Finding Determinants of NBA Shot Probability using Interpretable Machine Learning Methods

By: Avyay Varadarajan[1]

[1] Mission San Jose High School, Fremont, CA

Abstract

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.

Introduction

- We wish to use an interpretable machine learning approach to determine the relative importance of various features that affect shot probability.
- We first train an XGBoost model to calculate individual shot probability given various features ('Distance Between Shooter And Defender','Shot Distance', 'Difference Between Shooter And Defender Height', and '3PT%')
- **XGBoost** is an efficient gradient boosting algorithm with high performance.
- We then wish to interpret this model using shapley values.
- **Shapley values** are a concept in game theory which provide a concrete way to understand how much a specific player or feature affects the outcome.
- In our case, we essentially use shapley values to represent the relative importance of our model's features.
- To extract these values, we use TreeSHAP.
- **TreeSHAP** allows us to calculate shapley values from our model's learned weights.

Data

- We use two data sources for our analyses: NBA play-by-play Data, obtained from stats.nba.com, and archived SportVU NBA Player Tracking Data.
- **Play-by-play data**
  - Use this data to determine the time when each shot occurred.
- **SportVU data**
  - Player and ball positional coordinates captured 25 times a second
  - We use this data to calculate 'Distance Between Shooter And Defender','Shot Distance', 'Difference Between Shooter And Defender Height'.

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

Thus, to model the shot probability, we use the following:

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

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

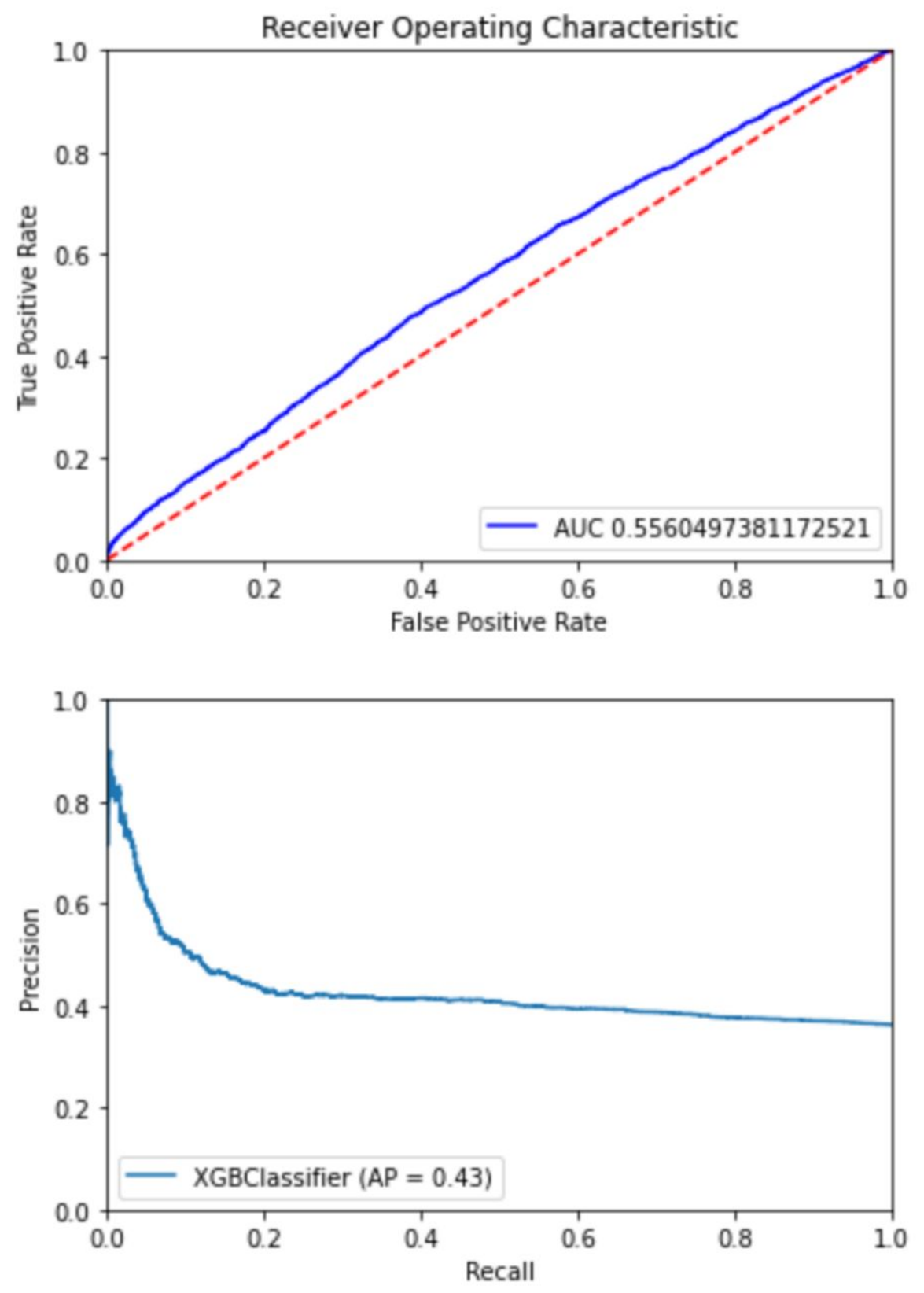
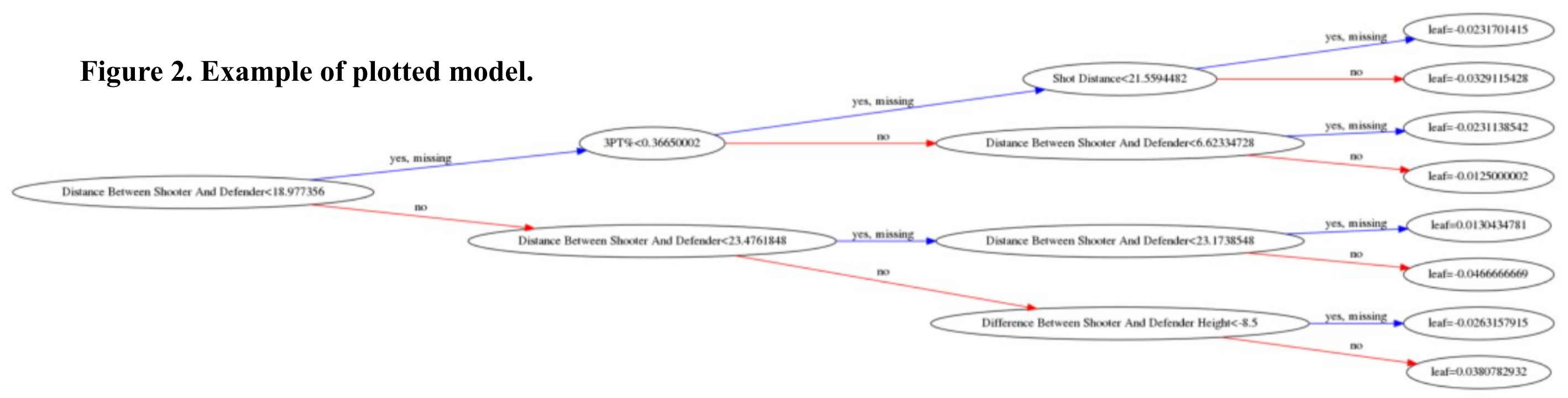


Figure 1. Model Metrics

Although our model's predictive power isn't extremely strong (AU-ROC=0.56, AU-PRC=0.43), 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.


Distance Between Shooter And Defender<18.977356
yes, missing
no
3PT%<0.36650002
Distance Between Shooter And Defender<23.4761848
Shot Distance<21.5594482
Distance Between Shooter And Defender<6.62334728
Distance Between Shooter And Defender<23.1738548
Difference Between Shooter And Defender Height<-8.5
leaf=-0.0231701415
leaf=-0.0329115428
leaf=-0.0231138542
leaf=-0.0125000002
leaf=0.0130434781
leaf=-0.0466666669
leaf=-0.0263157915
leaf=0.0380782932

Interpretation

We are now 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. 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 in to the model ('Distance Between Shooter And Defender', 'Shot Distance','Difference Between Shooter And Defender Height', and '3PT%').

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.

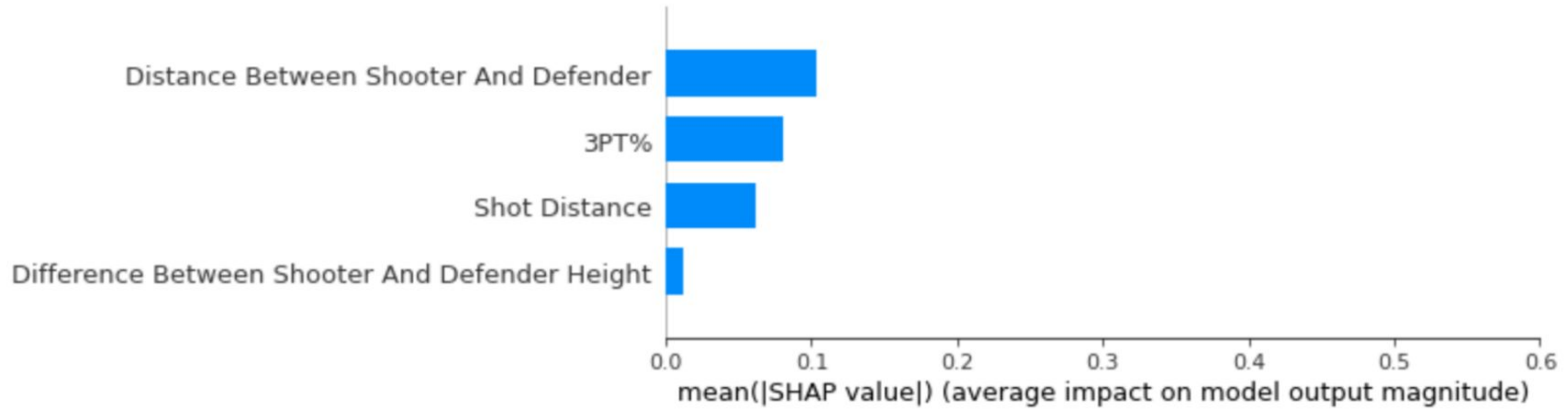
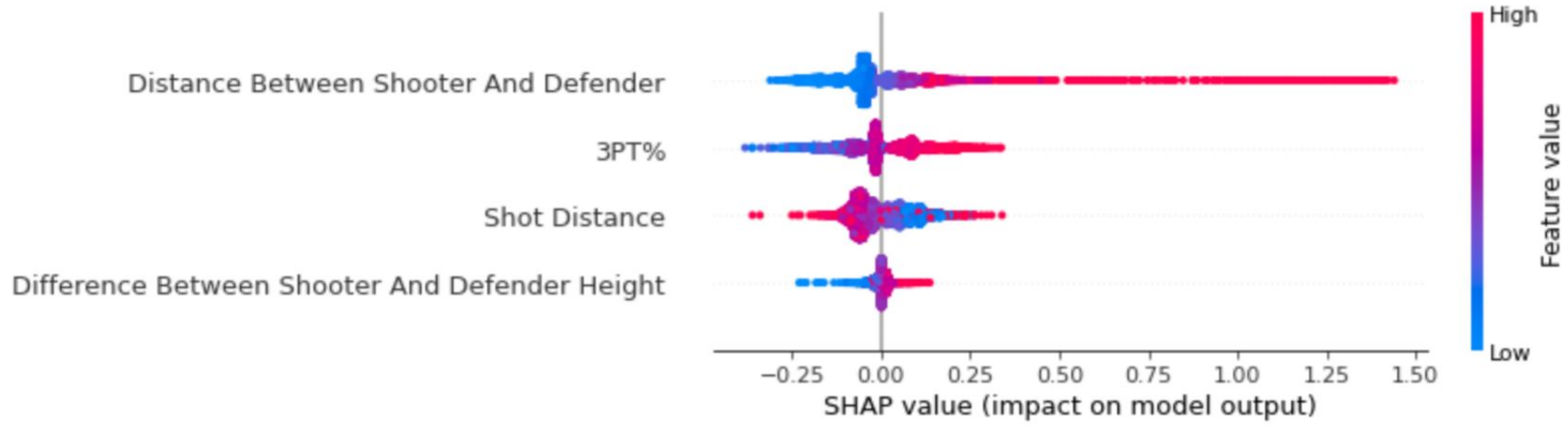


Figure 3. Summary of calculated shapley values

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%.

Interpretation Cont.

Figure 3. 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'.

Discussion/Conclusion

- 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.
- Our ideology is similar to that of Matt Ploenzke's Submission to the Big Data Bowl (Ploenzke, 2019).
- This can not only be aggregated over a season, but over specific games as well, allowing us to answer questions like: "How well did Anthony Davis play on defense during game 6 of the finals?"
- This can also help isolate offensive achievement as well.
- But beyond this research, we hope that our methods can show that shapley values area 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.
- Further, 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.
- Some improvements on this research include
  - Improving model performance
  - Comparing our shot probability model to existing shot probability models

Acknowledgements

- Alok Pattani (Google) for early feedback on this project.
- Vignesh Varadarajan (California Institute of Technology) for providing feedback.

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.