playing at home, as well 'SCORE_DIFFERENCE' that showed the difference in score by the end of the game.

2.3 Exploratory Data Analysis

Through studying the ACTION_GROUP feature, we found that dunk shots were the most accurate type of shot at 0.905 accuracy whereas jump shots were the least accurate with 0.374 accuracy.

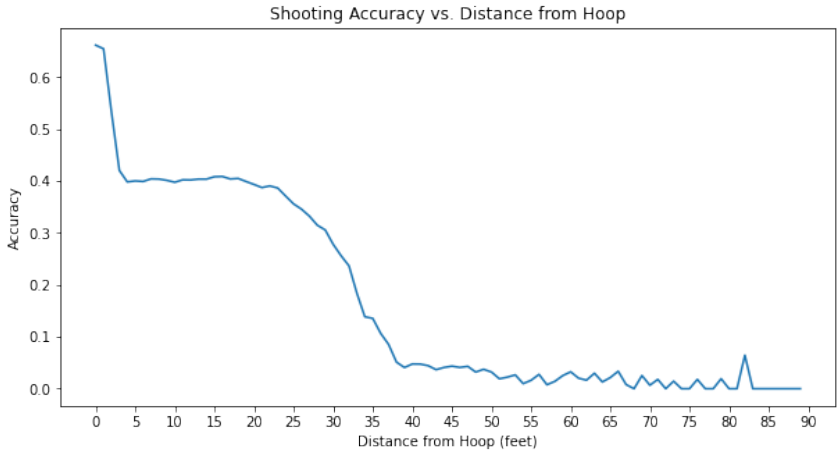


Figure 1: Accuracy vs. Distance

Additionally, we analyzed shooting distance compared to shooting accuracy. As seen in figure 1, there is a steep decline from 0.66 at no feet away from the hoop to about 0.4 at 5 feet. The accuracy

then plateaus and stays at roughly the same until around 22 feet. This is striking because the three-point line is 23 feet and 9 inches away, which means that a three-point shot at the three-point line has a close, albeit slightly lower, accuracy than a two point shot between 5 and 22 feet.

However, this slight decrease in accuracy fails to compensate for the increase in points gained from shooting a three-pointer. This is most likely why the NBA has seen a steady increase in three-point shots taken over the past two decades, as seen in figure 2. Whereas, three-point shots made up less 20% of attempted shots in the 2003-2004 season, the proportion of these shots reached almost 40% in the 2022-2023 season.

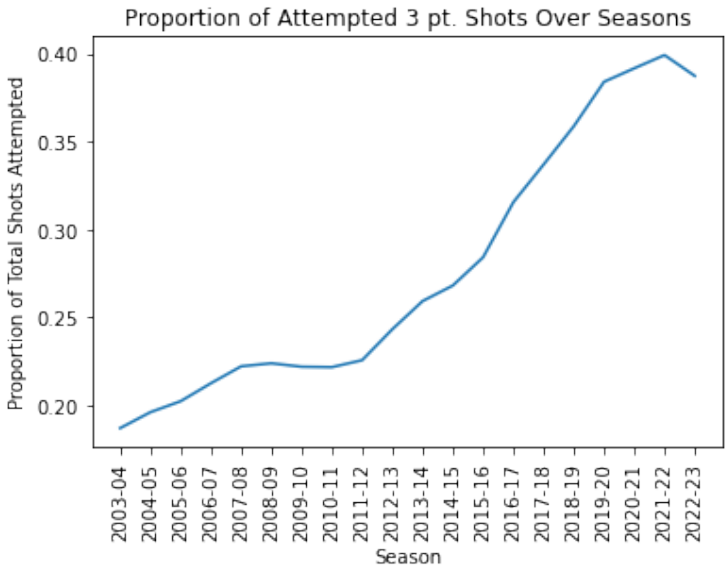


Figure 2: 3-Point Attempts

3 Model Description and Training

To reduce training times and eliminate any issues stemming from the way the game has changed over the last 20 years, we decided to only go back to the 2017 season for our data. For all models, we trained and validated the models on 2017 to 2022 data, and then we tested the models on 2023 data. Furthermore, unless otherwise stated, all models used the following variables:

Feature	**Description**	**Type**
POSITION_GROUP	Guard, forward, or center	object
SHOT_TYPE	2-point shot or 3-point shot	object
BASIC_ZONE	Description of the area on the court	object
LOC_X	Horizontal component of court location	float
LOC_Y	Vertical component of court location	float
SHOT_DISTANCE	Distance from the hoop	integer
ACTION_GROUP	Description of how shot was attempted (i.e. dunk shot)	object
TIME_LEFT_QUARTER	Seconds remaining in the quarter	integer
TIME_LEFT_GAME	Seconds remaining in the game	integer
DEFENSIVE_RATING	Opposing team season-long defensive rating	float
BUZZER	Is shot taken in the last 5 seconds of a quarter	boolean
CLUTCH	Is shot taken in the last 2 minutes of the game	boolean
HOME	Is shot taker on the home team	boolean

Table 1: List of predictive features, descriptions, and types used in modeling task.

We explored three classes of predictive models: classification on the shot outcome, probability of the shot outcome, and regression on the shot value.

Outcome	Description	Type
SHOT_OUTCOME	False if shot missed, True if shot made	boolean
SHOT_VALUE	0 if shot missed, 2 if made two-pointer, and 3 if made three-pointer	integer

Table 2: Outcome features, descriptions, and types used for prediction.

3.1 Classification on Shot Outcome

As more of an exploratory measure, we built a decision tree to create an accuracy baseline of classifying the boolean outcome of whether a shot was made.

3.1.1 Decision Tree Classifier

The focus of this decision tree was not to achieve the highest accuracy on the dataset, but rather give a better understanding of the relative importance of features in shot outcome prediction. To this end, we set the min_impurity_decrease hyperparameter of scikit learn's DecisionTreeClassifier to be 0.001. Hence, a split would only occur if it led to a substantial decrease in Gini impurity.

3.2 Probability of Shot Outcome

We trained a wide variety of models on our primary task, predicting the probability that a shot was made. Our focus was on this section because the results could be transformed into both of our other problems. Setting a threshold on the probability provides classification results. Alternatively, multiplying by 2 or 3 depending on if the shot was a 2-pointed or 3-pointer yields a shot value prediction.

3.2.1 (Regularized) Logistic Regression

We trained a basic logistic regression model as a baseline using only original variables to the shots dataset. Then, we trained a model using grid search to test different regularization methods, regularization parameters, and number of iterations using the same data. After, we again used grid search on the same parameters but this time with the new variables included. We found the best parameters were a ridge regression model with a C of 0.01 and max iterations of 300.

Also, for the classification aspect of this model, we explored changing the threshold for where to classify as 0 and where to classify as 1. We got a threshold of 0.42, as that is where precision and recall met. This gave us a slightly lower accuracy, but we could see from it that lowering the threshold gave us more makes from further away from the basket, which we liked. Ultimately, since we were primarily focused on the probability aspect of this model rather than the classification aspect, we shifted our focus more towards analyzing the probabilities of the shots.

3.2.2 Decision Tree Regressor

Again, we implemented a decision tree with the goal of providing a highly interpretable model, not one with the best performance. To predict shot probability, scikit learn's DecisionTreeRegressor was trained with SHOT_OUTCOME as the response. The min_impurity_decrease hyperparameter was set at 0.003.

3.2.3 Tree Ensemble Models

In our exploration of predicting the probability of a successful shot, we employed 3 tree ensemble models (Random Forest, XGBoost, AdaBoost). Unlike aiming for the highest accuracy (whether the shot goes in or not), our primary objective was to accurately predict the probabilities that a shot goes in. To achieve this, we fine-tuned the hyperparameters of the models testing different combinations. We observed that the default hyperparameters were either the best or very close to the best, so we used those.

First, the **Random Forest**, a versatile ensemble model known for its robust performance. LOC_X, LOC_Y and Defensive Rating were the 3 most important features for predicting the probabilities, but the results were not satisfactory (MSE: 0.2565).

Next, we trained an **XGBoost** model on the same dataset resulting in a much improved MSE of 0.2291. The most important features were the shot distance and the action group of the shot (especially if it was a dunk).

Finally, the **AdaBoost** model was the best out of the 3 with an MSE of 0.2279 and similar results with the XGBoost model.

Table 3 shows the results of the 3 tree ensemble models we tried:

Model	**MAE**	**Brier Score (MSE)**	**RMSE**
Random Forest	0.4596	0.2565	0.5065
XGBoost	**0.4543**	0.2291	0.4787
AdaBoost	0.4571	**0.2279**	**0.4774**

Table 3: Model Evaluation Metrics

3.2.4 Artificial Neural Network

Trained using TensorFlow, our feed-forward neural network architecture comprises three layers with 128, 64, and 32 nodes, utilizing ReLU activation for capturing non-linear patterns. The Adam optimizer with a learning rate of 0.001 facilitates efficient weight optimization. To prevent overfitting, a dropout rate of 0.2 is applied. The mean squared error (MSE) serves as the loss function, aligning with the regression task's objective.

The model is trained for 50 epochs with a batch size of 128. Early stopping is not implemented in this configuration. These choices aim to balance complexity and generalization, allowing the neural network to learn and represent intricate patterns within the data.

3.3 Regression on Shot Value

We chose the best performing probability prediction model based on MSE and compared it to a baseline linear regression for the shot value prediction problem.

3.3.1 (Regularized) Linear Regression

Similar to the logistic regression approach from earlier, we trained a basic linear regression model as a baseline using only original variables to the shots dataset. Then, we trained a model using lasso regularization on the same data. We used grid search to determine the best regularization parameter. Then, we did the same thing on data that included the new variables, so we could determine if our new variables had any effect on the model. We found that the best C was 0.01, and the model selected the new variable, DEF_RATING.

3.3.2 Artificial Neural Network

The artificial neural network trained on the probability task was scaled corresponding to the value of the shot, resulting in a prediction of the shot value. We hypothesized that a network trained on the shot value outcome itself would provide similar performance to our scaled results, and we were correct in that assessment.

4 Experimental Results

To discuss the results of our models of shot prediction, it is helpful to first visualize every shot attempted and it's result in figure 3.

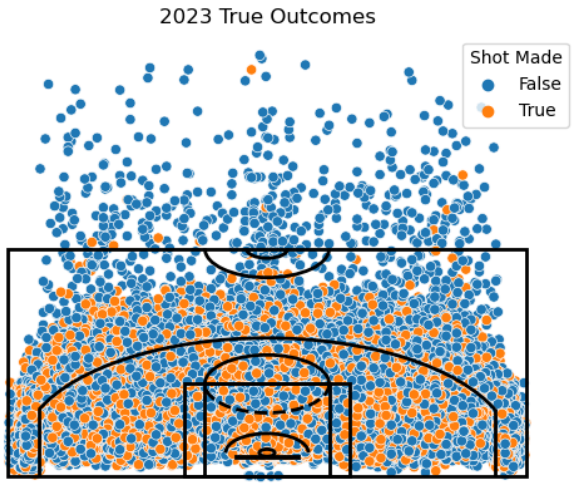


Figure 3: True Shot Outcomes in 2023

4.1 Classification

The decision tree classification revealed an accuracy of 0.6235. Despite the appearance of more splits based on information gain, only one split was observed to have an impact on the classification results: $\text{SHOT\_DISTANCE} \leq 2.5$. This simple yet impactful split demonstrated a good baseline accuracy. Further splits involved shot distance and a boolean variable signaling if the shot was a dunk. While they changed the probability of the shot being made, they did not impact the classification result.

4.2 Probability

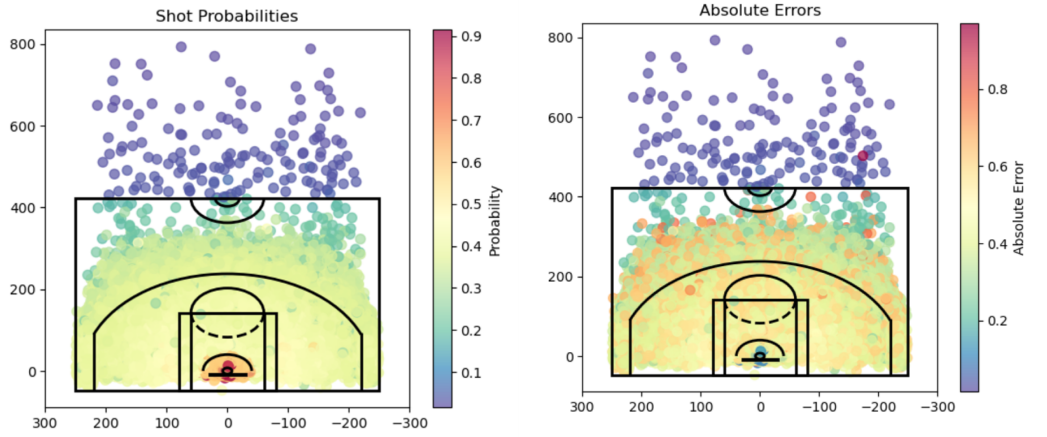
We chose the selection metric for our models to be MSE. In Figure 4, we can see that the neural network achieved the lowest MSE of our probability-based models. Therefore, we decided to use that model to predict our shot values by multiplying the probabilities by 2 or 3 depending on where the shot was taken. The accuracies listed in figure 4 are based on a 50% threshold.

Model	Accuracy	Brier Score (MSE)	RMSE	MAE
Decision Tree Classifier	0.6235	--	--	--
Logistic Regression	**0.6297**	0.2276	0.4770	0.4555
Regularized Logistic Regression	0.6291	0.2273	0.4770	0.4555
Decision Tree Regressor	0.6235	0.2286	0.4781	0.4563
Random Forest	0.6223	0.2565	0.5065	0.4596
XGBoost	0.6268	0.2291	0.4787	**0.4543**
AdaBoost	0.6265	0.2279	0.4774	0.4571
Neural Network	0.6262	**0.2269**	**0.4764**	0.4579

Figure 4: Results of Probability/Classification Models

We produced plots of the probabilities of our highest accuracy model, logistic regression, overlaid on a basketball court as seen in Figure 5a. These results make sense because the general pattern is

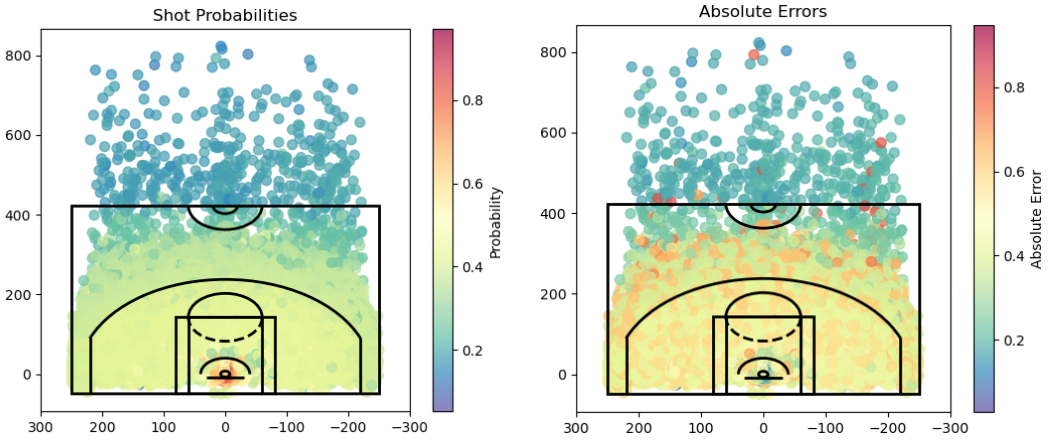
the further from the basket, the lower the probability of the shot being made. We also produced an absolute errors plot in Figure 5b for each shot.



(a) Shot Probabilities (b) Absolute Errors

Figure 5: Logistic Regression Shot Results

As the best performing model according to MSE, the feed-forward neural network provides similar results in prediction. Figure 6a displays the probability of each shot being made. Again, probability generally decreases with distance from the basket, but there is also a noticeable drop in probability about five feet from the hoop. Additionally, figure 6b shows the absolute error of each neural network probability to the true result. We see the best performance at the extremes, next to the hoop or far away.



(a) Shot Probabilities (b) Absolute Errors

Figure 6: Neural Network Shot Results

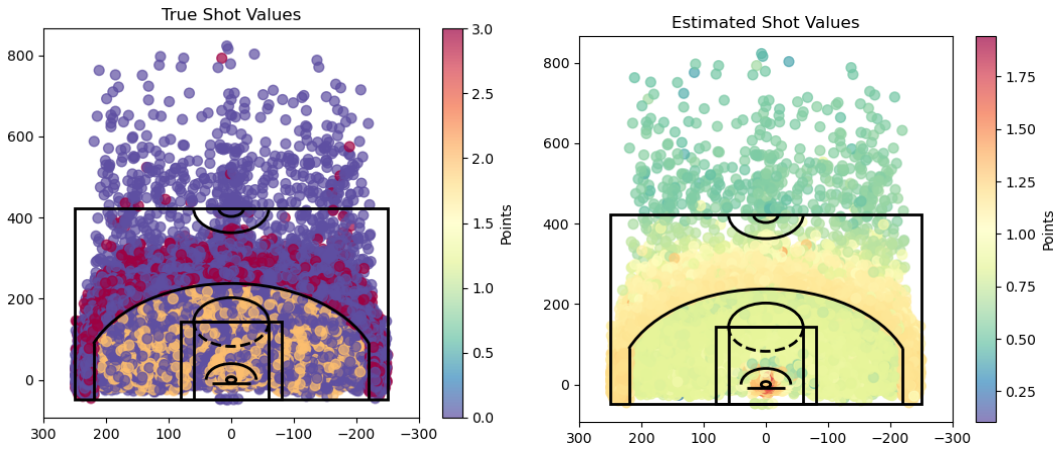
4.3 Regression

Again, the selection metric for our shot value regression models was MSE. In Figure 7, we can see that the neural network achieved a lower MSE than both simple and regularized linear regression, but not by a significant margin. Still, when comparing player performance to expected results, we chose the Neural Network model to provide the expectation. An interesting note, the mean absolute error these models has a very practical meaning. The MAE represents the average error measured in actual points between the true shot value and estimated value.

Model	MSE	RMSE	MAE
Linear Regression	1.3542	1.1637	**1.0895**
Lasso Linear Regression	1.3594	1.1660	1.0999
Neural Network	**1.3532**	**1.1633**	1.0962

Figure 7: Results of Regression on Shot Value Models

In figure 8, we compare the true shot values against the feed-forward neural network predicted values. The true values are either 0, 2, or 3 as seen in the three colors on display. More importantly, the shot value prediction results in more expected points on shots immediately outside the three-point arc. A practice that is commonly adopted in modern basketball and reinforced by this figure is that the most efficient shots are either directly next to the basket or outside the three-point arc.



(a) True Shot Values

(b) Neural Network Predicted Shot Values

Figure 8: True and Predicted Shot Value Results

5 Insights and Contributions

In our exploration of NBA shot value prediction for the 2023 season, we utilized our neural network to forecast the expected value (EV) of every shot taken by NBA players during the 2022-23 season.

5.1 Scoring Difference: A Novel Metric

To assess player performance comprehensively, we introduced a novel basketball statistic: Scoring Difference. This metric represents the difference between the total points made by a player (minimum 250 points scored) and the cumulative expected values of their shots [Figures: 9, 10]. In essence, Scoring Difference gauges a player's ability to surpass or fall short of the model's predictive expectations, providing a unique perspective on offensive proficiency.

5.2 Comparative Analysis

Our analysis focused on comparing the predicted values to the actual outcomes of every shot, grouping the results by player exclusively for the 2023 season. This approach allowed us to identify the top and bottom performers.

Player_Name	Total_Points_Made	Total_EV	Score_Difference	Average_Points_Per_Shot	Average_EV_Per_Shot	Per_Shot_Difference
Nikola Jokic	1349	1034.75	314.25	1.32	1.01	0.31
Stephen Curry	1367	1141.53	225.47	1.23	1.03	0.20
Kevin Durant	1059	854.62	204.38	1.23	0.99	0.24
Domantas Sabonis	1185	1025.68	159.32	1.26	1.09	0.17
Joel Embiid	1503	1343.78	159.22	1.15	1.03	0.12
De'Aaron Fox	1483	1328.57	154.43	1.11	1.00	0.12
Luka Doncic	1623	1493.06	129.94	1.12	1.03	0.09
Kyrie Irving	1376	1246.91	129.09	1.14	1.04	0.11
Jalen Brunson	1308	1186.18	121.82	1.09	0.99	0.10
DeMar DeRozan	1360	1238.93	121.07	1.04	0.95	0.09
Jakob Poeltl	778	660.89	117.11	1.26	1.07	0.19
Kawhi Leonard	996	880.56	115.44	1.14	1.01	0.13
Deandre Ayton	1051	939.67	111.33	1.18	1.06	0.13
Bojan Bogdanovic	1005	896.72	108.28	1.14	1.02	0.12
Kevin Huerter	1045	941.24	103.76	1.21	1.09	0.12
Buddy Hield	1238	1136.13	101.87	1.19	1.09	0.10
Devin Booker	1165	1063.84	101.16	1.09	1.00	0.09
Nikola Vucevic	1315	1214.33	100.67	1.15	1.06	0.09
Donovan Mitchell	1603	1502.68	100.32	1.14	1.07	0.07
Luke Kennard	509	414.94	94.06	1.33	1.09	0.25

Figure 9: Top 20 Players by Scoring Difference

Player_Name	Total_Points_Made	Total_EV	Score_Difference	Average_Points_Per_Shot	Average_EV_Per_Shot	Per_Shot_Difference
Luguentz Dort	813	979.65	-166.65	0.93	1.12	-0.19
RJ Barrett	1141	1273.54	-132.54	0.97	1.08	-0.11
Killian Hayes	694	811.13	-117.13	0.85	1.00	-0.14
Paolo Banchero	1043	1145.41	-102.41	0.93	1.02	-0.09
Scottie Barnes	989	1090.74	-101.74	0.97	1.07	-0.10
Terry Rozier	1151	1252.70	-101.70	0.97	1.05	-0.09
Dillon Brooks	927	1027.60	-100.60	0.94	1.04	-0.10
Jalen Green	1319	1418.87	-99.87	0.97	1.04	-0.07
Tari Eason	651	744.51	-93.51	0.99	1.13	-0.14
Russell Westbrook	938	1029.90	-91.90	0.96	1.05	-0.09
Jabari Smith Jr.	848	939.81	-91.81	0.95	1.05	-0.10
Dorian Finney-Smith	506	594.95	-88.95	1.01	1.19	-0.18
Jaden Ivey	938	1026.58	-88.58	0.95	1.04	-0.09
Kelly Oubre Jr.	814	902.28	-88.28	0.99	1.10	-0.11
Josh Okogie	390	477.76	-87.76	0.94	1.15	-0.21
Jeremy Sochan	533	619.30	-86.30	0.97	1.12	-0.16
Dennis Smith Jr.	396	477.09	-81.09	0.88	1.06	-0.18
Bennedict Mathurin	926	1002.10	-76.10	0.97	1.05	-0.08
Fred VanVleet	1081	1153.38	-72.38	0.97	1.04	-0.07
Oshae Brissett	294	361.69	-67.69	0.92	1.13	-0.21

Figure 10: Bottom 20 Players by Scoring Difference

5.3 Pure Assessment of Scoring Ability

Crucially, this analysis evaluates a player's ability to score points above the model's expectation, shedding light on their capacity to overperform or underperform relative to predictive metrics. The focus on point differentials provides a clear indication of a player's impact on the scoreboard, making Scoring Difference a valuable metric in understanding offensive prowess.

In summary, our innovative approach to predicting shot values and introducing the Scoring Difference metric offers a unique perspective on player performance. This analysis aims to contribute valuable insights to the ongoing discourse on evaluating NBA players' offensive capabilities, enriching the understanding of their contributions beyond conventional statistics.

5.4 Surpassing Scope of Homeworks

Our project surpassed the typical bounds of a homework assignment in several aspects. Initially, we navigated the challenge of sourcing and merging different datasets, manipulating and introducing novel variables for our models. The shot level data scale was notably extensive, comprising over 4,000,000 data points—significantly surpassing the scope of typical homework assignments. Prior to model creation, we engaged in thorough exploratory data analysis, a step often overlooked in assignments. Unlike assignments where model choices are stipulated, we autonomously determined which models best suited our problem, exploring a more diverse set than typically assigned. With the large scale of data, we also needed to alter model parameters to keep the run-time low for cross validation. Additionally, for the probability models, we experimented with altering the threshold to construct a classification model—an endeavor not typically encountered in homework settings. Another departure from the norm was the in-depth analysis of our results, delving beyond mere accuracy assessment. We confirmed our modeling results further through practical analysis of player performance. Observing that star-level scorers performed well over model expectation asserts that the model is working for an average NBA player and also demonstrates its usefulness as a real-life statistical metric.

6 Conclusion

6.1 Contextual Considerations

While Scoring Difference emerges as a useful statistic, it is essential to interpret it within the context of other performance metrics. A well-rounded assessment should consider various factors, including defensive contributions, playmaking abilities, and overall team dynamics. Scoring Difference provides a specific lens on offensive output, but a comprehensive understanding of a player's overall impact requires integration with broader statistical and contextual analyses.

Let's consider the case of Giannis Antetokounmpo, widely regarded as one of the best players in the NBA. Despite having a very slight negative Scoring Difference, the analysis reveals intriguing insights [Figure: 11]:

- Giannis consistently takes shots with a high average expected value, aligning closely with the model's predictions.
- His scoring output is in line with the expected metric, indicating efficiency in converting shots into points.
- Beyond shot outcomes, Giannis excels in drawing fouls, evident from his impressive 12 free throw attempts (FTA) per game. This ability contributes significantly to his elite scorer status, with an impressive 31.1 points per game (PPG), even in situations where he may perform less optimally according to our Scoring Difference metric.

Player_Name	Total_Points_Made	Total_EV	Score_Difference	Average_Points_Per_Shot	Average_EV_Per_Shot	Per_Shot_Difference
Giannis Antetokounmpo	1461	1466.87	-5.87	1.14	1.15	-0.0

Figure 11: Giannis Antetokounmpo Stats

This example underscores the importance of considering a player's diverse skill set, such as drawing fouls and overall scoring proficiency, when assessing their contribution to the team. While Scoring Difference provides valuable insights, the broader context, including a player's style of play and unique skills, is essential for a comprehensive understanding of their impact on the court.

6.2 Future Work

One way we could improve our model is by using more variables that the NBA protects such as closest defender and shot clock data. We believe these variables could have a large impact on whether a shot goes in or not. With a more accurate model, the conclusions from our analysis would be more sound. We could further involve some analysis with free throw data to determine the players best at drawing fouls and contributing scoring value that way.

Our work provides other avenues for studying player performance. For example, shot selection (the quality of shots taken by a player) could be examined through the EV per shot of players. A player with a higher EV per shot would generally have better shot selection.

Overall, we believe this style of analysis has the potential to become a much better metric to judge a player's scoring ability than current shooting efficiency metrics such as FG%, eFG%, TS%. While these percentages do incorporate the value of shots, they fail to include the expectation component that our analysis provides.

References

DomSamangy. Nba shots dataset. https://github.com/DomSamangy/NBA_Shots_04_23.

Nathan Lauga. Games dataset. https://www.kaggle.com/datasets/nathanlauga/nba-games/.