Finding Determinants of NBA Shot Probability using Interpretable Machine Learning Methods

Anonymous[†,*]

[†] Anonymous

1. Introduction

Since the advent of basketball analytics, a metric that is accurately able to determine the relative worth of player's defense has been widely sought after. It is widely regarded that features like shot defense are key to a player's defensive identity, but regularized on-off metrics like RAPM (Sill, 2010) are unable to take this into account. Using player-tracking data, we are able to extract information about shot defense.

We determine the relative importance of a of a set of offensive and defensive factors on individual shots near the 3 point line. Using 2015-16 SportVU data, where player and ball positional coordinates are captured 25 times a second, and the accompanying play-by-play data, we extract the following features: 'Distance Between Shooter And Defender', 'Shot Distance', 'Difference Between Shooter And Defender Height', and '3PT%'. 3PT% is calculated for the entirety of the 2015-16 season.

We then train a gradient boosting model to predict the shot success probability of a given shot. Although this can be useful on its own, it does not directly provide the relative importance of each of the input features.

To this end, we use interpretable machine learning techniques, specifically shapley values. Using TreeSHAP (Lundberg et al., 2020), we determine the importance scores for each input feature, per shot. Aggregating these values over all games in our dataset, we estimate the relative importance of each feature.

[*]N/A

2. Data

We use two data sources for our analyses: NBA play-by-play Data, obtained from stats.nba.com ("NBA Stats", 2020), and archived SportVU NBA Player Tracking Data.

2.1. **Player Tracking Data.** We obtain archived SportVU data from the 2015-16 NBA Season (Seward, 2018). This data captures the X-Y coordinates of every player on the court and the ball 25 times a second across the course of an NBA game.

First, we use the play-by-play data to determine the approximate time when 3PT shots occurred. We then determine the exact time of the 3PT shot, using the player tracking data.

We then apply a Savitsky-Golay filter (Schafer, 2011) on the height of the ball. This increases the precision of the SportVU-measured ball height, making the proceeding calculations of velocity and acceleration more accurate.

We are then able to determine the velocity, and the acceleration of the ball height denoted by v(t) and a(t) respectively: concretely defined as:

$$(2.1) \qquad\qquad v(t) = \frac{dx}{dt}$$

$$(2.2) \qquad\qquad a(t) = \frac{dv}{dt}$$

Using the acceleration and velocity, we are able to determine the exact time of the shot. Specifically, before the ball reaches its maximum point, we determine its highest acceleration.

We then wish to extract the features: 'Distance Between Shooter And Defender', 'Shot Distance', 'Difference Between Shooter And Defender Height', and '3PT%'.

Specifically, we define distance to be the standard Euclidean distance:

$$(2.3) \qquad\qquad d(p_1, p_0) = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}$$

,where $p_1 = (x_1, y_1)$ are the coordinates of body 1 (the hoop or player), and $p_0 = (x_0, y_0)$ are the coordinates of the other body.

The primary defender is defined as the player closest to the shooter, using these Euclidean distances. We define a shot near the three-point line to be a shot where $18 \leq d(\text{hoop}, \text{player}) \leq 35$. In the rest of this paper, we will merely refer to a shot near the three-point line to be a 3PT shot.

	Dist. Between Shooter And Defender (ft.)	Shot Dist. (ft.)	Shooter - Defender Height (in.)	3PT%
0	1.341465	20.852259	-1	0.338
1	5.086128	20.823000	-1	0.366
2	3.881703	21.964743	-1	0.378
3	3.329384	19.799617	-3	0.326

Table 1. Example rows and columns of Post-Processed Data
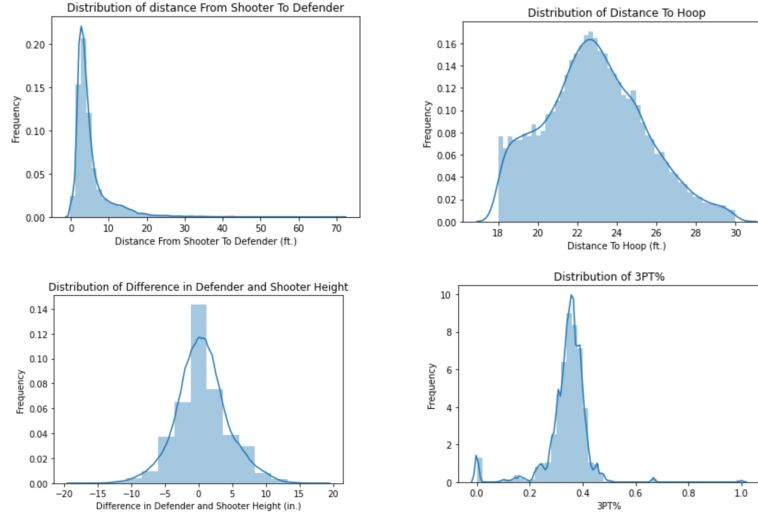
2.2. **Play-by-play.** We use play-by-play data from stats.nba.com for games from the 2015-16 season. This play-by-play is the accompanying data for the aforementioned SportVU dataset.

	GAME_ID	PCTIMESTRING	PERIOD	SCORE
0	0021300694	12:00	1	None
1	0021300694	12:00	1	None
2	0021300694	11:32	1	None
3	0021300694	11:32	1	None
4	0021300694	11:09	1	2 - 0

Table 2. Example rows and columns of play-by-play Data

This play-by-play data can help us narrow down the time frame when the shot occurred.

3. Methods

3.1. **Shot Probability Model.** Our preliminary goal is to devise a method to statistically determine the probability of a shot being made. Mathematically, we define:

- μ = Shot made (0 or 1)
- d_d = Distance to defender (ft.)
- d_s = Distance to hoop (ft.)
- d_h = Difference between shooter and defender height

Then, the probability of making a shot is

$$(3.1) \qquad\qquad P(\mu \mid d_d, d_s, d_h, 3PT\%)$$

We use XGBoost to model shot probability. Based on a hyperparameter search, we use the following hyperparameters: learning rate=0.05, max depth=3, n estimators=100, basescore=0.45, colsample bytree=1, subsample=0.8, gamma=0. Our chosen booster is 'gbtree'.
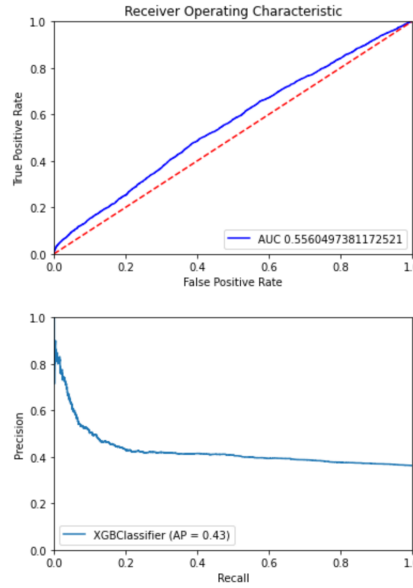

Figure 1. Model Metrics

Although our model's predictive power isn't extremely strong, we still perform better than if we only used 3PT% to make predictions. The league average 3PT% was 0.35, so a random estimator would have an AU-PRC of 0.35. We specifically want to deduce what the model is learning within this improvement above 0.35.

Figure 2. Example of Plotted Model

3.2. **Interpretation.** We are then able to interpret the model's predictions. Specifically, we wish to concretely determine which features the model finds to be the most useful to predict the shot probability.

To this end, we use shapley values (an idea from cooperative game theory): a concrete way of "splitting up" contributions among features. Shapley values assign specific negative and positive values, which signify whether a specific trait positively or negatively the model's predictions. The higher the shapley value for a given feature, the more the model's prediction was affected by that feature.

In order to solve for our shapley values, we use TreeSHAP. For each given datapoint (a single shot), we able to extract the shapley values for the aforementioned features fed into the model $\mu, d_d, d_s, d_h, 3PT\%$ ('Distance Between Shooter And Defender', 'Shot Distance', 'Difference Between Shooter And Defender Height', and '3PT%').

4. RESULTS

Out of all of the individual 3-point shots points, we select 70% for the train set and 30% for the test set.

Based on the trained XGBoost model and the TreeSHAP [2] interpretations, we are able to generate a summary and a detailed visualization of the calculated shapley values.
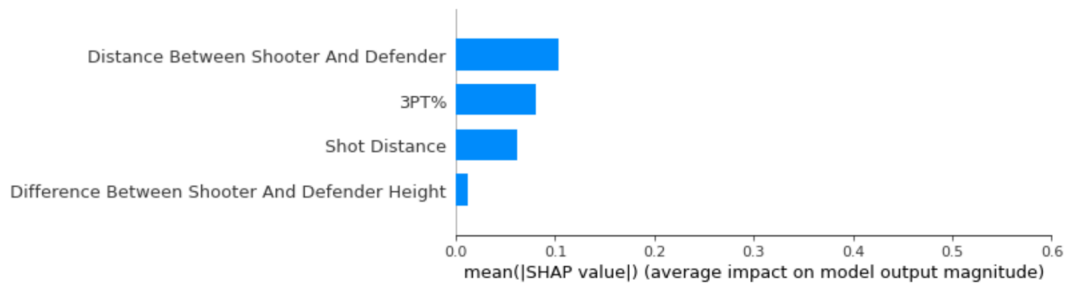

Figure 3. Summary of Shapley Values

In the above plot, we see the average shapley value for all of the data points. Specifically the distance between a shooter and their defender is more important than the 3PT%, Shot Distance, and the Difference between the shooter and defender height. In addition, the difference between a shooter's height and a defender's height has little to no significance when determining the probability of a made 3PT shot. Finally, the shot distance on a 3PT shot seems to be less significant than the Distance Between Shooter and Defender and the shooter's 3PT%.
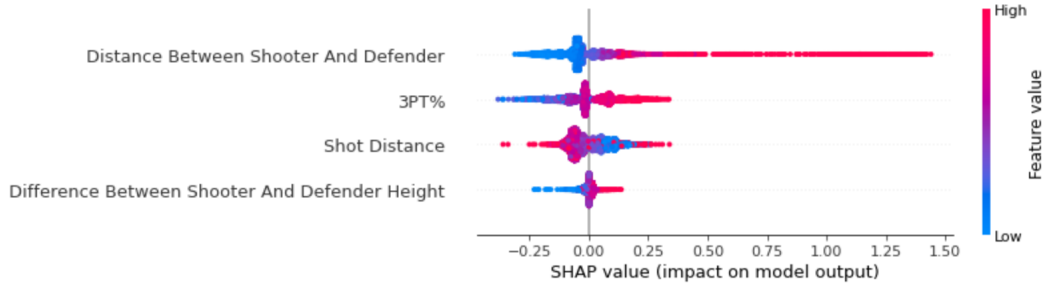
Figure 4. Detailed Shapley Values

In the more detailed version of our shapley value plot, we are able to pinpoint the trends for each of the features. For instance, in the 3PT% plot, we notice that the higher the 3PT%, the higher the shapley value. Although this specific information is fairly intuitive, it serves as a sanity check for how our model actually was able to learn.

Similarly we can determine the distribution of shapley values. For instance, there is not much variance in the shapley values for 'Difference Between Shooter and Defender Height', while there is significant variance in the 'Distance Between Shooter and Defender'.

5. Discussion

We believe that our ideology can help coaches adjust their strategies, optimizing for specific shooter situations.

In our specific research, we hope to calculate shapley values for specific players. For instance, if we can determine the associated shapley values for a given player on defense, the summation of all of these values across the season can bring us closer to a unified defensive statistic. This can also help isolate offensive achievement as well.

But beyond this research, we hope that our methods can show that shapley values are a field worth exploring in sports. Whether discussing the relative importance of specific attributes on a shot, or discussing lineups as a whole, we believe that the calculation of shapley values can help us understand the relative importance of features.

Our ideology is similar to that of Matt Ploenzke's Submission to the Big Data Bowl (Ploenzke, 2019). Specifically, we hope that our method can show the benefits to intepretable machine learning methods in general. Generally, machine-learning methods are considered to be black-box learning methods, but we believe that concepts like shapley values, can help decipher these methods. This can help us understand the way that these models are learning, allowing us to better understand sports as a whole.

References

Lundberg, S. M., Erion, G., Chen, H., Degrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., & et al. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Nba stats. (2020). https://stats.nba.com/

Ploenzke, M. (2019). Nfl big data bowl sub-contest.

Schafer, R. (2011). What is a savitzky-golay filter? [lecture notes]. *IEEE Signal Processing Magazine*, *28*(4), 111–117. https://doi.org/10.1109/msp.2011.941097

Seward, N. (2018). Sealneaward/nba-movement-data. https://github.com/sealneaward/nba-movement-data

Sill, J. (2010). Improved nba adjusted plus-minus using regularization and out-of-sample testing. *MIT Sloan Sports Analytics Conference*.